UNIONE EUROPEA
Fondo Sociale Europeo

UNIVERSITÀ DEGLI STUDI DI TORINO

DIPARTIMENTO DI INFORMATICA

DOTTORATO DI RICERCA IN INFORMATICA

CICLO: XXXVII

TITOLO DELLA TESI: Analysis of Public Administration Procurement and Expenditures Related to Energy Efficiency Improvements

TESI PRESENTATA DA: Roberto Nai

SUPERVISORI: Rosa Meo, Emilio Sulis

COORDINATORE DEL DOTTORATO: Viviana Patti

ANNI ACCADEMICI: 2021/2022, 2022/2023, 2023/2024

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA*: INF/01

(*N.B. Nel caso di più settori disciplinari interessati, deve essere indicato quello più presente nella trattazione della tesi).

La borsa di dottorato è stata cofinanziata con risorse del **Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI2014IT16M2OP005), risorse FSE REACT-EU, Azione IV.5 "Dottorati su tematiche Green"**

I dedicate this thesis to my maternal grandparents for their unwavering support; may these pages reach their hands, written in the desire to share every word with them.

Acknowledgements

The research presented in this dissertation would not have been possible without the fortunate meeting with my co-supervisor, Emilio Sulis.

All my work has also been made possible thanks to my PhD advisor, Rosa Meo, to whom I am deeply grateful for her patience and trust, especially during the challenges of the final year.

I also thank Viviana Patti for her support as PhD coordinator and my discussant, Ugo De' Liguoro, who is always available.

Thanks to the thesis reviewers Chiara Di Francescomarino and Livio Robaldo, whose suggestions improved my final work.

I am grateful to the Department of Management for their support in providing legal domain experts in public tenders, particularly Prof. Gabriella Margherita Racca and Dr. Francesco Gorgerino, for their contribution to the data analysis.

A heartfelt thanks to Andrei Buliga, Patrizio Bellan, Laura Genga, Chiara Ghidini, Francesca Meneghello, Massimiliano Ronzani, and Mozhgan Vazifehdoostirani for their valuable support throughout my research on Process Mining.

Looking back in time, I extend my heartfelt thanks to Daniele Codetta Raiteri, the first person who made me believe in pursuing a PhD. After 20 years, I can finally express my gratitude.

Last but not least, to the school principal, Maria Elena Dealessi, and all the staff and colleagues at I.T.I.S. "A. Volta", especially Cinzia, Laura, Giuseppe, Teresa and Virginia, for always making me feel at home. To Antonella, Enrico, Francesca, Mara, Marco, Graziella, Roberto, Silvia, and Simone, extraordinary colleagues who enriched my days and journey. To the exceptional Professors Bernardelli, Kostopoulos, Lesina, Masini, Porcelli, and Punta, who taught the true value of education.

Abstract

Administrative law governs the interactions between public authorities and private entities, including the rules for awarding public contracts. Public procurement law, a subset of administrative law, regulates the procedures for public tenders to ensure transparency, competition, and equal treatment. Public tenders, as a core mechanism for resource allocation and procurement of goods and services by governments and public institutions, play a critical role in fostering transparency, efficiency, and fairness within procurement processes.

In this context, ensuring and enhancing procurement processes' transparency, efficiency, and compliance remains a significant challenge. Informatics in the legal domain offers promising solutions, addressing key aspects such as data management and transparency through data standardisation, interoperability, and protection; automation and AI for decision-making, including tools to evaluate bids or predict contract risks; fraud detection and compliance through data mining, machine learning and process mining to uncover irregularities or prevent corruption; and legal text analysis to automate the processing of legal documents. This thesis tackles these open research problems by proposing an integrated approach that leverages advanced computational techniques to address these challenges effectively.

To investigate the analysis of the public administration procurement process and expenditures related to energy efficiency improvements, the thesis focuses on leveraging national datasets, such as those provided by the National Anti-Corruption Authority and the Italian Administrative Justice. By integrating these datasets, the thesis establishes a structured analytical approach that enables the application of machine learning techniques to identify patterns and factors associated with complaints. A relevant perspective concerns the construction of predictive models in the public procurement domain. The resulting models support proactive risk management and enhance transparency in public procurement processes, addressing critical challenges in this area.

To investigate green and energy issues, it has been considered important to focus attention on the broader analysis of issues related to public tenders, among which relevant sustainability issues concern the areas of healthcare, renewable energy, and expenditures on services and works of public agencies. The thesis also employs process mining techniques to analyse procurement workflows, leveraging data from the Tenders Electronic Daily dataset alongside

national sources to transform procurement information into event logs. Additionally, natural language processing techniques are utilised to enrich the event logs by extracting new events embedded in the textual descriptions of tenders. This approach enables the identification of bottlenecks, the evaluation of process performance, and the detection of areas for improvement in public tender procedures. Advanced techniques, e.g., Large Language Models, are extended beyond the enrichment of event logs to include exploring a recommendation system for automating the drafting of tender documents. This approach leverages historical data to support the generation of consistent, data-driven content, enhancing efficiency in public procurement.

The methods investigated in this thesis can also be used in other research areas. Extending beyond the legal domain, the thesis also explores the application of these techniques in the educational sector. By analysing event logs and predictive models, the research examines learning processes and showcases the adaptability and broader utility of the proposed methodologies. By addressing key challenges in data integration, process transparency, and decision support, this work contributes to advance the management of public procurement processes. It demonstrates the versatility of data-driven approaches across multiple domains.

By addressing these challenges, this thesis contributes to exploring data-driven solutions for the transparent, efficient, and compliant management of public procurement processes, offering potential applications across various domains.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ANAC National Anti-Corruption Authority

API Application Programming Interface

ARM Advanced RISC Machine

A.R.T. Accountability, Responsibility, and Transparency

AUC Area Under the Curve

BDAP Database of Public Administrations

BI Business Intelligence

BPM Business Process Management

BPMN Business Process Model and Notation

CA Contracting Authorities

CAI Cohere AI

CNN Convolutional Neural Network

CPV Common Procurement Vocabulary

CRM Customer Relationship Management

CV Cross-Validation

DB Database

DCG Discounted Cumulative Gain

DL Deep Learning

DNN Deep Neural Network

DT Decision Tree

EDU Education

EL Event Log

EO Economic Operators

ETL Extract, Transform, Load

FK Foreign Key

FPR False Positive Rate

GDP Gross Domestic Product

IAJ Italian Administrative Justice

ICT Information and Communication Technology

IDCG Ideal DCG

IE Information Extraction

IR Information Retrieval

KNN K-Nearest Neighbours

KPI Key Performance Indicator

LLM Large Language Model

LMS Learning Management System

LR Logistic Regression

MOOC Massive Open Online Course

NB Naive Bayes

NDCG Normalized Discounted Cumulative Gain

NLP Natural Language Processing

NUTS Nomenclature of Territorial Units for Statistics

OAI Open AI

OD Open Data

OECD Organisation for Economic Co-operation and Development

PHP Hypertext Preprocessor

PK Primary Key

PM Process Mining

PNNR Piano Nazionale di Ripresa e Resilienza

PPM Predictive Process Monitoring

RAG Retrieval-Augmented Generation

RF Random Forest

ROC Receiver Operating Characteristic

RS Recommendation Systems

SOA Società Organismi di Attestazione

SPA Single Page Applications

SVM Support Vector Machine

TED Tenders Electronic Daily

TNR True Negative Rate

TPR True Positive Rate

XAI Explainable Artificial Intelligence

XGB eXtreme Gradient Boosting

XSD XML Schema Definition

Chapter 1

Introduction and Background

Public administration greatly impacts a Country's economy by purchasing large quantities of goods and services to implement policies and provide public services. Such expenditure is a quote of the total value of all goods and services produced within a Country in a specific period, usually a year, i.e. Gross Domestic Product (GDP). In the Organisation for Economic Co-operation and Development (OECD) countries, public tender expenditure as a share of GDP increased over the last decades, from 11.8% of GDP in 2007 to 12.9% of GDP in 2021 [166]. In Italy, this quote reached the value of 11.8% of GDP in 2023 [166]. Analysis of government procurement and expenditures related to energy efficiency improvements can be carried out using automatic methods, extending the scope of application to various "green" areas where the role of the public sector is relevant, extending the scope of research to sustainability issues in a more general sense.

Comprehensive and clear knowledge of the public tender situation is paramount for a national State from a management perspective, increasing transparency and good governance. Moreover, better knowledge of public tender builds trust in institutions, increases accountability, and leads to better quality services for citizens. Businesses can benefit from having a picture of the current situation to understand the possible market space. Regulatory bodies can obtain valuable information to detect cases of corruption and fraud at an early stage. Furthermore, academic research can use such data across various disciplines, from economics to management and law.

The widespread dissemination of data through open standards has even more significant effects. A related aspect of interest concerns the combating of possible fraud cases [171]. According to the OECD [136], Open Data (OD) can help to design better anti-corruption policies and monitor their effective implementation. Increasingly, public administrations typically offer their data in institutional repositories, which can be freely accessed online.

The widespread availability of online public tender data fosters transparency and trust and presents opportunities for data science techniques to enhance efficiency and sustainability. As this digital shift progresses, the field of computer science is expected to play an increasingly critical role in supporting public administration for good governance.

1.1 Problem definitions

The tendering sector faces a range of challenges that hinder its efficiency, transparency, and sustainability. One of the main difficulties lies in the inherent complexity and unpredictability of public tenders. These processes are often characterised by a high likelihood of legal complaints, which can delay projects, inflate administrative costs, and erode trust in public institutions [91]. Such challenges are further exacerbated by fragmented legal frameworks across jurisdictions, evolving regulatory requirements, and the substantial volume of data involved. Traditional management tools frequently struggle to cope with these demands, resulting in inefficiencies and operational bottlenecks [266].

Legal complaints represent a particularly critical issue within the tendering process. These disputes often arise due to unclear or inconsistent tendering procedures, creating a need for predictive tools capable of identifying potential risks early in the process. Addressing these risks proactively could reduce delays and foster a smoother tender lifecycle [91, 26]. However, despite advances in legal informatics, areas like predictive analytics and process optimisation in public procurement remain relatively unexplored [68].

Another pressing challenge is the inefficiency in managing tender workflows. Organisations need tools that can provide accurate estimations of the time required to complete various stages of the process. This need highlights the importance of analysing and forecasting procedural workflows to identify bottlenecks and optimise timelines [248]. Moreover, the increasing complexity of public procurement emphasises the necessity for systems that not only predict outcomes but also explain them in a transparent and interpretable manner. Explainable AI (XAI) methods are particularly important in this context, as they ensure that predictive models gain the trust and acceptance of stakeholders [9].

In addition to these issues, the manual effort required to generate new tender documents presents a significant challenge. This process is time-consuming and prone to errors, underscoring the need for intelligent systems capable of learning from historical data to automate and optimise document drafting. Such systems would ensure consistency and relevance, particularly in contexts where sustainability and other domain-specific criteria play a crucial role [29].

Finally, the issue of data integration poses a persistent obstacle. Public procurement datasets are often dispersed across multiple unconnected repositories, making data access cumbersome and increasing the risk of data loss. Consolidating these datasets into a unified framework would enable more robust analysis and support innovative applications in the field [68].

1.2 Thesis Focus and Key Contributions

This thesis presents a comprehensive framework that leverages Machine Learning (ML), Process Mining (PM), and Large Language Models (LLMs) to tackle significant challenges in public procurement and extend their applicability to other domains. The analysis of the *green sector*[1] was extended to the entire spectrum of public contracts to address sustainability issues across multiple areas.

Firstly, the research applies ML algorithms to legal datasets to predict the likelihood of legal complaints in public tenders. The research started integrating legal datasets from multiple public repositories, including the National Anti-Corruption Authority (ANAC), the National Institute of Statistics (ISTAT) and the Italian Administrative Justice (IAJ). This predictive approach enabled early identification of potential disputes, supporting more informed decision-making and proactive risk management throughout the tender lifecycle. The study explores a range of classification algorithms, including Decision Trees (DT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and XGBoost (XGB), using evaluation metrics such as precision, recall, ROC/AUC, and F1-score to assess their performance. This effort facilitates robust analyses and lays the groundwork for future research in public procurement.

Another key area of focus is the application of Process Mining (PM) to optimise workflows within the tendering process. By transforming Tender Electronic Daily (TED) datasets into event logs, the research uncovers patterns and inefficiencies. Techniques such as Process Discovery and Variant Analysis provide actionable insights that help organisations streamline operations and manage timelines more effectively.

The research also investigates the potential of LLMs to create recommendation systems to support experts in drafting and optimising public tenders. The study evaluates the application of various models and embeddings, assessing their performance through domain-expert-defined queries. Among these queries, particular attention is given to topics such as sustainable procurement, including, for example, searches like the acquisition of electric

[1]The green sector in Italian public contracts encompasses activities promoting environmental sustainability, such as energy efficiency, waste management, renewable energy, and eco-friendly construction.

or hybrid transport vehicles. These systems leverage the ability of LLMs to process and analyse vast amounts of textual data, delivering intelligent, data-driven recommendations that enhance both the efficiency and accuracy of legal and tender management practices.

The application PM and Predictive Process Monitoring (PPM) techniques were further explored in different domains, including education (EDU) environments. A case study illustrates how these techniques can be utilised not only within the legal domain but also to enhance understanding of teaching and educational processes, showcasing the versatility and effectiveness of PM across various sectors.

Lastly, the research emphasises the importance of transparency and interpretability in predictive models. By employing XAI techniques such as SHAP (SHapley Additive exPlanations), the thesis ensures that stakeholders can understand and trust the results produced by automated systems, which are particularly critical in the context of public procurement.

The key contributions of this thesis lie in its application of ML, PM, and LLMs to real-world public procurement challenges, showcasing how these data-driven approaches not only improve current processes but can also inspire future advancements in procurement management.

1.3 Thesis Outline

This thesis is organized as follows. Chapter 2 presents the methods employed in this research, focusing on the key techniques used, including Information Extraction (IE), ML for classification, and PM techniques. The chapter also discusses the role of NoSQL databases, particularly Elasticsearch, in managing unstructured data.

Chapter 3 delves into the application of ML for predicting legal complaints in public tenders. It describes the datasets used, the predictive models developed, and the results of applying ML techniques to the Italian public procurement domain.

Chapter 4 explores the use of PM to assess transparency and efficiency in European public tenders. The chapter applies PM techniques to event logs derived from public procurement data, identifying process patterns and potential areas for improvement.

Chapter 5 investigates the development of recommendation systems based on LLMs and their application in public tenders, focusing on how LLMs can support the drafting and optimisation of tender documents.

Chapter 6 extends the application of PM and PPM techniques to EDU environments. It presents a case study on using PM to enhance e-learning by analysing short-term tutorials, identifying patterns in learning processes, and predicting outcomes.

Chapter 7 concludes the thesis by summarising the main findings, discussing their implications for public procurement, and proposing directions for future research in the field.

1.4 Related Publications

The following lists present the scientific publications related to the research activities described in this thesis. They are organised by chapter and present the publications related to this thesis, including their methodologies, experiments, and findings.

Chapter 2:

- Meo, R., Nai, R., and Sulis, E., "Explainable, Interpretable, Trustworthy, Responsible, Ethical, Fair, Verifiable AI... What's Next?" in *Advances in Databases and Information Systems - ADBIS 2022* [139].

Chapter 3:

- Nai, R., Sulis, E., Pasteris, P., Giunta, M., and Meo, R., "Exploitation and Merge of Information Sources for Public Procurement Improvement" in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* [161].
- Nai, R., Sulis, E., and Meo, R., "Public Procurement Fraud Detection and Artificial Intelligence Techniques: a Literature Review" in *23rd EKAW Int. Conf. proc.* [159].
- Nai, R., Meo, R., Morina, G., and Pasteris, P., "Public tenders, complaints, machine learning and recommender systems: a case study in public administration" in *Comput. Law Secur. Rev.* [154].
- Nai, R., Fatima, I., Morina, G., Sulis, E., Genga, L., Meo, R., and Pasteris, P., "AI Applied to the Analysis of the Contracts of the Italian Public Administrations" in *Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops* [153].

Chapter 4:

- Nai, R., Sulis, E., and Genga, L., "Automated Analysis with Event Log Enrichment of the European Public Procurement Processes" in *Advances in Conceptual Modeling - ER 2023 Workshops* [156].

Chapter 5:

- Nai, R., Sulis, E., Fatima, I., and Meo, R., "Large Language Models and Recommendation Systems: A Proof-of-Concept Study on Public Procurements" in *Natural Language Processing and Information Systems - NLDB 2024* [155].

Chapter 6:

- Nai, R., Sulis, E., and Genga, L., "Enhancing E-learning effectiveness: a process mining approach for short-term tutorials" in *Journal of Intelligent Information Systems* [157].
- Nai, R., Sulis, E., Marengo, E., Vinai, M., and Capecchi, S., "Process Mining on Students' Web Learning Traces: A Case Study with an Ethnographic Analysis" in *Responsive and Sustainable Educational Futures - EC-TEL 2023* [158].

Appendix A:

- Nai, R., Sulis, E., and Meo, R., "ITH: an open database on Italian Tenders 2016–2023" in *Scientific Data* [160].

Chapter 2

Methods

This chapter presents all the computational methodologies applied in the research outlined in this thesis, building on the foundational challenges and objectives introduced in Chapter 1. It provides an overview of the main methods, techniques and tools employed in investigating data collection and preprocessing, as well as in selecting models that are most suitable for research objectives, considering factors such as data characteristics, algorithmic requirements, and predictive performance, and evaluating their effectiveness through well-known metrics and validation strategies. Specifically, these methodologies encompass various approaches, from IE and data storage to ML algorithms and PM techniques. The strategies chosen were selected and tailored to address the research objectives with precision and reliability, ensuring robust results across various datasets and application domains. The methodologies outlined below in this chapter are critical not only for understanding general AI and ML applications but also for addressing specific challenges in public procurement, such as analysing tender documents and predicting compliance issues.

Section 2.1 introduces key *IE methods*, such as web scraping and browser automation, and discusses how the extracted data are stored using SQL and NoSQL databases, with a particular focus on the Elasticsearch tool. These methods were employed to construct the datasets analysed in Chapters 3, 4, and 5.

Section 2.2 focuses on *classification*, a key task in supervised learning, and provides a detailed overview of the techniques used in ML to assign data points to predefined classes. It outlines different types of classification problems, such as binary, multiclass, and multilabel, and introduces key algorithms. Additionally, the section delves into crucial evaluation metrics, including Accuracy, Precision, Recall, the F1-score, the ROC/AUC curve, and cross-validation. These methods were applied to predictions in the analysis of legal processes (Chapter 3) and educational processes (Chapter 6).

Section 2.3 examines *explainability* in ML, highlighting the importance of making models interpretable and transparent, particularly in complex systems where decision-making accountability is critical. These methods were implemented to explain the resulting prediction model in Section 3.4.4.

Section 2.4 introduces *PM discipline*, including the concepts and techniques of event logs, process discovery, variant analysis, and PPM, showing how data-driven insights can optimise business processes.

Section 2.5 explores the role of *LLMs in natural language processing tasks*, focusing on their architecture, applications, and the ethical considerations surrounding their use.

High-resolution images for this chapter are available at https://bit.ly/4eRONLT.

2.1 Information Extraction

In the digital age, vast amounts of information are generated and stored online in both structured and unstructured formats; extracting meaningful data from this vast expanse of information is critical for applications in fields such as data science, business analytics, and academic research [62]. *Information Extraction* (IE) refers to the process of automatically retrieving relevant data from various sources, transforming it into structured formats, and storing it for further analysis [173]. The primary challenge in information extraction lies in dealing with the heterogeneous nature of online data. While some data are stored in structured formats like tables or databases, a significant portion exists in unstructured formats such as text, images, and videos. IE techniques aim to bridge this gap, enabling the transformation of unstructured or semi-structured content into structured datasets that can be easily queried and analysed [6].

This chapter provides an overview of the key technologies and methods used in information extraction, focusing on how these techniques enable the creation of datasets from online data. The discussion will also introduce the role of *web scraping* [51] as a critical tool for data extraction and highlight how the collected data can be then managed in SQL or NoSQL databases [225].

Technologies for Information Extraction

Various tools and technologies are employed to implement IE techniques, focusing on Application Programming Interfaces (APIs) interaction, web scraping, browser automation, regular expressions, and data parsing to efficiently extract and structure information from static and dynamic unstructured sources.

1. *API Interaction*: when available, *APIs* [167] provide a more structured way of extracting data compared to web scraping. APIs allow direct access to the underlying datasets of web services and are often used alongside scraping tools to gather data in a more efficient and legal manner. When APIs are unavailable, scraping becomes the fallback method.

2. *Web Scraping*: web scraping is the primary method for extracting data from websites that do not provide structured access through APIs. Tools like *MechanicalSoup*[1] and *BeautifulSoup*[2] automate the process of downloading and parsing HTML content from web pages; these tools are crucial for gathering large datasets from dynamic or static websites and parsing them without big programming efforts.

3. *Browser automation*: for websites that use dynamic JavaScript [73] to load content, tools like *Selenium*[3] simulate human interaction with a browser, allowing for the scraping of content that only appears after certain actions, such as scrolling or clicking, have been performed.

4. *Regular Expressions (Regex)*: Regex [74] is a powerful tool for pattern matching within the text, often used to extract specific pieces of information from scraped data. For example, regex can be employed to capture patterns like email addresses, dates, or phone numbers from web content. Although it lacks the sophistication of other methods, it is highly effective for straightforward extraction tasks due to its speed and simplicity.

5. *Data Parsing Libraries*: tools like *pandas*[4] in Python are frequently used after scraping to clean, process, and structure the extracted data for analysis. These libraries assist in converting raw web content into structured datasets that can be easily manipulated for further IE tasks.

Transition to Web Scraping and Database Technologies

Web scraping is an invaluable tool for academic research, particularly in fields that rely on large-scale data collection from online platforms. Researchers in disciplines such as social sciences, linguistics, economics, and computer science have employed web scraping techniques to collect datasets for empirical studies that would be otherwise impossible or

[1] https://pypi.org/project/MechanicalSoup
[2] https://pypi.org/project/beautifulsoup4
[3] https://pypi.org/project/selenium
[4] https://pypi.org/project/pandas

prohibitively time-consuming to obtain manually. Once data are extracted and structured, it must be stored in a way that supports efficient querying and analysis. This is where database technologies come into play. Two main types of databases are typically used to store extracted data: SQL and NoSQL databases. The well-known SQL databases are suited for storing data with fixed structures, such as financial records or customer information, while NoSQL databases are more flexible and can handle unstructured or semi-structured data, making them ideal for scenarios where data schemas are subject to change. Thus, NoSQL databases are often employed to store web scraping results, as the structure of the extracted data may vary across different web pages.

2.1.1 Web Scraping

Web scraping refers to the process of automatically extracting data from websites and it's an essential tool for researchers and developers who seek to gather large datasets from the web without the need for manual copying [272].

This chapter explores the technical principles of web scraping and discusses its applications and challenges.

Web Scraping - Basis techniques

At its core, web scraping involves sending an HTTP [88] request to a server to retrieve the HTML [152] content of a web page, followed by parsing this content to extract the required data. The general workflow for web scraping can be broken down into several steps:

1. Sending HTTP Requests: tools such as MechanicalSoup are commonly used to send HTTP requests to the web server, which then responds by returning the HTML code of the web page.

2. Extracting Data: after identifying the relevant HTML elements, the scraper extracts and downloads the necessary information, including text, images, links, or other forms of data embedded in the page.

3. Parsing HTML Content: once the HTML is obtained, a parser (such as BeautifulSoup) is used to navigate the *Document Object Model* (*DOM*) [93] structure of the web page and identify the elements containing the required data.

4. Storing Data: the extracted data are typically saved in a structured format, such as a Comma-Separated Values (CSV) [219] or JSON [176] file or database, making it ready for further analysis.

Web Scraping - Advanced techniques

While the basic process is straightforward, more advanced techniques are often required to deal with complex websites. Modern websites may load content dynamically using JavaScript, which can prevent traditional scrapers from accessing the full data. In such cases, tools like Selenium can be employed to automate a browser and retrieve content after JavaScript has been executed. Other techniques include handling pagination to scrape data spread across multiple pages and using proxies or rotating user-agent strings to avoid detection and blocking by websites that employ anti-scraping mechanisms. Selenium is a popular tool used to automate web browsers, allowing scrapers to interact with websites in a manner similar to how a human user would. Selenium supports multiple web browsers, such as Chrome, Firefox, and Safari, and can be used to load dynamic content, handle JavaScript execution, and navigate interactive websites. This makes it an ideal tool for scraping content from single-page applications (SPAs) that load elements asynchronously using technologies like AJAX.

Web Scraping - Challenges

Although increasingly established and widely used, web scraping presents several challenges: changing website structures that break static scrapers, IP blocking that requires rotating proxies or headless browsers, and legal uncertainty surrounding its practice, which may involve potential legal risks. The limitations of web scraping are outside the scope of this research, and for further details on its challenges, see related studies such as [221] on ethical web scraping practices.

Conclusion

Web scraping is a valuable and versatile tool in the era of big data, enabling the extraction of large datasets from websites for analysis and research. However, it comes with technical challenges, as well as ethical and legal considerations that must be carefully navigated. The evolution of scraping technologies, combined with the increasing use of anti-scraping mechanisms by websites, suggests that this area will continue to evolve. Future developments may include more sophisticated tools for scraping dynamic content and more robust legal frameworks to regulate its use.

2.1.2 NoSQL tools and Elasticsearch

NoSQL databases are designed to handle large volumes of unstructured or semi-structured data that traditional relational databases struggle with. Unlike SQL databases, they offer more flexibility by allowing schema-less data models and horizontal scaling, making them well-suited for modern web applications and big data environments [225]. NoSQL databases can be categorised into four main types:

1. Document Stores (e.g., MongoDB): these store data in document format (usually JSON or BSON [138]), which allows for flexible schema designs [11].

2. Key-Value Stores (e.g., Redis): these are optimised for simple lookup operations, storing data as key-value pairs [33].

3. Column-Family Stores (e.g., Cassandra): these are designed for reading and writing large amounts of data in distributed systems, using tables with rows and dynamic column families [98].

4. Graph Databases (e.g., Neo4j): these focus on relationships between data points and are used for graph-based models like social networks [202].

Among the diverse NoSQL databases, Elasticsearch (ES) [87] is particularly prominent within the Document Stores category. According to DB-Engines[5] rankings, it consistently ranks as one of the most widely adopted solutions for search and text analysis, with a significant market share. ES is distributed free of charge[6].

Elasticsearch

ES is a distributed, RESTful [169] search engine built on top of Apache Lucene [137]. While it is often classified as a search engine, ES functions as a NoSQL database, offering highly efficient indexing and search capabilities for semi-structured and unstructured data, typically JSON documents. These documents are indexed and stored in a distributed manner, allowing for efficient querying and retrieval. However, ES stands out due to its optimisations for full-text search and analytics. Unlike traditional NoSQL systems, which focus on data storage, ES is designed with a primary emphasis on high-performance text search and real-time analytics, making it a powerful tool for applications that involve large-scale text data. Finally, ES supports *embedding* [140] storage, enabling advanced similarity searches and ML-based applications. In conclusion, ES is an essential tool for handling large-scale, semi-structured or unstructured datasets, particularly when full-text search and efficient querying are required.

[5] https://db-engines.com/en/ranking/search+engine
[6] https://www.elastic.co

2.2 Machine Learning

ML, as a subfield of artificial intelligence (AI), focuses on developing algorithms that allow systems to learn from data and make decisions without being explicitly programmed [150].

2.2.1 Classification - Definition and Algorithms

Classification is one of the most important tasks in *supervised learning* [127], where the goal is to predict a discrete label or category based on input data [84]. Classification problems are ubiquitous, ranging from medical diagnosis, spam detection, and sentiment analysis to image recognition. This section explores the foundations, key algorithms, evaluation metrics, and final considerations in classification tasks.

Problem Definition

Classification involves assigning an input to one of two or more predefined categories. Formally, given a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_i represents the input features and y_i represents the corresponding label, the objective is to learn a function $f : X \rightarrow Y$, where X is the input space and Y is the label space. The function f is designed to predict the label for unseen data, ideally generalising well beyond the training set [97].

Types of Classification

- *Binary Classification*: the simplest form of classification, where the goal is to classify inputs into one of two categories (e.g., spam or not spam) [97].
- *Multiclass Classification*: this involves classifying inputs into more than two categories (e.g., recognising digits from 0 to 9) [151].
- *Multilabel Classification*: in this variant, each input can belong to multiple categories simultaneously (e.g., tagging an image with multiple objects) [242].

Classification Algorithms

Several ML algorithms are designed for classification tasks, each with distinct mechanisms and assumptions about the data. Below is an overview of some of the most widely used classification algorithms:

Logistic Regression

Logistic regression (LR) [268] is a linear model used primarily for binary classification. Despite its name, it is a classification algorithm rather than a regression model. The hypothesis for logistic regression is that the probability of an instance belonging to a particular class is a logistic function (sigmoid) of an input features' linear combination:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Where $\beta_0, \ldots, \beta_p$ are the model parameters learned from the data. LR is particularly effective for linearly separable data.

Decision Trees

A decision tree (DT) [112] is a non-parametric model that partitions the input space into regions defined by a series of decisions based on feature values. Each internal node represents a test on a feature, and each leaf node corresponds to a class label. DT are easy to interpret but prone to *overfitting*[7].

Random Forest

Random Forest (RF) [28] is an ensemble method [56] built on DTs. By constructing multiple trees on different random subsets of the data and aggregating their predictions (via majority voting for classification), RF reduces overfitting and improves accuracy. This method is particularly robust in handling large datasets with many features.

Support Vector Machines

Support Vector Machines (SVM) [263] are powerful classifiers that work by finding the hyperplane that best separates the data into different classes. SVMs use support vectors (key data points) to maximise the margin between the decision boundary and the nearest data points from each class. SVMs can be extended to handle nonlinear classification tasks using kernel functions (e.g., radial basis function, polynomial kernel).

[7] Overfitting in ML occurs when a model learns not only the underlying patterns in the training data but also the noise and irrelevant details. This causes the model to perform well on the training data but poorly on new, unseen data because it has become too tailored to the specifics of the training set [55].

k-Nearest Neighbours

k-Nearest Neighbours (k-NN) [175] is a simple, instance-based learning algorithm that classifies new data points by looking at the k closest examples from the training set. The majority label among the nearest neighbours is assigned to the new data point. Although effective for small datasets, k-NN can be computationally expensive for large datasets.

XGBoost

eXtreme Gradient Boosting or XGBoost (XGB) [41], is a highly efficient and scalable implementation of gradient boosting designed for supervised learning tasks, including classification. XGB builds an ensemble of decision trees in a sequential manner, where each subsequent tree attempts to correct the errors made by the previous trees. XGB is especially effective in dealing with large and sparse datasets. It is often the top choice in ML competitions because of its flexibility and accuracy.

2.2.2 Classification - Evaluation Metrics

The performance of classification models is typically evaluated using a variety of metrics, particularly when dealing with imbalanced datasets [82]. The following are some commonly used evaluation metrics for classification tasks.

Accuracy, Precision, Recall, F1-Score

Accuracy is the ratio of correctly predicted instances to the total number of instances [223]. Although widely used, accuracy can be misleading for imbalanced datasets, where the model may predict the majority class and achieve high accuracy.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Precision is the ratio of correctly predicted positive instances to the total predicted positives [223].

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Precision is important in situations where the cost of false positives is high. For example, in email spam detection, incorrectly flagging legitimate emails as spam (false positives) can lead to important communications being missed.

Recall (also known as Sensitivity or True Positive Rate or TPR) is the ratio of correctly predicted positives to all actual positives [223].

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is crucial in scenarios where it is important to capture all positive cases, even at the cost of more false positives. For example, in medical diagnostics for life-threatening conditions, it is better to have more false positives than to miss a true positive case.

Specificity (or True Negative Rate or TNR) measures the proportion of actual negatives that are correctly identified as such by the model [223].

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Specificity is particularly important in contexts where it is crucial to avoid false positives, such as in medical testing, where misclassifying a healthy person as sick can lead to unnecessary interventions.

F1-Score is the harmonic mean of precision and recall, offering a balanced measure when the classes are imbalanced [223].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-Score is particularly useful when an uneven class distribution and a balance between precision and recall is desired. For example, in fraud detection, missing fraudulent transactions (false negatives) and flagging legitimate transactions as fraudulent (false positives) have significant costs, so the F1-Score provides a balanced metric.

ROC/AUC Curve

In addition to metrics like Accuracy, Precision, Recall, and F1-Score, another important tool is the *Receiver Operating Characteristic* (ROC) curve. This curve visualises the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR), which is equivalent to (1 - TNR). Thus, the *Area Under* the ROC *Curve* (AUC) is used to summarise a classifier's performance across different thresholds. The higher the AUC, the better the model is at predicting negative classes as negative and positive classes as positive [67]. Figure 2.1 displays three examples of ROC/AUC curves: Figure 2.1.a represents an ideal situation in which the model is perfectly able to distinguish between positive class and negative class, Figure 2.1.b represents a moderate classifier where AUC is 0.8, so it means there is an 80% chance that the model will be able to distinguish between positive class and negative class,

Figure 2.1.c represents a poor classifier with AUC approximately 0.5, so the model has no discrimination capacity to distinguish between positive class and negative class.

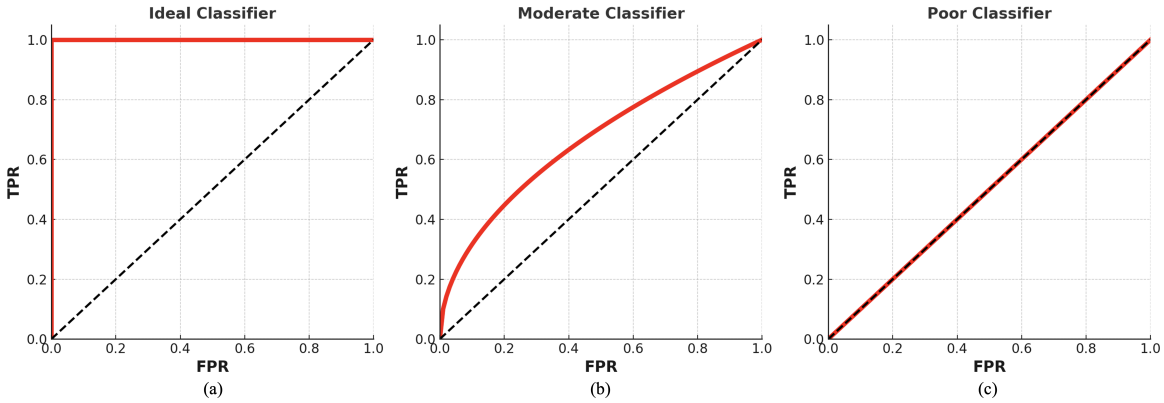


Fig. 2.1 ROC/AUC curve sample: (a) represents an ideal classifier, (b) represents a moderate classifier, (c) represents a poor classifier

Cross-Validation

While individual metrics such as Accuracy, Precision, Recall, and the F1-score are essential for evaluating model performance, ensuring these metrics generalise well to unseen data is equally important. This is where *cross-validation* (CV) comes into play. Cross-validation is a robust technique that addresses this need by splitting the dataset into several parts, or *folds*, and training the model on one portion while testing it on another. This process is repeated multiple times to provide a more reliable estimate of the model's performance and prevent issues such as overfitting. Common strategies, such as *k-fold* CV, offer approaches to balance bias and variance in model evaluation [191].

k-Fold CV is one of the most widely used techniques in ML for evaluating model performance while ensuring that it generalises well to unseen data. The key idea behind k-fold CV is to divide the dataset into k equally sized folds. The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time with a different fold acting as the test set, and the performance is averaged across the k iterations [255]. For example, considering a dataset with 1000 data points and a 5-fold CV setup: (i) the dataset is divided into 5 equal parts, each containing 200 samples; (ii) in the first iteration, 800 samples are used for training, while 200 samples form the test set; (iii) this process is repeated for the remaining folds, with a different fold serving as the test set in each iteration; (iv) the performance metrics from all five iterations are then averaged to provide a reliable estimate of the model's overall performance [163].

Conclusion

Classification remains a cornerstone of ML, with wide-ranging applications across industries. From basic algorithms like LR to sophisticated algorithms such as XGB, various tools exist to tackle classification problems. The choice of algorithm depends on various factors, including the nature of the dataset, performance requirements, and the need for model interpretability. Understanding classification techniques and evaluation metrics is crucial for developing robust and real-world ML models.

The classification algorithms discussed above are particularly suited for tasks like identifying high-risk tender documents. These methods ensure that predictive models can adapt to the specific characteristics of procurement data, improving decision-making processes in public tenders [115].

2.2.3 Deep Learning

Deep Learning (DL) is a subset of ML that focuses on models known as deep neural networks. It has gained significant prominence due to its ability to handle complex patterns and data structures. Unlike traditional ML algorithms, DL models are designed to automatically learn feature representations from large datasets, making them particularly effective in tasks such as image recognition, natural language processing, and speech analysis [116].

Deep Learning for Classification Tasks

Classification is one of the primary tasks in ML, where the goal is to assign input data into one of several predefined categories. DL has proven to be highly effective for classification problems, outperforming traditional methods in complex scenarios. Convolutional Neural Networks (CNNs) [117, 113] and Feedforward Neural Networks (FNNs) [100] are commonly used as architectures in classification tasks, especially for structured data like images and tabular data. The depth of these networks enables them to capture hierarchical patterns, which are crucial for tasks like image recognition and natural language understanding [116].

Recurrent Neural Networks and Long Short-Term Memory Networks

Recurrent Neural Networks (RNNs) [63] are designed to handle sequential data by maintaining an internal state that evolves over time, making them particularly useful for tasks such as time series prediction, language modelling, and speech recognition. However, RNNs suffer from the vanishing gradient problem, where gradients become too small to effectively update the model weights during training, particularly for long sequences. To address this,

Long Short-Term Memory (LSTM) networks were introduced by [99]. LSTMs incorporate mechanisms known as gates (input, forget, and output gates), which allow the network to selectively retain or forget information across time steps. This makes LSTMs particularly powerful for capturing long-term dependencies in data.

Applications of LSTM in Classification

LSTMs are widely used in classification tasks that involve sequential data, such as text classification, sentiment analysis, and speech recognition. For example, in sentiment analysis, LSTMs are able to learn the context of a sentence by retaining important information from previous words, improving accuracy compared to other models. Moreover, they outperform standard RNNs in tasks where long-term dependencies are crucial [90].

Conclusion

Despite the success of DL in classification tasks, training these models is computationally expensive and often requires large datasets. Additionally, tuning hyperparameters and preventing overfitting are significant challenges. In conclusion, DL has revolutionized classification tasks, providing models that can automatically learn from raw data. Among these, LSTM networks have proven particularly useful for sequence classification tasks by solving the limitations of traditional RNNs. As research continues, new architectures and techniques are likely to push further the boundaries of what DL can achieve.

2.3 Interpretability and Explainability in AI

Explainability in AI and ML has become an essential topic in recent years, driven by the increasing deployment of these technologies in critical applications such as healthcare, finance, and autonomous systems. Complex AI models, particularly DL models, often act as *black boxes*, making decisions that are difficult to interpret or explain. This lack of transparency challenges trust, accountability, and ethical use of AI systems [145].

Explainable AI (XAI) seeks to provide insight into how models arrive at their decisions, enabling stakeholders—ranging from engineers and researchers to end-users—to understand better, trust, and interact with AI systems [198]. A core focus in XAI is balancing the trade-off between model performance and interpretability, as more accurate models tend to be less interpretable [130].

This section explores various dimensions of explainability in AI and ML, from fundamental concepts to practical methods. The current challenges and limitations in the field are discussed, along with emerging trends and future directions for research.

2.3.1 Explainable AI

AI refers to computational systems whose actions and decisions resemble human intelligence, including functions typically associated with intelligence, such as learning, problem-solving, planning, and acting rationally, as defined by Russell and Norvig [208]. AI is interpreted broadly to include closely related areas such as ML. Systems that heavily use AI, have had a significant impact in domains that include healthcare, transportation, finance, social networking, e-commerce, and education. These "intelligent" systems have almost pervaded all the areas of our modern society. This growing societal impact has brought a set of risks and concerns, including the mistakes that AI systems can make. As a response, researchers are trying to design and deploy a new generation of systems that are trustworthy, i.e, meritable of trust from human beings and more robust to errors in software, resilient to cyber-attacks, and secure, in presence of incomplete scenarios.

The ingredients for a trustworthy AI are manifolds. This is related to the deployment of an AI product: sometimes the output of an AI system is used to support the decision-making, and in this case, end-users will need to trust the outcomes of the artificial model. Other times the system is used to inform the user about the inner structure of the instances coming from the application domain. In these cases, the end-user needs to be convinced that the system has grasped a meaningful organization of the application domain examples.
Systems whose outcomes cannot be well-interpreted are difficult to trust, especially in sectors, such as healthcare or self-driving cars, in which the impact of an erroneous decision has moral and fairness implications [123]. This need for models that are trustworthy, fair, robust with respect to missing data, high-performing in the real-world applications led to the revival of eXplainable AI (XAI) [92]. This field focuses on the understanding and interpretation of AI systems' behavior. The popularity of the search term "Explainable AI" in the last five years, as measured by Google Trends, is illustrated in Figure 2.2. The noticeable spike in recent years reflects also the increased research output of the same period.

XAI is not a monolithic concept: it reflects several related notions. The *explainability* and *interpretability* terms are usually used interchangeably [109, 22]. However, while they are very closely related, some works identify differences among related concepts [124]. They will be distinguished in the following sections. XAI has numerous applications: model validation, model debugging, and knowledge discovery [61]. The obtained explanations should show whether a ML model is grounded upon the possible biases in the training data

or when the learned models ignore important parts of the input data and instead rely on irrelevant ones. They could show that the flaws of the models could be caused by flaws in the training data.

As the demand for more explainable ML models with interpretable predictions rises, so does the need for methods that can help to achieve these goals. XAI is centered on the challenge of demystifying the black boxes but also implies *Responsible AI* as it can help to produce transparent models. Responsible AI takes into account societal values and moral and ethical considerations. Responsible AI has three main concepts: *Accountability*, *Responsability*, *Transparency*; these are called the A.R.T. of AI [57]. Finally, XAI is a part of a new generation of AI technologies called the *third wave AI* [270]. One of the objectives of this ambitious "wave" is to precisely generate models that can explain themselves.

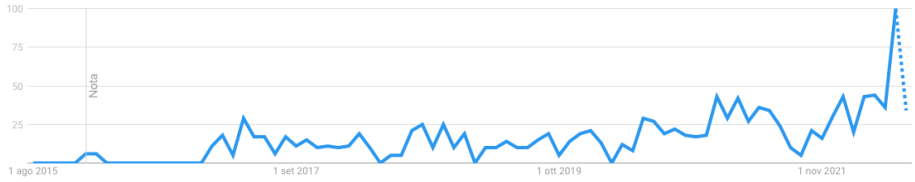


Fig. 2.2 Google Trends popularity index of the term "Explainable Artificial Intelligence" over the years 2017–2022

2.3.2 Explainability Techniques

Explainability is more related to the techniques thought to convince the end-user about the validity of the model outcomes. The most common methods are providing post hoc explanations or recalling from the domain similar instances to the given one in input [61]. These post hoc explanations are local, and specific to single instances and can be model-agnostic or specific to the single method. The model-agnostic ones treat the model to be explained as a black box and assume the predictions of the global model can be approximated as the application of many interpretable white-box models, valid locally, in a small neighbourhood of each input. Then, they sample the feature space in the neighbourhood of each instance to prepare a training set that is passed to train a white-box model, such as a sparse linear model (Lasso), or if the local behavior is non-linear using if-then rules. Another approach is to determine the importance of each feature on the model by measuring the impact of features' perturbations on the output score. The results may be interpreted as counterfactual explanations that describe a causal relationship between the input X and the output Y. They have the form: "If input X had not occurred, output Y would not have occurred".

Explanation approaches designed for a specific type of model leverage on the characteristics of the model to explain them. For instance, in the case of Deep Neural Networks (DNN), their structure must be treated as a white box, with a detailed description of their components. There are three methods: back-propagation methods (top-down) compute the gradient of specific outputs with respect to the input and back-propagate it to derive the contribution of each feature. This method can be efficiently implemented in software libraries (PyTorch or TensorFlow) as a modified gradient function but can give noisy explanatory results. Perturbation methods work bottom-up (with mask perturbations in an optimization framework) and learn a perturbation mask that preserves the contribution of each feature and can be trained by an additional DNN. The intermediate methods either transform the representations at the higher layers of the DNN into a synthetic image together with an encoding of the target object in a mask, or they adopt a prediction's decomposition through the additive contribution of the hidden vectors in the DNN corresponding to each input (e.g, a word in the textual input to a Recurrent Neural Network). Therefore, each component of the decomposition quantifies the contribution of each input to the DNN output.

2.3.3 Interpretability

One of the most popular definitions of interpretability is the one of Doshi-Velez and Kim, who, in their work [60], define it as "the ability to explain or to present in understandable terms to a human". Interpretability is more focused on the task of model properties' exploration with the goal of providing transparency to humans. For instance, clarifying the meaning of the components of a black-box model, like a deep neural network or a Support Vector Machine with the goal of understanding the model. The most common technique is to put aside an obscure model a "white-box" model, trained on the same instances. The latter model incorporates interpretability directly into its structure. This is the case of logical models (decision tree or rule-based model), linear models (that accompany features with coefficients whose magnitude informs their impact on the model outcome), attention model (for natural language, referred to as the words in the context).

One of the more interesting goals of learning an interpretation of a black box model is to understand the representations of the input (images) captured by the Deep Neural Network (DNN) model like a Convolutional Neural Network (CNN). The reference here is to the internal network nodes of CNN, as it is known that they encode artefacts learned from the input images. One of the most effective methods is finding the inputs that best activate neurons at a specific layer [61]. The optimization should be regularized using natural image priors produced by a generative model (GAN). Instead of directly optimizing the image, these methods optimize the latent space codes of the GAN to find an image that activates a

given neuron. The visualization results provide several interesting observations. The neurons from the first layer to the last layer learn representations at several levels of abstraction, from general to task-specific. The second interesting learned issue is that a neuron is multifaceted, i.e., it could respond to different images semantically related to the same concept (i.e. faces). CNN learns distributed code for objects and learns objects by the representation of their parts that can be shared across different categories [61].

Based on the above, interpretability is mostly connected with the intuition behind the outputs of a model [4] and the idea that the more interpretable a ML system is, the easier it is to identify cause-and-effect relationships within the system inputs and outputs. Doshi-Velez and Kim [60] proposed the following classification of evaluation methods for interpretability: application-grounded, human-grounded, and functionally-grounded; Figure 2.3 shows the taxonomy proposed. Application-grounded evaluation concerns itself with how the results of the interpretation process affect the human, domain expert, and end-user in terms of a specific and well-defined task or application. Human-grounded evaluation is similar to application-grounded evaluation; however, there are two main differences: first, the tester, in this case, does not have to be a domain expert but can be any human end-user, and secondly, the end goal is not to evaluate a produced interpretation with respect to its fitness for a specific application but rather to test the quality of the produced interpretation in a more general setting and measure how well the general notions are captured. Functionally grounded evaluation does not require any experiments that involve humans but instead uses formal, well-defined mathematical definitions of interpretability to evaluate the quality of an interpretability method. This type of evaluation usually follows the other two types of evaluation: once a class of models has already passed some interpretability criteria via human-grounded or application-grounded experiments, then mathematical definitions can be used to rank the quality of the interpretability models further.

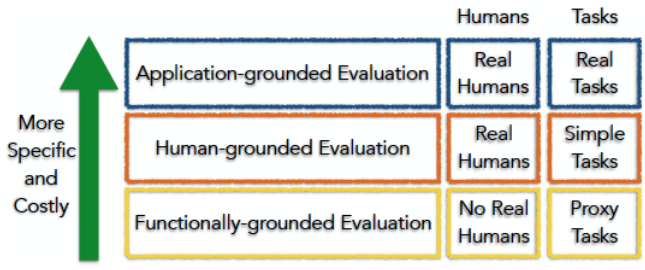


Fig. 2.3 Taxonomy of evaluation approaches for Interpretability [60]

2.3.4 The problem of bias

AI explanations might reveal that decisions are influenced by factors not aligning with explicit organizational policies. Amazon cancelled a plan to use AI to identify the best job candidates for technology positions upon discovering the models were biased against women because the training data consisted predominantly of males, reflecting historic hiring practices [46]. The example above explains that biases in AI mean biases in predictions. The ethical consequences of AI systems' algorithmic decision-making are a great concern. The emergence of biases in AI-led decision-making has seriously affected the adoption of AI. To build an unbiased system, a strong sense of justice needs to be in place to help decision-makers act fairly without having any prejudice and favoritism [14].

The survey presented in [224] discusses the different sources of bias. According to the authors, Figure 2.4 shows a taxonomy of bias sources.

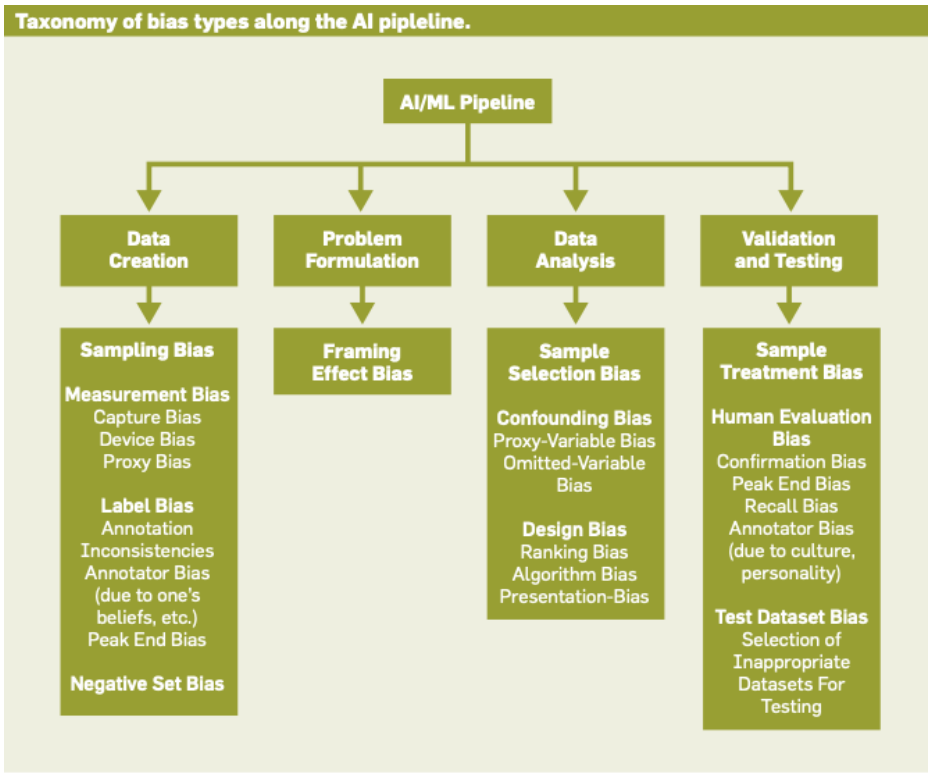


Fig. 2.4 Taxonomy of biases [224]

As it is well known, the knowledge discovery process in AI stems from a pipeline composed of different steps: data source cleaning, integration, feature selection or feature construction, model training, selection, validation, and finally, outcome presentation. All these steps might be the source of some bias. Some might be due to the users/analysts insufficient knowledge/preparation that comes out under the multiple forms of employing a

sampling bias, showing a capture bias, a device bias, a measurement bias, or a negative set bias (insufficient examples for the negative class), or a confirmation bias (that leads to ignore some relevant issues in the domain). All these examples of bias could lead to unsuitable choices in the data preparation.

Other biases could come from the presence of an ill-posed domain problem: the framing effect bias is sometimes due to the need to formulate the problem so that the experimentally measured results could reflect some business objective. Another source of mistakes in the AI model is the confounding bias that exists when an omitted feature is not included in the training data: this makes it impossible to measure the correlation between causes and effects. Another example is the inclusion of a proxy feature that is the source of indirect discrimination (e.g., zip code could be correlated to the ethnic condition), a sensitive feature that should be omitted to avoid discrimination based on ethnic conditions).

Other biases are algorithm biases: which influence on the model outcome by how they explore the hypothesis and evaluate constraints or how they present their results in a ranking to the users waiting for feedback. The users could not be impartial or ready in their evaluation due to a recall bias. Finally, the deployment of the AI model might, in turn, influence and alter the studied scenario.

2.3.5 Conclusions

A summary was provided on the rapidly evolving field of explainable and interpretable AI. While many application areas of AI systems require trust, fairness, and responsible principles to guide automated decisions, other applications, such as Cyber-Physical systems and autonomous driving, also necessitate formal methods to ensure verifiable systems and guarantee software security against cyber-attacks.

2.4 Process Mining

PM is a data-driven methodology that sits at the intersection of data mining, business process management (BPM), and computational intelligence. It aims to extract meaningful insights about business processes from event logs generated by enterprise systems, such as ERP, CRM, or workflow management systems. By leveraging real-world data, PM enables organisations to understand, analyse, and improve their processes in ways that traditional BPM approaches may not support [250].

The discipline of PM was pioneered by Wil van der Aalst and has grown significantly as more organisations have embraced digital transformation. Event logs collected by information

systems can reveal hidden details about how processes are executed in practice, often deviating from designed workflows or organisational guidelines. PM provides a suite of techniques for discovering these deviations, analysing process performance, and predicting future process behaviours based on historical data [250].

2.4.1 Event Log

The foundation of any PM activity is the event log. An event log consists of records of events, each representing a step in the execution of a business process. These records are typically captured automatically by enterprise information systems. An event log is not simply a list of actions; it is a structured dataset that contains specific attributes needed for PM tasks.

Each event in an event log is usually characterised by three key attributes:

- *Case ID*: a unique identifier representing a particular process instance (e.g., an individual purchase order or a customer service ticket).
- *Activity*: the name of the action or task executed (e.g., "Order Received", "Invoice Sent").
- *Timestamp*: the exact date/time the event occurred (e.g., 2023-08-19 12:17:55).

In addition to these, event logs can contain other information, such as the resource responsible for performing the activity (e.g., an employee or a system), as well as additional data attributes (e.g., cost, product ID). The structure and quality of an event log are critical for ensuring that PM algorithms work effectively. Poor quality logs with missing or inconsistent data can lead to unreliable results [134].

The proper preprocessing of event logs, including cleaning, filtering, and aggregating data, is a vital step in the PM pipeline. Without clean and well-structured event logs, process discovery and subsequent analysis would be flawed or incomplete [134].

2.4.2 Process Discovery

One of the most powerful and widely used techniques in PM is *process discovery*. Process discovery is the task of constructing a process model based on an event log without any prior knowledge of the process. Unlike traditional business process modelling approaches, which often rely on interviews, workshops, or manual documentation, process discovery is entirely data-driven [119]. The goal is to automatically derive an accurate model that describes the real execution of the process.

Various algorithms have been developed for process discovery, with each having its strengths and weaknesses:

- *Alpha Miner*: one of the earliest process discovery algorithms designed to create a Petri Net [247] representation of the process. While it works well for simple processes, it struggles with noise and complex control flow patterns [252].
- *Heuristics Miner*: this algorithm extends the Alpha Miner by incorporating heuristics to handle noise and infrequent behaviour. It is better suited for real-world event logs that contain deviations from the standard process [267].
- *Inductive Miner*: a more advanced algorithm that focuses on ensuring the process model is sound (i.e., free of deadlocks and other issues). It produces structured, hierarchical models and handles complex processes with greater accuracy [118].

Process discovery is a critical step in understanding how processes are executed in reality. The resulting models can be used to compare the designed process with the actual process, allowing organisations to identify inefficiencies, bottlenecks, and compliance violations [119].

2.4.3 Variant Analysis

Variant analysis in PM is the examination of different pathways, or "variants", that process instances may take. In any business process, especially those involving human decision-making or complex interactions, multiple variants of the process are likely to exist. Variant analysis allows process analysts to compare these variants to identify differences, inefficiencies, or potential improvements [24].

For instance, in a purchase order process, one variant may involve all steps being followed in sequence (e.g., "Order Placed" $\rightarrow$ "Order Confirmed" $\rightarrow$ "Payment Processed" $\rightarrow$ "Item Shipped"), while another variant may have skipped steps or out-of-sequence events due to exceptions (e.g., "Order Placed" $\rightarrow$ "Item Shipped" before "Payment Processed"). Through variant analysis, companies can detect deviations from the standard process, often caused by specific circumstances or operational inefficiencies [24].

By visualising and quantifying these variants, organisations can better understand the root causes of process variations, assess their impact, and prioritise corrective actions. This is especially useful in highly regulated industries, where compliance with standard procedures is essential.

2.4.4 Predictive Process Monitoring

As organisations collect more event data over time, *PPM* has become an increasingly valuable capability in PM. Unlike process discovery or conformance checking, which focuses on past events, PPM aims to forecast future events and outcomes based on historical data.

PPM leverages ML algorithms and statistical models to predict various aspects of ongoing process instances. For example:

- Will a particular case finish on time, or is it likely to face delays?
- What is the probability that a case will deviate from the standard process?
- What are the chances that a process will fail to meet certain performance indicators (e.g., customer satisfaction, cost targets)?

Various predictive techniques are applied in this context, ranging from traditional models like decision trees and support vector machines to more advanced approaches such as DL models (e.g., LSTMs). These models are trained on historical event logs, and learning patterns that correlate with specific outcomes (e.g., long cycle times, missed deadlines). Once trained, they can be applied to ongoing cases, providing real-time insights that help organisations take proactive measures to optimise process performance [231].

For example, in a customer service process, predictive models could alert managers if a particular case is likely to exceed its service level agreement (SLA) deadline, allowing them to intervene early. This capability can significantly improve operational efficiency and customer satisfaction by preventing issues before they occur.

2.4.5 Conclusion

PM is a powerful tool that helps organisations extract actionable insights from their data, particularly in complex and dynamic business environments. From constructing detailed process models through discovery to uncovering inefficiencies via variant analysis and even predicting future outcomes, PM enables organisations to make data-driven decisions that improve operational efficiency, compliance, and customer satisfaction. As the field continues to evolve, it is expected that PPM, in particular, will play a more significant role in enabling real-time optimisation of business processes, helping companies to stay agile and competitive in an increasingly data-driven world. PPM can be applied to tender management workflows to anticipate potential delays, non-compliance risks, or deviations from standard processes, offering actionable insights to procurement officers. This predictive capability enhances decision-making [38], optimizes resource allocation, and improves overall process efficiency in tender management.

2.5 Large Language Models

LLMs have revolutionised natural language processing (NLP) by demonstrating the ability to generate coherent text, answer complex questions, and even exhibit some level of reasoning across various domains. LLMs, such as GPT-3 / GPT-4, BERT, and T5, are trained on vast amounts of data, enabling them to perform a wide range of tasks with impressive fluency and versatility. These models are built on deep neural networks, typically using transformer architectures, which allow them to capture long-range dependencies and context in text. The rapid advancements in LLMs have opened new possibilities in language understanding, generation, and translation, making them foundational tools in modern AI applications [256].

2.5.1 Transformer Architecture

At the core of most LLMs lies the transformer architecture, introduced by Vaswani et al. [256]. Unlike traditional recurrent neural networks (RNNs), transformers rely on self-attention mechanisms, which allow the model to focus on different parts of the input sequence dynamically. This architecture significantly reduces training time and improves the model's ability to handle long sequences, making it ideal for large-scale language modelling tasks [52]. The self-attention mechanism is key to the model's ability to understand the relationships between words in a sentence, regardless of their positions, which enhances its performance in language understanding tasks.

2.5.2 Pre-training and Fine-tuning

LLMs are typically pre-trained on massive corpora of unlabelled text, allowing them to learn a general representation of language. Pre-training is often unsupervised, using tasks such as masked language modelling (MLM) in BERT or autoregressive language modelling in GPT [189]. After pre-training, these models can be fine-tuned for specific tasks such as text classification, sentiment analysis, or machine translation. Fine-tuning requires significantly less data than pre-training and enables the model to adapt its general knowledge to a particular domain [128].

2.5.3 Applications of LLMs

LLMs have proven effective in a wide variety of applications. One of the most prominent is in conversational AI, where models like GPT-3 are used to build chatbots capable of engaging in human-like dialogue [29]. They are also widely used in machine translation,

where models like mBERT and XLM have demonstrated improvements in cross-lingual understanding. Additionally, LLMs are employed in content generation, summarisation, and even code generation, where they help automate tasks in industries such as marketing, law, and software development [190].

2.5.4 Ethical Considerations

While LLMs offer numerous benefits, their deployment raises important ethical concerns. One major issue is the potential for these models to generate harmful content, including biased or inappropriate language [13]. The large datasets used in training can inadvertently encode societal biases, which the models may reproduce in their outputs. Moreover, the opaqueness of LLM decision-making processes poses challenges for explainability, making it difficult to determine how a model arrived at a particular result [25]. As LLMs become more integrated into real-world applications, addressing these ethical challenges becomes critical.

2.5.5 Future Directions

The future of LLMs is likely to involve even larger models and more sophisticated techniques for reducing their computational and environmental costs. Efficient LLMs are emerging, with research focusing on distillation, quantisation, and sparse attention mechanisms to reduce the size and energy requirements of these models [213]. There is also increasing interest in multi-modal LLMs, which can process and generate not only text but also images, video, and audio, expanding their potential applications beyond traditional text-based tasks [187]. Additionally, researchers are exploring ways to improve the interpretability and fairness of LLMs, ensuring that their outputs are transparent and unbiased [58].

2.5.6 Conclusion

LLMs represent a major leap forward in the field of natural language processing, providing unparalleled capabilities for language understanding and generation. However, their widespread use also brings significant challenges, particularly around ethics, explainability, and efficiency. As LLM research continues to evolve, addressing these concerns while pushing the boundaries of what these models can achieve will be critical for their responsible development and application.

Chapter 3

Machine Learning for Law: Predicting Complaints in Italian Tenders

3.1 Introduction

With the proliferation of e-procurement systems in the public sector, valuable and open information sources have become accessible. This research aims to investigate methods for improving the quality and accuracy of the public procurement process, enhancing the efficiency of administrations, reducing the time spent by economic operators, and lowering public administration costs.

The study focuses on exploring various legal Open Data, particularly analysing the dataset from the National Anti-Corruption Authority in Italy on public procurement and court rulings related to public procurement, published on the website of the Italian Administrative Justice from 2007 to 2022. The primary objective was to train ML models capable of automatically identifying which procurement processes resulted in disputes and subsequent complaints to the Administrative Justice, by recognising the relevant features of procurement that correspond to specific anomalies.

The remainder of this chapter is organized as follows: section 3.2 introduces the legal datasets used in this work. In section 3.3, a literature review about AI techniques applied in public procurements is presented, while in section 3.4.1, the objectives of this study are exposed; in section 3.4.3, the proposed methodology is described, while section 3.4.4 provides insights about the results of the research. Section 3.4.4 briefly discusses some explanations of the predictive outcomes. Finally, section 3.5 concludes the chapter.

High-resolution images for this chapter are available at https://bit.ly/4eRONLT. The source code for the experiments conducted in this chapter is available in the repository: https://github.com/roberto-nai/PhD-THESIS.

3.2 Dataset Overview

This section describes the resulting effort to build a main dataset, merging the national OD from 2016 to 2022. Following the advice of domain experts, the time frame is reasonably consistent with the validity of the Italian public procurement code, which came into force in April 2016 and was repealed in 2023[1]. The selection of the period also concerns reasons of both consistency, related to the abundance of complete data, and expediency, according to the legal experts' suggestions. The following sections describe the methods adopted, presenting the main indicators and possibilities for use. A full discussion of the data model obtained can be found in Appendix A.

3.2.1 Legal Data Sources

This research is based on two legal datasets involving the public procurement process in Italy. The first dataset was obtained from ANAC, an independent Italian administrative authority whose task is to prevent corruption in the Italian public administration, implement transparency and supervise public contracts[2]. ANAC collects data on calls for procurement from the public contract authority and provides a catalogue of OD describing public procurement, public authorities (public administrations which create the procurement), and economic operators (contractors who win the procurement). Currently, the ANAC website[3] provides data on approximately 3 million of public procurement collected from 2016 to 2022 whose amount is above 40 thousand euros.

The second dataset was obtained from the IAJ that contains the judges' sentences related to the public procurement complaints; currently, the IAJ website[4] provides about 67,850 sentences collected from 2007 to 2022.

The third and fourth datasets are obtained from the Italian National Institute of Statistics (ISTAT)[5] and the Open Database of Public Administration (BDAP)[6], which provide comprehensive data on Italy's economy, population, public finances, and administration.

[1] Legislative Decree n. 50 of 2016, replaced by the Legislative Decree n. 36 of 2023
[2] https://www.anticorruzione.it/en/mission-e-competenze
[3] https://dati.anticorruzione.it
[4] https://www.giustizia-amministrativa.it/web/guest/dcsnprr
[5] http://www.istat.it/en
[6] https://openbdap.rgs.mef.gov.it

These platforms enhance transparency and support informed decision-making by offering researchers, policymakers, and the public access to various statistical and financial information.

3.2.2 National Anti-Corruption Authority (ANAC)

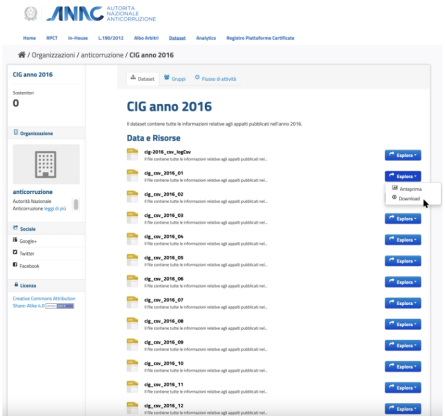
The main dataset concerns the government body responsible for collecting public tenders in Italy, ANAC. A specific section of the ANAC website[7] provides access to data in a standard view where data can be selected for categories and downloaded in compressed files (Figure 3.1.a). In the ANAC OD catalogue, the main dataset is the one related to the creation of a *Tender Notice*. Four other relevant datasets available include the list of the *Contracting Authorities* (CA) that have created a tender, the list of tenders that received an *Award*, the *Economic Operators* (EO) that awarded a tender, and the *activities* related to a tender after the awarding process (e.g. *contract-start*, *contract-end*, *subcontract*, etc.). A brief description is provided for each of these four other datasets. CAs are public bodies or entities that act on behalf of one or more public bodies and are responsible for acquiring goods, services or works through tendering procedures. These authorities are in charge of managing and supervising the entire public tender process. CAs can be of three types: Central (e.g. ministries), Regional and Local (e.g. municipalities), and other entities (e.g. hospitals). Awards contain the list of awarded tenders with the final amount awarded, the date of the award and the EOs that was awarded the tender. EOs can be sole enterprises, artisans, partnerships or capital companies, cooperatives, etc. Each tender and its related activities are identified by a 10-character alphanumeric feature called *CIG*. CAs and EOs are identified by their *tax code* (an alphanumerical string); CAs also have a unique *ISTAT code*. Figure 3.1.b details an excerpt of *Tender Notice* and *Award* where not all tenders are awarded to an EO (cells *DATE_AWARD* and *TENDER_AWARD* empty).

ANAC data overview

This section presents a short description of the most relevant features from three tables of ANAC, whereas their features are listed in Table 3.1.

In `TENDER_NOTICE` table, each tender is identified by an alphanumerical value called *CIG* (the key ID value), used to connect most of the remaining tables. The main distinction between tenders is their *type* and *sector*: types can be “Services” (S), “Supplies” (U) or “Works” (W), while sectors can be “Ordinary” (O) or “Extraordinary” (E) based on whether they are planned or due to extraordinary events (e.g. floods, earthquakes, etc.). The CPV code

[7] https://dati.anticorruzione.it/opendata



(a)

CIG	OBJECT	CONTRACTING AUTHORITY	DATE PUBLICATION	DATE AWARD	TENDER AMOUNT	TENDER AWARD	CITY	TENDER TYPE
5686540940	ATTREZZATURE SERVIZIO ANESTESIA E RIANIMAZIONE	AZIENDA UNITA' SANITARIA LOCALE N. 6 SANLURI	13/03/2014	27/03/2014	€48.980,00	€48.980,00	N/A	FORNITURE
997686731A	P23-216-TLP-LATRONICO-PZ-RIPRISTINO STATICO FUNZIONALE DEL SOLAIODI COPERTURA	3 REPARTO GENIO A.M.BARI	26/07/2023	N/A	€87.397,16	N/A	Bari	LAVORI
A040977869	POLIZZA ASSICURATIVA ALL RISKS BENI IMMOBILI E MOBILI 24 MESI + 12 MESI RINNOVO	COMUNE DI LONGARE	19/12/2023	N/A	€60.357,72	N/A	Vicenza	SERVIZI
A03FEA7C2E	CONVENZIONE PER LA FORNITURA MATERIALI DI PULIZIA, MONOUSO ED IGIENE PERSONALE PER I COMANDI/ENTI DELLA GIURISDIZIONE DEL COMANDO INTERREGIONALE MARITTIMO SUD E DELLE UU.NN. DIPENDENTI DALL'AREA OPERATIVA DI CINCNAV	DIREZIONE DI COMMISSARIATO MARINA MILITARE - TARANTO	19/12/2023	N/A	€215.000,00	N/A	Taranto	FORNITURE
0075606809	SS.SS. NN.33-336-341 E N.S.A. N.26	ANAS - SOCIETA' PER AZIONI	17/09/2007	N/A	€50.000,00	N/A	Roma	LAVORI
99124035B9	MESSA IN SICUREZZA DELL'AUDITORIUM SCUOLA 'ALDO MORO' DI VIA XXIV MAGGIO - OPERE EDILI	COMUNE DI CISLAGO	23/06/2023	18/07/2023	€129.089,87	€119.690,12	Varese	LAVORI
6861416B98	S.C. 20016693 CAVO 3X1X185 MM² ARE4(P1)H5EX ISOLATO IN XLPE (O HPTE) 12/20 KV, TIPO ASPIRALE VISIBILE, CONDUTTORI E SCHERMO IN ALLUMINIO, ISOLAMENTO ESTRUSOA SPESSORE RIDOTTO, GUAINA TERMOPLASTICA RESISTENTE ALL'URTO EALL'ABRASIONE, ALTRE CARATTERISTICHE SECONDO S.T. ACEA DISTRIBUZIONEDMX7, ED. LUGLIO 2016 (IMPORTO COMPRENSIVO DELL'OPZIONE DI ESTENSIONE DEL 20%)	ARETI S.P.A.	13/12/2016	23/12/2016	€285.465,60	€229.488,00	Roma	FORNITURE

(b)

Fig. 3.1 (a) A page of the ANAC website for downloading CSV files (by year and month); (b) CSV file data preview

describes all the types of tender, i.e. *Common Procurement Vocabulary*[8]. These categories are organized into an ontology (in a hierarchical organization) whose elements are identified by codes; using the first two digits of the codes (that correspond to the upper part of the ontology and the coarsest grain categories) they provide the CPV *divisions* useful to distinguish the product categories purchased by CAs (e.g. "90" represents cleaning services while "9040" sewer cleaning). A tender can be defined inside a *framework agreement*, meaning that the CA and EO have a previous agreement to provide services for further tenders for a defined duration (e.g., 1 - 5 years). Often, a tender is split into *lots*, with a lower amount. Finally, each tender has a well-defined *selection criterion* to choose the EO who will be awarded, and *implementation criterion*, which the winning EO will have to comply with.

The table `AWARDS` contains the list of relevant features related to the tender award, with the awarding entity, date, amount, etc. As expected, non-awarded tenders are not reported in this table (so they are only available in the `TENDER_NOTICE` table).

The table `CONTRACTING_AUTHORITIES` includes information about the name of the CA as well as the main keys to link to the other tables. In this respect, the *tax code* is used to know the type of CA by joining this table with `CONTRACTING_AUTHORITIES_BDAP` table from Registries, while *ISTAT code* is used to join this table with `ISTAT_DIMENSIONS` and

[8] https://simap.ted.europa.eu/web/simap/cpv

ISTAT_NUTS to investigate the geographical dimensions. A complete data model of the ANAC data integrated with the ISTAT (Section 3.2.4) and BDAP (Section 3.2.5) datasets is described in Appendix A.

Table 3.1 Main features of the tables TENDER_NOTICE (T_N), AWARD (AW), and CONTRACTING_AUTHORITIES (C_A)

Table	Feature	Description
T_N	CIG	PK: alphanumeric value
	Tender object	Textual summary of the tender
	Framework agreement between PA and EO	1 if yes, else 0
	Number of lots	Integer value {1..n}
	Tender type	Supplies (U) Works (W) Services (S)
	Tender area	Ordinary (O) Special (S)
	Tender amount	Float value
	Date of publication	Date in format yyyy-mm-dd
	EO selection mode	Integer value {1..122}
	Execution mode	Integer value {1..19}
	Region	Italian region names + Central Government
	CPV	String ID (XX000000-Y)
	CPV division code (first two digits of CPV)	String ID (XX)
	PNNR flag	1 if yes, else 0
AW	CIG_{FK} + AWARD_ID	PK: alphanumeric value
	EO consortium (group of EOs)	1 if it's a group of EOs, else 0 (individual)
	Award date	Date in format yyyy-mm-dd
	Awarded amount (bid amount)	Float value
	Awarded amount drop (bid drop)	Float value
	Number of bids admitted	Integer value {1..n}
	Subcontracting admitted	1 if yes, else 0
C_A	Tax Code	PK: alphanumeric value
	ISTAT Code	Alphanumeric value
	CA denomination	Textual string

ANAC data validation

This section presents a validation of the ANAC dataset to assess its technical quality. First, the dataset was examined to ensure completeness in essential fields, such as those listed in Table 3.1, since these fields are mandatory in the submission forms compiled by Contracting Authorities (CAs).

Further investigation was conducted on the issue of missing values, as several empty fields were identified in certain tables. In summary, two types of missing data were observed: (i) fields that are expected to be missing, as the corresponding values were not required, and (ii) fields with missing values likely due to operator error during data entry, which is a common issue in data quality. Both cases are explored in detail below.

i) *Expected missing data*. An in-depth analysis with domain experts confirmed that it is normal for certain fields to remain empty. For example, some entries, such as the PNRR code, were introduced later, meaning earlier notices could not include this information. In the main table `TENDER_NOTICE`, the PNRR flag was missing in 76% of the records. Discussions with domain experts confirmed that these values were correctly missing, as the PNRR only commenced in 2022. In the `AWARDS` table, two features with a high proportion of missing values were also correctly identified as empty. These include the minimum and maximum discount offered compared to the tender amount (55%) and the award criterion. According to domain experts, the missing discount values can be considered unapplied.

ii) *Omission in data entry*. In some minority cases, the missing fields are likely due to omissions by the operators responsible for inputting the data into the system. For instance, the ISTAT code for geographic aggregation in the `CONTRACTING_AUTHORITIES` table shows 4% missing values. These missing values can be reconstructed by linking other fields within the same dataset, such as using the location name in place of the corresponding primary key (PK).

3.2.3 Italian Administrative Justice

The IAJ is a textual data set containing the judges' sentences saved in HTML format, DOC/DOCX, and PDF files. In addition to the texts, the sentence files contain some useful metadata: the European Case Law Identifier (ECLI) code[9] of the sentence, the court region (that corresponds to the region of the public authority that created the tender), the year and the progressive number of the judge's sentence. Thanks to the ECLI code, it is possible to trace the metadata of complaints related to the sentences: the complaint object, the year, and the progressive number (from which the complaint started). Figure 3.2 presents a screenshot

[9] https://e-justice.europa.eu/content_european_case_law_identifier_ecli-175-en.do

of the IAJ website where the texts and metadata of the complaints are available. For this dataset, all fields in the metadata were complete, with no missing information. Similarly, all the associated texts were fully available and contained no gaps or omissions. This ensures the dataset is both comprehensive and reliable in terms of metadata and content.

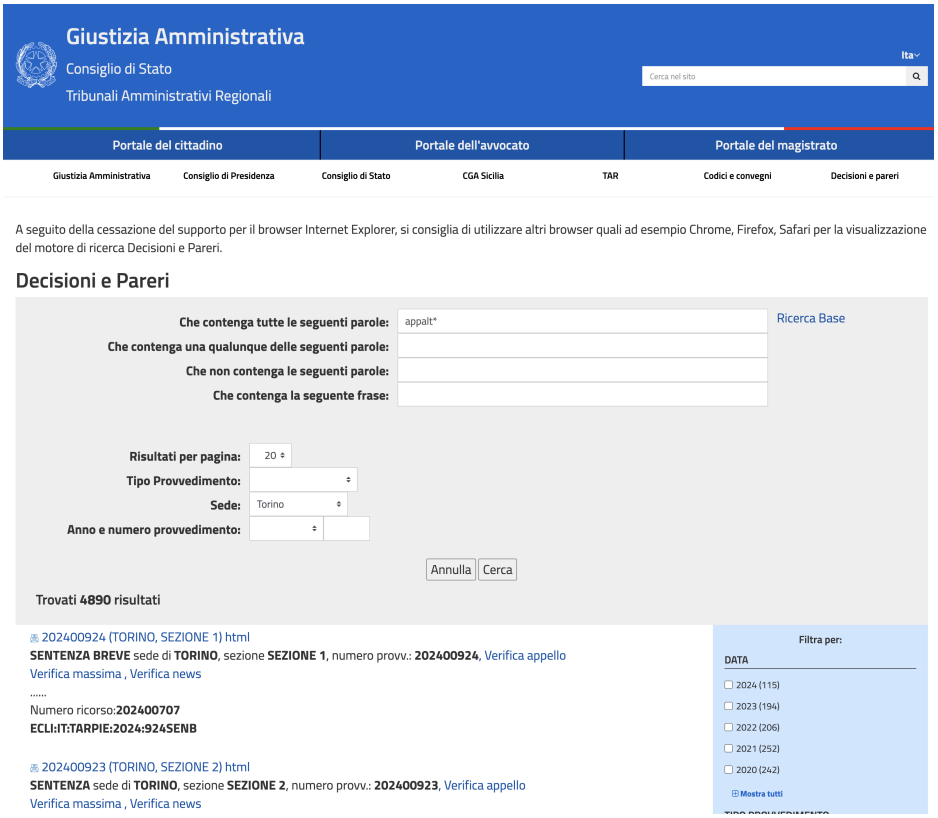


Fig. 3.2 A page of the IAJ website with complain metadata and text download

3.2.4 Italian National Institute of Statistics

A relevant aspect concerns identifying the scope of each tender's administrative aggregation. ISTAT provides a wide range of statistical information concerning Italy; among them, it's possible to find the *Nomenclature of Territorial Units for Statistics* (NUTS)[10] and the distribution of *population per municipality*. The NUTS system is organized into three hierarchical levels: NUTS 1 includes socio-economic regions, such as large economic areas; NUTS 2 concerns a smaller region for applying regional policies, like provinces or large metropolitan areas; NUTS 3 involves the smallest areas, such as regions, provinces and

[10]https://www.europarl.europa.eu/factsheets/en/sheet/99/nomenclatura-comune-delle-unita-territoriali-statistiche-nuts-

municipalities[11]. The distribution of inhabitants can be of interest for understanding, for example, the quote of investment per population in a particular area[12]. Including NUTS and population facilitates, for instance, the comparison and analysis of territorial investments at the region/province/municipality level, whereas NUTS and population are identified by the *ISTAT code* (an alphanumerical string) of the corresponding municipality.

3.2.5 Database of Public Administrations

The BDAP includes relevant information on several aspects of public administrations, such as the organizational structure (e.g. municipality, school/university, hospital, etc.), geographical coverage (North, Central or South Italy), and budgetary information. Such information aims to enhance transparency, improve operational efficiency, and support policy development. Including DBAP facilitates, for instance, comparison and analysis of investments by type of authority; categorizations described above are identified by the *tax code* (an alphanumerical string) of the municipality they refer to.

3.3 Public Tenders Fraud Detection and Artificial Intelligence Techniques: a Literature Review

Organizations are increasingly focused on mitigating the chances of experiencing fraud, which represent a significant loss of revenue. For instance, an accredited biennial 2020 study carried out by the Association of Certified Fraud Examiners claims that on average 5% of a company's revenue is lost because of unchecked fraud every year. Among the reasons for these large losses is that it takes about 14 months for a fraud to be discovered and that audits capture only 3 percent of actual fraud. This necessitates the use of better tools and processes to quickly and inexpensively identify potential criminals [2].

Researches in political science, economics and sociology investigated the field, trying to highlight possible flaws in the systems that lead to such risks, with a view to prevention. Recently, the new possibilities offered by information technology allow for new studies in the area of fraud detection as well. Recent changes include the availability of large data sets at low cost, the use of increasingly powerful computing devices and the development of applications that enable the training of ML models [254, 148].

Public organizations are also subject to fraud risks, starting with public procurement. A major challenge is to be able to detect potential fraud automatically, through appropriate

[11] https://www.istat.it/it/archivio/6789
[12] https://www.istat.it/it/archivio/156224

artificial intelligence (AI) techniques. ML methods, in particular, have proven very effective in a wide range of practical applications [274, 227, 228]. In addition, the most recent methodologies have also developed Natural Language Processing (NLP) [20] techniques, as well as neural networks [142].

A systematic review of the research literature is proposed to systematise the existing research on AI techniques applied to fraud detection in public procurement. The objectives are summarised by exploring the following three research questions:

- *RQ1: Which disciplinary areas are more interested in investigating fraud in public procurement?*
- *RQ2: What AI techniques are being applied to investigate fraud in public procurement contracts?*
- *RQ3: Which research studies are most influential in the field?*

The remainder of the section is organised as follows: section 3.3.1 introduces related works. Section 3.3.2 describes the proposed methodology, while section 3.3.3 presents insights from the review results. Finally, section 3.3.4 summarises the main research, and section 3.3.5 concludes the section.

3.3.1 Related works

There are previous works focused on the application of techniques to public procurement. In [235], 102 articles published between 2015 and 2019 have been selected from Scopus and WoS databases, focusing on the primary data mining techniques used to prevent corruption. It is observed that the main techniques of AI are those based on the theorem of Bayes, neural networks, Support Vector Machines (SVM), decision trees, Random Forest, logistic and linear regression.

In another recent survey, 147 articles published between 2015 and 2019 have been selected from Scopus and WoS [237]. These works focused the following types of corruption: fraud (77.49%), overpricing (7.05%), bribery (5.05%) and favouritism (4.66%) generate greater citations in the articles. A large part of research analyses business intelligence literature for bank fraud using text mining. For the geographical distribution of the authors, the first author of the publication was considered; the leading countries with articles on data mining and corruption are the United States (16,3%), China (10,9%) and the United Kingdom (8,9%).

Another work focused on the most used methods to detect different types of "corruption" [141], exploring 23 articles published between 2016 and 2021. Data mining and ML methods are used in this segment over a large amount of data collected from different data

sets, such as contract registers, blacklist economic operators, business registers and so on. The methods include classification techniques, with the aim of detecting connections between economic operators and contracting authorities, but also for finding companies that participated in collusion, as well as associations rules, and graph databases algorithms.

Another recent review is focused on Social Network Analysis (SNA) to capture the contributions of the scientific community to the topic of corruption in public procurement [131]. Authors identified the most recurrent authors, their interactions, number of citations, identification of keywords, and their repetitions. Authors analyzed 18 articles from 2011 to 2021. To perform network analysis on the collected dataset and represent the interactions between the actors or nodes of the graph, the open-source engine VOSviewer[13] was used; the tool allowed the authors to identify the publications, authors, journals, institutions, keywords, and countries with the most significant impact on the research in repositories of scientific articles (from parameters such as centrality degree and edge weight).

This work builds on previous research by focusing on recent studies (2016-2021) related to fraud detection using ML methods and techniques.

3.3.2 Methodology

For this survey, the methodology outlined in [238] is followed, dividing the process into three phases: planning, conducting the review, and reporting the review. The workflow in figure 3.3 resumes the activity of this research.

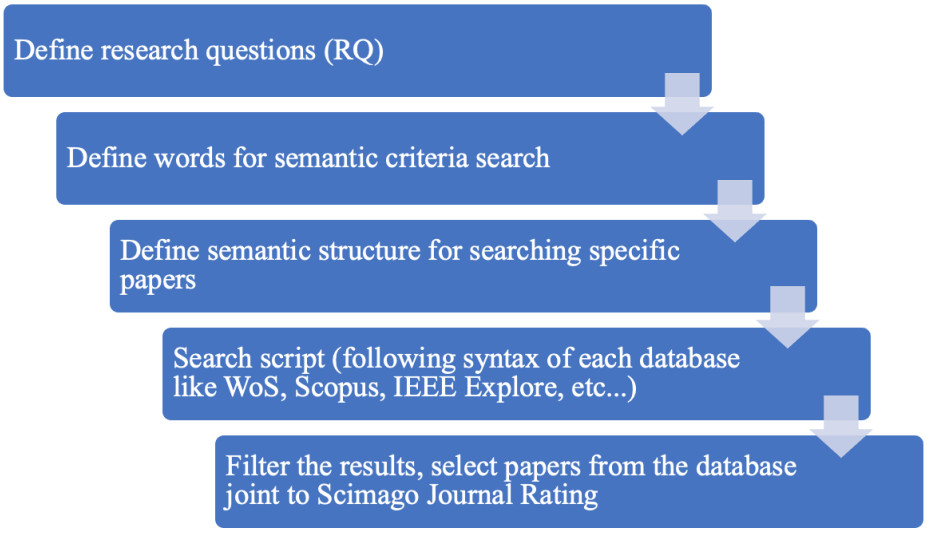


Fig. 3.3 Workflow for the planning and conducting the review

[13] https://www.vosviewer.com

Table 3.2 describes the objective of the survey through the research questions and the expected output variables.

Table 3.2 Key research questions addressed in the study, focusing on disciplinary areas (RQ1), AI techniques (RQ2), and influential research (RQ3)

Question	Type of answer sought
RQ1: Which disciplinary areas are more interested in investigating frauds in public procurement?	List of disciplinary areas
RQ2: What AI techniques are being applied to investigate fraud in public procurement contracts?	List of AI techniques
RQ3: Which research studies are most influential in the field?	Weighted graph of citations between articles

Semantic structure of the search

In the first phase of the workflow, the following keywords have been defined: public tenders, public competitions, public procurement, e-procurement, state laws, fraud detection, corruption, crime, criminal, prediction, predictive, modelling, detection, artificial intelligence, machine learning, deep learning, neural networks. In the second phase, various combinations of the keywords have been prepared as queries for the scientific databases (Scopus, WoS and IEEE Explore). In the third phase, specific scripts have been prepared following the syntax of each database. For instance, one of the queries in Scopus (very similar in WoS) has been:

```
TITLE-ABS-KEY ((("PUBLIC TENDER" OR "PUBLIC PROCUREMENT"
OR "E-PROCUREMENT" OR "public competitions"
OR "public regulations" OR "state laws")
AND ("DETECTION" OR fraud OR corruption OR crime OR criminal)
AND (prediction OR predictive OR "machine learning"
OR "deep learning" OR "neural networks" OR "modeling"
OR "artificial intelligence" )))
```

Inclusion and exclusion criteria

The results have been filtered by categories: Computer Science, Engineering, Business, Politics and Business Management, joint with Scimago Journal Rating (Quartile and H-Index). The search has been automated in Python, combining the Scopus APIs, WoS and

IEEE Explore to automatically get the list of results joined with Scimago rating indexes[14]. Finally, 15 studies were selected for this survey.

3.3.3 Results

The main results of the survey are summarised by examining certain features of interest for each paper. The disciplines involved, the dataset used in input (Input), the AI techniques, methodologies and technologies adopted (Methods).

Table 3.3 and table 3.4 resume the results about RQ1. Disciplines involved are mostly Computer Science, but also Business, Management and Accounting or Engineering, from different venues, e.g. Conferences or Journal in Quartile 1 or 2 (Q). Regarding the geographical distribution of the authors, all the authors are considered. Most researchers come from Europe (Spain, Portugal, Italy) and America (Brazil, U.S.A., Paraguay).

Finally, the research summarized in Table 3.5 and Table 3.6 has been selected with respect to the RQ2 research question. Most frequent methods concern typical ML supervised and unsupervised algorithms. Regarding technologies, Python emerges as the most used programming language (libraries such as Scikit-learn are becoming the state-of-the-art in the subject), but also Java or R. Among the tools used, Neo4J and KNIME are notable. Finally, some works explore the adoption of neural networks, as well as social network analysis methods.

A co-occurrence network has been generated in figure 3.4, starting from the main keyword *public procurement*. By applying the community detection algorithm (Louvain) [183], the densest groups of terms are detected. In particular, the first group includes terms related to corruption, risk, and governance with respect to public administration, public spending, public sector. A second group includes terms about ICT technologies (blockchain, e-procurement, information and communication, data processing). Interestingly, a third group includes Artificial Intelligence, data mining, and social networks (and fraud detection). Fourth, another group of terms includes network methods: semantic web, knowledge graphs, semantic technologies (and anomaly detection). These results confirm the effectiveness of the approach, and these categories will be used in the presentation of the research in section3.3.4.

[14]Python script using Scopus APIs and Scimago Journal Rating is available here: https://github.com/roberto-nai-unito/scopus-api

Table 3.3 Disciplinary area of the selected papers with citation count (#Cit), the Venue, and the Journal's Quartile (JQ)

Paper	#Cit	Venue	Disciplines	JQ
[236]	-	Electronics (Switzerland)	Computer Science Engineering	Q2
[76]	3	International Journal of Forecasting	Business and International Management	Q1
[257]	7	International Transactions in Operational Research	Business Management and Accounting Computer Science Decision Sciences	Q1
[122]	-	Conference paper	-	-
[177]	3	Conference paper	-	-
[68]	3	Governance	Public Administration	Q1
[81]	-	Automation in Construction	Engineering	Q1
[142]	2	Conference paper	-	-
[185]	7	Proceedings	Computer Science	-
[162]	1	Proceedings	Computer Science	-
[35]	1	Proceedings	Computer Science	-
[80]	8	Complexity	Computer Science	Q1
[179]	4	Business and Politics	Business, Management and Accounting	Q1
[53]	5	Conference paper	-	-
[49]	10	SSRN Electronic Journal	-	-

3.3.4 Summary of main research

The main results of the papers can be summarised by grouping them into three classes. Most of the papers adopt typical ML methods, while two smaller groups deal mainly with neural networks and network analysis.

Typical machine learning methods. In [236], a multi-phase model was used (the identification of anomalies and generation of the detection model), which uses different algorithms, such as clustering (K-Means), Self-Organizing map (SOM), Support Vector Machine (SVM) and Principal Component Analysis (PCA). Following this methodology, a semi-supervised learning model is built for the detection of anomalies, which obtains an accuracy of 95%, allowing the detection of procedures where the aim is to benefit a particular supplier by means of the qualification assignment parameters.

Two ML models have been used in [76], to predict whether a contract will result in malfeasance, breach of contract, or inefficiency: a lasso classification model [233] and a gradient boosting classification model [75]. The methods used allow to describe which

Table 3.4 Geographical distribution of the authors contributing to the study, indicating the number of authors from each country

Country	Count
Spain	12
Brazil	11
United States	8
Portugal	5
Italy	4
Paraguay	3
Croatia	2
Australia, Austria, Colombia, Slovenia, United Kingdom	1

variables —and in which way these variables— contribute to the likelihood that a contract will be problematic, which is very useful from the perspective of policymakers; for instance, variables associated with projects such as their size or duration were important predictors of malfeasance. Also, the time lag between adjudicating the contract and the nearest election showed high predictive value.

Alternative predictive models were estimated in [68]; the article traces the organization of corruption in public procurement, by theoretically and empirically assessing the contribution of Extra-legal Governance Organizations (EGO) to supporting it. They used traditional regression and supervised machine-learning methods for identifying and validating proxy indicators for EGO presence in public procurement, such as single bidding or municipal spending concentration. The predictive models included both traditional regression analysis and ML: binary logistic regression, random forest, and Gradient Boosting Machines (GBM). Testing prediction accuracy on unseen data, GBM achieves 85%. Looking at external validity, the model's predicted EGO score also significantly and moderately strongly correlates with established indicators of organized criminality both within Italy and across Europe.

The accuracy of eleven ML algorithms for detecting collusion using collusive datasets obtained from Brazil, Italy, Japan, Switzerland and the United States is tested in [81]; while the use of ML in public procurement remains largely unexplored, its potential use to identify collusion is promising. The three top-performing ML algorithms have been the Extra Trees, Random Forest and Ada Boost (ensemble methods). In the scenario where all auction information was available, these algorithms' accuracy (detection rates) ranged between 81% and 95%, with a balanced accuracy generally above 73% (excluding the US dataset).

In [177], a prototype called SALER is proposed. Inside SALER, several internal and external data sources are analysed and assessed to explore possible irregularities in budget and cash management, public service accounts, salaries, disbursement, grants, subsidies, etc.

Table 3.5 Papers selected for the literature review, listing the input data sources and the methods and technologies used in the analysis (part 1/2)

Paper	Input data	Methods & Technologies
[236]	Public Procurement System (SERCOP) of Ecuador	Clustering (K-Means), Self-Organizing map (SOM), Support Vector Machine (SVM), Principal Component Analysis (PCA). Technologies: Python Scikit-learn library, MiniSom, AZURE ML.
[76]	Sistema Electronico de Contratación Pública (SECOP) of Columbia	Lasso classification model, gradient boosting classification model (GBM). Technologies: n.a.
[257]	Various from the states of Brazil	Graph theory, network analysis, clusterization, regression analysis. Technologies: n.a.
[122]	Diario Oficial da Uniao (DOU) of Brazil	Bottleneck deep neural network and Bi-LSTM. Technologies: n.a.
[177]	Public procurement open data from Spain	ML, pattern detection. Technologies: Python, R, Neo4j.
[68]	Italian dataset managed by the ANAC	Binary logistic regression, random Forest, and Gradient Boosting Machines (GBM). Technologies: R.

SALER combines descriptive and predictive ML models and the results can be accessed with a web interface. Finally, the authors mention two frameworks similar to SALER: zIndex[15], a public procurement benchmarking tool for rating contracting authorities which is being developed in the Czech Republic by researchers from the Charles University of Prague and Arachne[16], considered by the European Commission as a good tool amongst anti-fraud measures; this risk-scoring tool generates more than 100 risk indicators sorted into specific risk categories to help managing authorities and intermediate bodies to prevent and detect errors and irregularities among projects, beneficiaries, contracts and contractors.

The relation between the award price and the bidding price is investigated by [80]. It proposes an award price estimator that uses the random forest [27] regression method over the Spanish open data from 2012 to 2018. Finally, a similar analysis, employing a dataset from European countries (TED[17]), is presented to compare and generalise the results. The

[15] https://www.zindex.cz

[16] https://ec.europa.eu/social/main.jsp?catId=325&intPageId=3587&langId=en

[17] https://data.europa.eu/data/datasets/ted-csv?locale=en

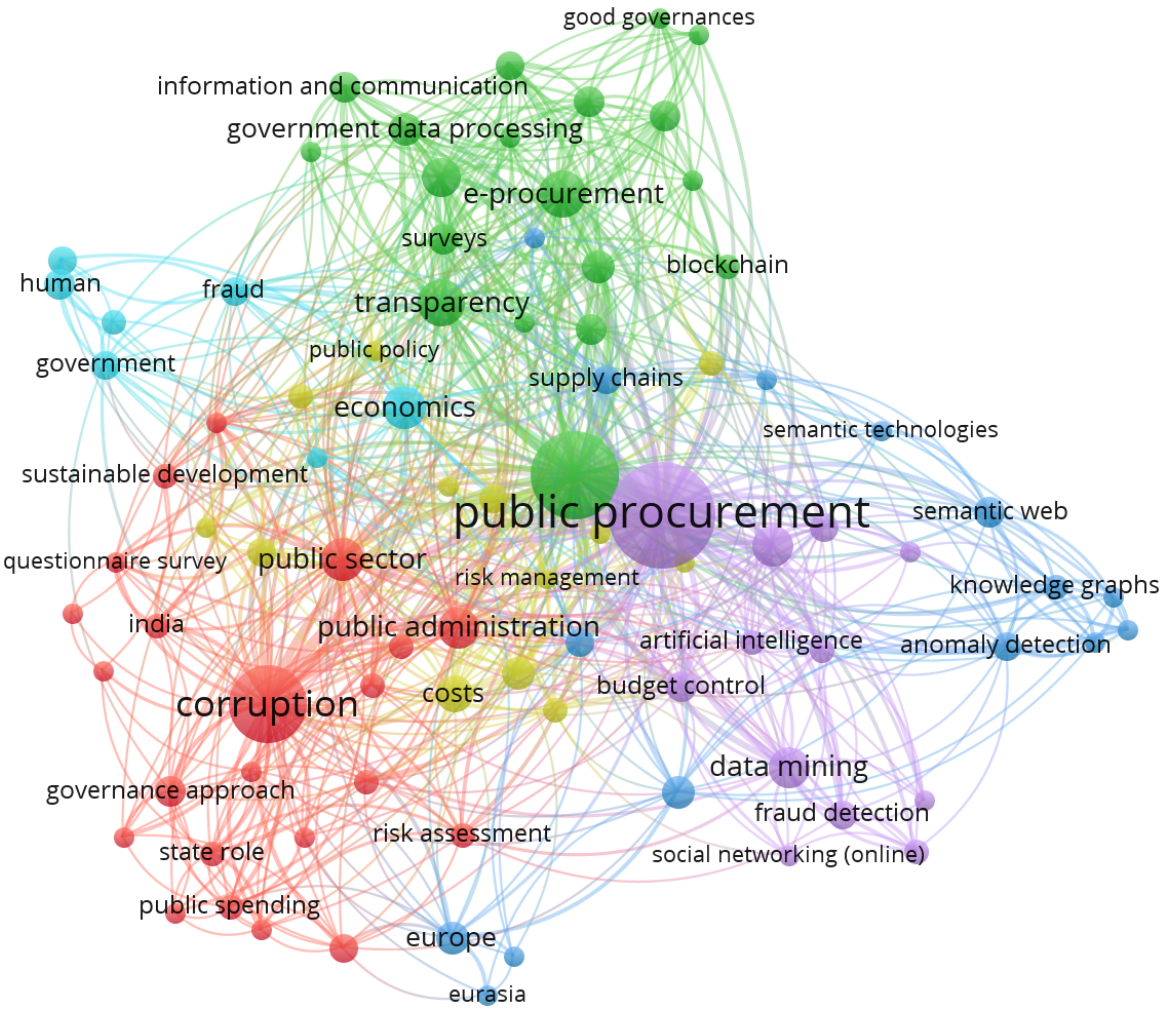


Fig. 3.4 Visualization of a term co-occurrence network

article illustrates how a ML algorithm can be useful. Particularly, random forest predicts the award prices with less uncertainty, adapting to the real market.

ML tools are used to analyze by [179] to analyze a large dataset of public contracts from across Europe, in order to identify the conditions under which close connections, defined both in terms of repeated interaction, as well as geographical dispersion, appear. In this case, random forest models have been used.

In [49], three main results through detailed data on the content of calls for tenders involving roadwork contracts in Italy are presented. The prediction capability of the various corruption indicators using standard ML algorithms has been tested: lasso, ridge regression, and random forest. The article shows that, among ML methods, the random forests algorithm provides the most accurate prediction. At a more general level, the article suggests that a

higher standardization of call for tenders documents can contribute to reduce corruption risks. For this purpose, sector authorities or specialized public bodies can play a crucial role.

In [162] the initial results of an anomaly detection experiment by applying the Isolation Forrest algorithm to a publicly available dataset, i.e. the public procurement of Paraguay, are discussed. An in-depth study of the diversity of ties between buyers and sellers in public contracts adopted a statistical analysis with Random Forest models starting with 3.3 million European Union contracts between 2009 and 2015. The effectiveness of the model is validated with local known anomalous procurement processes, which are: a) processes protested by entities involved in the contracting process, which were determined in favour of the protestant, and b) complaints about the contracting process from external entities with the possibility of anonymity. The results show an accuracy of over 90% in detecting these known anomalies as early as in the tender stage and during the contracting stage.

Network analysis and text mining. A Decision Support System (DSS) is proposed in [257] to allow law enforcement agencies to establish priorities concerning the companies to be investigated. This DSS incorporates data mining algorithms for quantifying dozens of corruption risk patterns for all public contractors inside a specific jurisdiction, leading to improvements in the quality of public spending and the identification of more cases of fraud. These algorithms combine operations research tools such as graph theory, clusterization, and regression analysis with advanced data science methods to allow the identification of the main risk patterns. Starting from various datasets and social network analysis (graph model-based), an unsupervised learning model has been developed for clustering fraudulent employees by [53].

In [142] and [185], the use of advanced text mining to improve the procurement process is explored. Based on Public Procurement of Croatia[18]. The authors introduce the use of NLP to improve the research of frauds, comparing common classification algorithms: Naïve Bayes (NB), Logistic regression (LR) and Support Vector Machines algorithm (SVM). The models have been trained and tested on all data, and by groups of procurement lots (food, medical equipment, construction, IT services, etc.) defined in the unique Public Procurement Dictionary (CPV). Groups such as IT services, repair and maintenance services, and health and social work services have good prediction results; conversely, groups such as architecture, construction, engineering and inspection services provide bad metrics, precisely because of the lack of information on technical and professional abilities.

Neural Networks. The types of fraud investigated by [122] are mainly collusion (bid-rigging), overpricing, and delivery fraud (quality and quantity of services and materials). To evaluate the reference dataset, bottleneck Deep Neural Networks and Bidirectional

[18]https://eojn.nn.hr

Neural Networks [214] were chosen. Deep neural network models were built using the Tensorflow [1]. Both bottleneck deep neural networks and Bi-LSTM proved to be competitive with traditional classifiers and achieved better precision, which is more desirable (over recall) in a criminal fraud investigation. In [35], starting from the Portuguese Public Procurement portal, a graph-oriented user interface is proposed to support decision-making, using Cypher queries. Besides this, supervised ML methods are used to find suspicious procurement.

Final remarks. The methods employed in this review are chosen for their complementary strengths and alignment with the research objectives. However, a comparative analysis of their advantages and limitations can further contextualise their selection and performance.

RF and GBM are ensemble methods that combine multiple decision trees to improve accuracy and robustness. RF excels in handling structured data and provides high predictive performance by reducing overfitting through randomisation [28]. GBM, on the other hand, builds trees sequentially to correct errors, offering even higher accuracy in many scenarios [75]. However, both methods are often criticised for their lack of interpretability, which poses challenges in fields like law and public procurement, where explainability is crucial.

LR, a simpler linear model, is widely used for its transparency and ease of interpretation [234]. It allows stakeholders to understand the contribution of each feature to the prediction. Nonetheless, LR struggles with complex, non-linear relationships in the data, which limits its applicability to more intricate problems.

Neural models, such as Bi-LSTM, are particularly effective for textual data due to their ability to capture long-term dependencies and context [99]. These models outperform traditional methods in tasks involving NLP but require large datasets and substantial computational power. Additionally, their black-box nature often makes their predictions less interpretable, which may not align with the needs of regulatory and legal environments.

Future research could explore hybrid approaches, combining the interpretability of simpler models like LR with the predictive power of ensemble or neural models.

A common issue, instead, is the need for a substantial amount of reliable historical data, some of which (especially the collusion-related) may not always be made available by competition commissions or law enforcement agencies [81].

3.3.5 Conclusions and future work

A review of the most recent studies on fraud detection for public organisations was conducted, identifying typical methods based on ML algorithms, with emerging interest in NN and XAI techniques.

As future work, the goal is to identify relevant authors in the field, following the approach outlined in [174]. Specifically, a bibliometric network analysis (both graph-based and

timeline-based) will be performed to examine the centrality and density of authors connected to the topic of ML and fraud detection in public procurement. The proposal to use bibliometric network analysis for identifying authors in the field is of significant practical relevance. This method allows the identification of key contributors, emerging research clusters, and pivotal topics, offering strategic insights for future research. For instance, [262] demonstrates the utility of co-authorship networks in mapping collaboration patterns, while [253] highlights how keyword co-occurrence networks can reveal research trends and gaps. These techniques are particularly valuable in interdisciplinary fields like public procurement and fraud detection, where understanding the intellectual landscape can guide impactful studies and foster collaborations.

3.4 ML-Based Detection of Anomalies in Public Procurement

Based on the datasets from the ANAC and IAJ described in the previous Section, an approach was developed to utilise the information from these sources to pursue one of the objectives of this research. The main goal was to train ML algorithms to automatically identify which procurement processes have led to disputes and subsequent complaints to Administrative Justice. By recognising the key features of procurement activities associated with specific anomalies, the aim was to enhance the monitoring and prevention processes in the field of public procurement.

In addition to this primary goal, the research sought to explore various ML techniques and investigate their applicability in this context. Particular attention was given to the explainability of the obtained models, as understanding the reasoning behind the predictions is crucial in legal and regulatory environments. Ensuring that the models perform well and offer transparent and interpretable insights was a key consideration, providing valuable support to decision-makers in identifying problematic procurement patterns.

Building on these premises, it becomes possible to formulate and explore the answers to the following research questions, which aim to address the core challenges and objectives outlined in this section:

- *RQ1: What methods and approaches can be utilised to effectively integrate various legal datasets into a single, comprehensive labelled dataset, ensuring consistency and relevance of the information?*
- *RQ2: To what extent is it feasible to design an experimental framework that predicts potential complaints to administrative courts based on the specific features and characteristics of public procurement processes?*

In line with the research questions outlined above, the first step involved establishing a connection between the two datasets to label cases as either involving complaints or not. Following this, ML algorithms were applied to assess the classifiers' performance and identify the features that greatly influence classification outcomes.

3.4.1 Related work

Two ML models have been used in [77] to predict whether a contract will result in malfeasance, breach of contract, or inefficiency: a Lasso classification model [233] and an eXtreme Gradient Boosting (XGB) classification model [75].

The study in [186] explores using advanced text mining to improve the procurement process. It is based on the Public Procurement of Croatia. The authors introduce the use of NLP to improve the research of frauds, comparing common classification algorithms: NB, LR, SVM.

The authors of [47] present three main results through detailed data on procurement content involving roadwork contracts in Italy. The prediction capability of the various corruption indicators using standard ML algorithms has been tested: Lasso, Ridge Regression, and RF.

The relation between the award and bidding prices is investigated by [79]. An award price estimator is proposed using the RF regression method [28] over the Spanish open data from 2012 to 2018.

The accuracy of eleven ML algorithms for detecting collusion using collusive data sets obtained from Brazil, Italy, Japan, Switzerland, and the United States is tested in [203].

The authors of [178] used ML tools to analyze a large data set of public contracts from across Europe to identify the conditions under which close connections among public administrations and economic operators appear, defined in terms of repeated interaction and geographical dispersion. In this case, RF models were used.

The authors of [264] used SVM and LR to find out the relationship between fraud risk, competition, and performance monitoring using the American SAM (System for Award Management) and FCMD (Federal Contractor Misconduct Database).

Alternative predictive models have been estimated in [68]; the article traces the organization of corruption in public procurement by theoretically and empirically assessing the contribution of Extra-legal Governance Organizations (EGO). They used traditional regression and supervised machine-learning methods for identifying and validating proxy indicators for EGO presence in public procurement, such as single bidding or municipal spending concentration.

In [168], the use of ML methods has been explored, specifically neural networks, to predict public procurement contract outcomes. A Python application was developed to classify public contracts in the pipe industry, identifying high-risk contracts with high non-performance risks.

In [170], a Neuro-Fuzzy neural network is presented for evaluating and predicting the success of a construction company in public procurement.

3.4.2 The objectives and rationale of this study

Building on previous research, which has demonstrated the applicability of ML and algorithms such as RF, XGB, and others to public data from local governments, this study seeks

to address a gap in the literature by evaluating ML models beyond the typical focus on detecting corruption or estimating appropriate award prices. If this system recognises that a public administration has faced a complaint due to a tender, future contracts could be at risk of being halted by Administrative Justice, potentially leading to increased costs and delays for both the CAs and the EOs, resulting in significant economic losses. To achieve this, the research proposes merging two legal datasets to extract the necessary knowledge, aiming to predict whether a public administration launching a tender or awarding a contract is at risk of receiving a complaint before the Administrative Justice.

3.4.3 Methodology

In response to RQ1, the task of joining the ANAC and IAJ datasets is complex, as the IAJ database consists of textual documents that may reference ANAC procurement in various ways: through the procurement identifier (CIG), the denomination of CAs, the Economic Operators EOs, the region, and the year of the court decision.

Regarding RQ2, the problem is defined as follows: to develop an ML model capable of predicting, with the highest expected accuracy, the likelihood of a complaint being filed with the IAJ courts based on the characteristics (features) of a tender and its corresponding label (positive/negative case).

Merge of ANAC and IAJ datasets

First, the extracted texts from the sentence files were indexed using specialised IR tools. *Lucene* [137] and its expansions, *Solr* [89] and *ElasticSearch* (ES) [59], represent the major open-source IR toolkits used in industry [10]. According to the DB-Engines Ranking of Search Engines[19] as of September 2024, ES is the most popular search engine software. Based on this indication, the texts and metadata of complaints and sentences were extracted from the documents and inserted into Newline Delimited JSON (NDJSON) files[20], one of the input formats accepted by ES for data indexing. The NDJSON schema, where the complaint and sentence texts have been serialized, is illustrated in Figure 3.5; these fields represent metadata and text relevant to legal cases, enabling efficient filtering and searching of court-related information within the ES index[21]. The mix of *keyword*, *text*, *date*, and

[19]https://db-engines.com/en/ranking/search+engine

[20]NDJSON is a convenient format for storing or streaming structured data that may be processed one record at a time; it is a format suitable for data exchange in software client/server applications. See also: https://dataprotocols.org/ndjson

[21]In ES, an *index* is a collection of documents (individual units of data) that share a similar structure, used to store, organise, and enable fast search and retrieval of data.

short types ensures that different kinds of data, from full-text to exact matches, are handled appropriately. Deep learning techniques were explored to improve the connection between

```
{
   "properties":
   {
        "court": { "type": "keyword" }, "ecli": { "type": "keyword" },
        "sentence_file": { "type": "keyword" }, "complaint_id": { "type": "keyword" },
        "complaint_num": { "type": "keyword" }, "complaint_year": { "type": "short" },
        "complaint_date": { "type": "date" }, "sentence_id": { "type": "keyword" },
        "sentence_num": { "type": "keyword" }, "sentence_year": { "type": "short" },
        "text_recurring": { "type": "text" }, "text_resistant_pa": { "type": "text" },
        "text_resistant_winner": { "type": "text" }, "sentence_title": { "type": "keyword" },
        "sentence_year": { "type": "short" }
   }
}
```

Fig. 3.5 JSON that represents the schema for an ES index used to store metadata and the text of complaints and related sentences. The schema defines various fields, each with specific types and properties, determining how the data is indexed and searchable within ES

the procurement data from the ANAC dataset and the sentences from IAJ. NLP techniques were adopted using the *LaBSE BERT* model [71] to create sentence embeddings of the *procurement object* from the ANAC `TENDER_NOTICE` table (Table 3.1).

Sentence embedding is a collection of techniques based on artificial neural networks that transform textual sentences into vectors of real numbers, allowing distance calculations. These real numbers represent the probabilities of words appearing in the sentence. The probabilistic nature of sentence embeddings explains their effectiveness when measuring the distances between vectors representing sentences, with lower vector distances indicating greater similarity in meaning.

The same technique was applied to the *complaint object* of sentences published by IAJ. Cosine similarity [210] and TF-IDF fuzzy matching [230] were then applied to sentence embeddings to collect the corresponding similar subjects of the *procurement object* and the *complaint object* together. Although ES reduces match results via the BM25 [201] score algorithm, exact matches in ES can be obtained using the *match phrase* query, which only returns documents that precisely match the searched phrase (this is stricter than a match query using the AND operator). This query type was used to avoid fuzzy matches. For the NLP phase, pairs of subjects procurement object, complaint object whose similarity and TF-IDF fuzzy matching exceeded 0.90 were considered.

When the match between the entries of the two datasets was successful (via IR or NLP), the presence of a complaint on procurement was considered a *positive case* for that procurement entry; otherwise, it was considered a *negative case*. Figure 3.6 summarises the workflow described above.

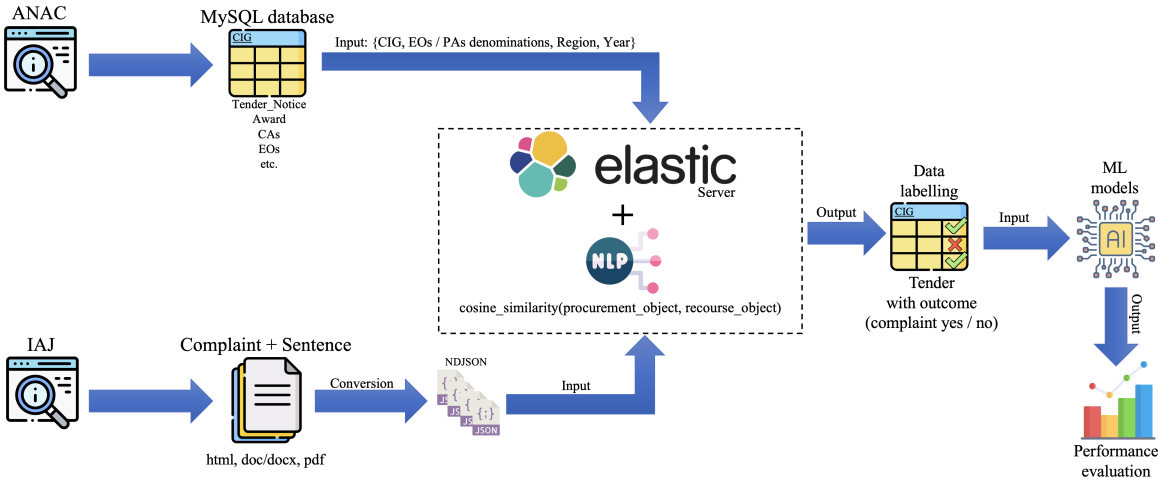


Fig. 3.6 Methodology workflow from data collection to merging, labelling and prediction evaluation

ML predictions

The No Free Lunch Theorem (NFLT) [5] states that no ML models work best in all situations and data sets. Following this, the best approach to finding out which ML model's prediction is the most accurate is to test multiple ML models and tune and compare them for the specific scenario. Since the problem is predicting whether a procurement will have a complaint, a binary classification algorithm will be used. Identified the solution as a supervised learning classification task [44], the following classifiers were explored: KNN, LR, NB, SVM, DT, RF, and XGB. The input features of the ML models are mainly described in Table 3.1 (Section 3.2.2), to which the dependent variable "complaint" needs to be added, which takes the value 1 if the procurement has experienced complaint, 0 otherwise (Section 3.4.3). The two classes are inherently imbalanced, with one possibly more prevalent than the other. To address this, the dataset was balanced by sampling equal numbers from each class. This ensures that the ML models do not favour the more frequent class, thus providing a fair evaluation. The balanced data was randomly divided into *training*, *validation*, and *test* sets, with an $80-10-10$ split, where the test set was reserved for unseen data to evaluate model performance on new, unobserved examples. To test their generalisation capabilities, the models were evaluated using *cross-validation* [110]. This method ensures that models perform well across different subsets of data, reducing the risk of overfitting the training set.

Dataset features were standardized and encoded to improve the performance of certain models, ensuring they operate effectively across varying scales of input data. In detail, based on the dataset described in Table 3.1, several preprocessing steps were applied to prepare the data for ML models. *Numerical fields*, such as the tender amount and awarded amount,

were normalized to bring all values within a standard range, ensuring that no single variable dominated the learning process due to differences in scale [172]. *Categorical fields* like Tender type were encoded using one-hot encoding [180], which converts the categories into binary vectors for compatibility with ML algorithms. For *boolean fields*, such as the framework agreement and subcontracting admitted, a boolean encoding was used, where 1 indicates true and 0 indicates false. These preprocessing steps are standard in ML to ensure that the data is in a format suitable for training, reducing bias and improving the accuracy of the models [95].

Hyper-parameter tuning was conducted on the models using the *Hyperopt* [15] library to optimize their performance and identify the best combination of parameters for accurately predicting procurement affected by variants.

On a technological level, the ML models were implemented in Python (version 3.12), the *Scikit-learn*[22] library (version 1.5.1) and *Hyperopt* library[23].

Explainable models

Quite often, ML models are black boxes, as is the case of ensemble methods like RF and XGB, because they make predictions that are difficult to explain or justify because the outcomes are due to a multitude of features. In particular, in the domain of justice, it is very important to justify the outcomes of the predictions since the end-users need to be informed about the reasons why the procurement risks a complaint.

According to surveys on Explainable AI models, such as [144], an explainable model is global if it explains the behavior of an ML model in its entirety, or it is local if it explains the predictions of individual instances. The locality is particularly useful in the legal domain because it allows a counterfactual explanation [45, 260]: in the legal domain, it explains which are the smallest changes in the feature values of procurement that change the prediction to a predefined output, e.g., might turn the presence of a complaint into an absence.

An ML model is transparent if its predictions are immediately explainable; it is post-hoc if the explanation is obtained a posteriori of the predictive model, with the adoption of further procedures. In this work, a post-hoc method has been adopted, called SHapley Additive exPlanations or SHAP [130] that can be used both for the explanation of prediction on single examples and for providing the features that are most decisive for the predictions of an ML model. SHAP is an explanatory method based on a solid mathematical foundation, which illustrates individual predictions based on the Shapley values of game theory. S. Lundberg and S. Lee in [130] reframe the problem of computing how each member of a team or

[22] https://scikit-learn.org/stable

[23] https://hyperopt.github.io/hyperopt-sklearn

coalition contributes to a coalition value, into the problem of computing how much a feature value in a given instance contributes to the model output. The idea is to explain the prediction of the original ML model (denoted here by f) on an instance x through a surrogate and simpler model, the explanation model (denoted by g), by a score computed as the sum of the contributions of a subset of the original features, each with a unit weight multiplied by a coefficient that is the Shapley value. Prediction is given by:

$$f(x) \approx g(x') = \Phi_0 + \sum_{i}^{M} \Phi_i \tag{3.1}$$

where x' represents the instance x projected on a subset composed by M original features and Φ_i are the Shapley coefficients.

In this way, the outcome score of the ML model is obtained by an additive formula that can be used both for explaining the prediction of single instances (in which the features with the highest Shapley coefficients weigh more) and for the global prediction.

The SHAP values have been computed for explaining the outcomes of the RF model using the Python SHAP library[24].

3.4.4 Results

Merge of ANAC and IAJ datasets

As outlined in Section 3.4.3, three types of searches were conducted to progressively improve the matches between IAJ sentences and ANAC tender records for labelling purposes: using {CIG}, using {EO participant denomination, EO winner denomination, CA denomination, Region/Court, Year}, and using the similarity between tender object, complaint object. The results in Table 3.7 demonstrate how the methods, applied incrementally, enhance the correspondence between the ANAC and IAJ datasets based on the available sentences (67,850).

ML predictions

The finally obtained labelled data set consists of 30,234 rows; the distribution of *positive cases* (tenders with a complaint) was analyzed in terms of type, region, CA type, CPV division, and year for discussing it with domain experts. The labelled data set was then divided into three smaller datasets containing procurement grouped by type: a data set for Works of 10,150 rows (5,075 positive/negative cases), one for Services of 15,028 rows

[24] https://shap.readthedocs.io/en/latest

(7,514 positive/negative cases) and one for Supplies of 5,232 rows (2,616 positive/negative cases). Table 3.8 illustrates the results in terms of Accuracy, F1-Score and ROC/AUC of the three best models, while Table 3.9 displays the results of the three best models with 10-fold cross-validation (a good standard value for k is 10, as suggested by [111]).The table demonstrates that XGB and RF consistently outperform the other models across the datasets. XGB performs best in two of the three smaller datasets (Services and Works), with the highest accuracy, F1-score, and ROC/AUC values. RF closely follows XGB, often taking the second spot and even outperforming XGB slightly in the Supplies dataset. This indicates that XGB and RF are the most robust classifiers in this experiment, consistently yielding better results than SVM and other models. Their superior performance highlights their effectiveness in handling different tender types and suggests they are well-suited for complaint prediction. Finally, Figure3.7 shows the ROC/AUC curve of the various models considered; the curves in the three graphs demonstrate that the XGB consistently performs best across all datasets, achieving the highest AUC scores. Specifically, XGB achieves an AUC of 0.928 for Services (Figure 3.7.a), 0.864 for Supplies (Figure 3.7.b), and 0.855 for Works (Figure 3.7.c). RF also shows competitive performance, closely trailing XGB in each case, particularly for the Supplies dataset (where outperformed XGB in Accuracy), where both classifiers achieve similar AUC values (0.864). Meanwhile, the other models, such as SVM, LR, and DT, show notably lower AUCs, indicating that they are less effective at distinguishing between classes in these datasets.

Explainable models

Figure 3.8 illustrates the SHAP values for the RF model applied to the Services data set; in the bee-swarm plot the features are ordered by their effect on prediction; it also shows how higher and lower values of the feature affect the result. Each dot in the plot represents a single observation; the horizontal axis represents the SHAP value, while the color of the point shows if that observation has a higher or a lower value when compared to other observations. In Figure 3.8, higher tender amount and the maximum frequency of EO selection criterion (of value 1) have an impact on the prediction of *positive cases*, while lower values have an impact on the prediction of *negative cases*; the third most important feature affecting the model is the bid award, whose logic is similar to the tender amount. In fourth place, the importance of the type of product purchased (CPV code) is reported. As also expected from the domain experts, the tender amounts (amount and bid-award) and the tender selection criteria (eo_selection), are the features that most influence the litigation of the tender; this consideration of the domain experts, reinforces the need for an increased explainability of the results of the IA models.

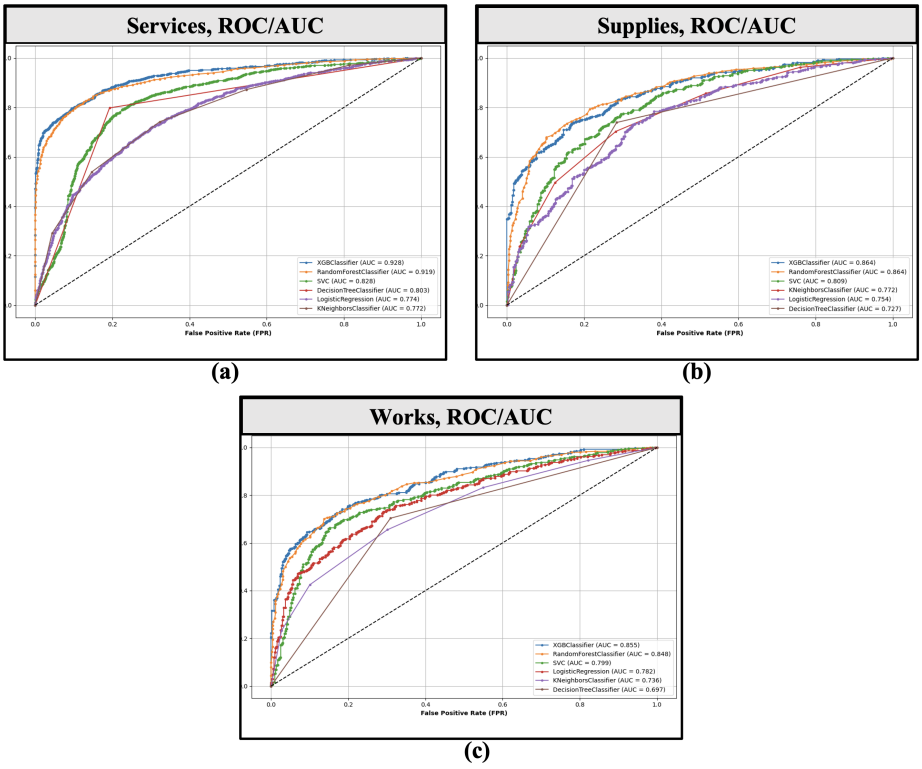


Fig. 3.7 ROC/AUC in the best case for the ML models applied to Services (a), Supplies (b), and Works (c) datasets

As a reminder, SHAP shows the contribution or the importance of each feature on the prediction of the model, it does not evaluate the quality of the prediction itself.

3.5 Conclusion and future work

The present study demonstrates the feasibility of managing a large juridical dataset from the Italian National Public Authority to automatically extract meaningful knowledge for addressing ML experiments (*RQ1*). A predictive experiment was conducted to estimate the likelihood of a complaint being filed before the administrative courts based on public procurement features (*RQ2*). Finally, an explanation of the prediction model was provided using SHAP, a method from game theory that calculates the impact of procurement features on the prediction outcome. The study illustrates how established methods and technologies (OD, IR systems, NLP, ML) can enhance systems, processes, and services in the public administration sector.

Future work will explore contemporary approaches such as few-shot learning, in-context learning, and LLM (e.g., GPT-4, T5) to assess whether they improve model performance,

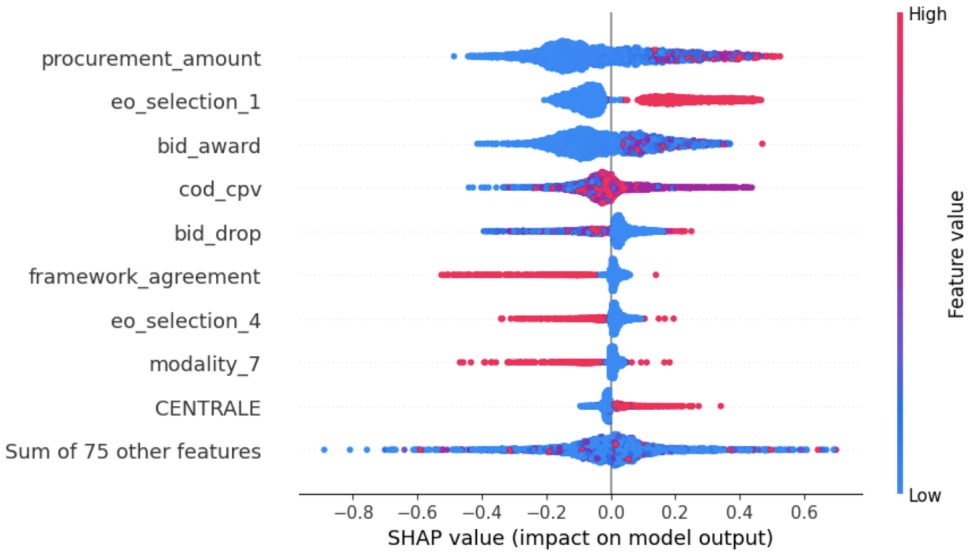


Fig. 3.8 SHAP summary plot showing the impact of each feature on the model's predictions. Features are ranked by importance, with red indicating higher feature values and blue representing lower ones. The SHAP values on the horizontal axis demonstrate the direction and magnitude of each feature's influence on the model's output

including enhancing the retrieval system to match cases between ANAC and IAJ. Further investigation into explainable AI techniques, such as extending SHAP to other ML models or using other methods, such as LIME, is also planned.

Table 3.6 Papers selected for the literature review, listing the input data sources and the methods and technologies used in the analysis (part 2/2)

Paper	**Input data**	**Methods & Technologies**
[81]	Public procurement open data from Brazil, Italy, Japan, Switzerland and USA	SGD, Extra Trees, Random Forest, Ada Boost, Gradient Boosting, SVC, K Neighbors, MLP, Bernoulli Naive Bayes, Gaussian Naive Bayes, Gaussian Process. Technologies: Python and Scikit-learn library.
[142], [185]	Electronic Public Procurement of Croatia	NLP, naïve Bayes, logistic regression, support vector machines. Technologies: Python.
[162]	Public procurement open data from Paraguay	Unsupervised learning model for anomaly detection based on the Isolation Forest algorithm. Technologies: KNIME framework.
[35]	Portuguese Public Procurement	Supervised ML, graph-oriented database. Technologies: Python Scikit-learn library, Neo4j.
[80]	Public procurement open data from Spain and EU	Random forest regression method. Technologies: Random Forest Regressor from Scikit-learn (Python).
[179]	European Economic Area members and associate countries	Random forest. Technologies: n.a.
[53]	Various (private and public)	Text Analytics, Social Network Analysis, Unsupervised learning, Online probabilistic learning. Technologies: Python, Java, DB/2.
[49]	Italian dataset managed by the ANAC	Lasso, ridge regression, and random forest. Technologies: n.a.

Table 3.7 Reference found between ANAC tender and IAJ sentences: as can be observed, multiple search techniques have increased the number of matches between tenders and judgments with complaints related to tenders

Reference found by feature	**Total**	**Overall percentage**
Procurement identifier: CIG	8,418	12.4%
Denomination: EO participant, EO winner, PA, Region/Court, Year	4,178	18.5%
Similarity: procurement object, complaint object	2,491	22.3%

Table 3.8 ML models performance measures in predicting a complaint of the three best models. In **bold** the best model for every tender type, underlined the second

Data set	**Classifier**	**Accuracy**	**F1-Score**	**ROC/AUC**
Services (15,028 rows)	**XGB**	**0.847**	**0.841**	**0.928**
	RF	0.847	0.841	0.919
	SVM	0.802	0.801	0.828
Works (10,150 rows)	**XGB**	**0.773**	**0.761**	**0.855**
	RF	0.768	0.756	0.848
	SVM	0.756	0.745	0.799
Supplies (5,232 rows)	**RF**	**0.780**	**0.774**	**0.864**
	XGB	0.778	0.770	0.864
	SVM	0.751	0.748	0.809

Table 3.9 ML models performance measures on cross-validation (average values) of the three best models. In **bold** the best model for every tender type, underlined the second

Data set	**Classifier**	**Accuracy**	**F1-Score**	**ROC/AUC**
Services (15,028 rows, $k = 10$)	**XGB**	**0.844**	**0.835**	**0.919**
	RF	0.836	0.830	0.909
	SVM	0.728	0.709	0.795
Works (10,150 rows, $k = 10$)	**XGB**	**0.743**	**0.726**	**0.815**
	RF	0.740	0.720	0.812
	SVM	0.707	0.679	0.769
Supplies (5,232 rows, $k = 10$)	**RF**	**0.759**	**0.747**	**0.828**
	XGB	0.758	0.735	0.823
	SVM	0.700	0.684	0.773

Chapter 4

Process Mining for Law: Comparing Transparency and Efficiency in European Tenders

4.1 Introduction

Knowledge management techniques have increasingly been recognized for their potential to enhance understanding of data and procedure within the legal sector [226]. The digitization of public administration facilitates the storage of legal datasets, which can be effectively managed with recent advancements in computational technology. This enables enhanced automation of administrative procedures. Integrating Artificial Intelligence (AI) into the analysis of administrative procedures offers an excellent opportunity to uncover latent patterns and features, thereby enhancing the management of complex legal workflows. AI techniques have the potential to make the law more coherent, fairer, and more transparent. [258]. The merging of knowledge-driven and data-driven AI into a cohesive knowledge-driven AI system plays a significant role in enhancing the overall quality of justice by combining the strengths of both approaches. This allows for more informed decision-making and improved outcomes within the legal framework [19].

Legal experts, as well as laypersons, may encounter challenges in extracting information on legal processes from large datasets, which, despite being designed for transparency and access to justice, can be difficult to navigate. Consequently, to address the issue, adopting automated legal process analysis techniques may represent an effective solution. This is particularly relevant for the work of public authorities, reflecting the growing demand for good administration. Citizens increasingly expect public services, including public

procurement, to be effective and easily accessible. In this respect, Public procurement encompasses the processes through which these entities acquire goods, services, or works from external suppliers or contractors, typically through competitive bidding or negotiation procedures. In this area, AI algorithms, when exploited to consider objective facts, can also help mitigate risks of corruption and abuse. More generally, legal processes present unique challenges, including handling procedural complexity, ensuring compliance with regulatory frameworks, and adhering to strict legal timelines [17]. The duration, variety, and complexity of legal processes are among contemporary justice systems' primary challenges [207].

These challenges are addressed by applying automated process analysis, specifically through PM techniques, to large legal datasets. From these datasets, legal event logs are extracted to analyze the timing of events within legal workflows. PM techniques can provide a detailed understanding of the course of legal processes, identify specific bottlenecks, and suggest targeted improvements [249]. A relevant challenge in PM is to extend the available set of events and features using unstructured sources, as in the case of extracting information from texts [232]. In the case study, the event logs are enriched by adopting LLMs to automatically extract relevant events and dates from legal texts. Finally, domain experts closely and carefully supervise the design and construction of the legal process analysis to facilitate the effectiveness of innovative AI technologies in the legal sector.

To demonstrate the approach's feasibility, a knowledge management framework is implemented in a case study focused on European Public Procurement. This study analyzes the Tender Electronic Daily (TED) dataset, which spans five European countries. First, the results obtained by applying process discovery techniques to data from public tenders are discussed. Then, the use of LLMs to automatically extract events and dates from tender texts is described, and this information is leveraged to improve the results of PM techniques. A larger dataset was obtained by merging European TED data by linking tenders to the national procurement agency (the Italian ANAC authority). This exercise allows a more fine-grained analysis, including extra events not appearing in the European dataset. In conclusion, domain experts evaluate the whole proposed approach, mostly focusing on the Italian case study, after following the methodological framework steps. The results of this study highlight the potential for examining legal workflows and propose solutions to improve the efficiency and effectiveness of legal process analysis.

In particular, this study seeks to address the following research question:

- *RQ1: How can PM techniques be applied to a legal dataset to identify procedural patterns in multiple countries and improve understanding of legal process efficiency?*

- *RQ2: How can legal processes in different countries be automatically compared, and what are the limitations and considerations in making such comparisons?*
- *RQ3: How can key information and significant events be extracted from a legal dataset to enhance the effectiveness of a PM model?*

In the remainder of this chapter, the analysis of related work is provided in Section 4.2, while Section 4.3 outlines the dataset overview, and Section 4.4 presents the methodology adopted. Section 4.5 covers the PM results. A discussion is offered in Section 4.6, and Section 4.7 wraps up the chapter.

High-resolution images for this chapter are available at https://bit.ly/4eRONLT. The source code for the experiments conducted in this chapter is available in the repository: https://github.com/roberto-nai/PhD-THESIS.

4.2 Related work

The methodological approach is grounded in state-of-the-art developments in the field, drawing on insights from pivotal research articles. One notable contribution is the study on the integration of the European Tender Electronic Daily (TED) dataset with a national dataset, as demonstrated by the French case reported in [181]. Similar to this research, the authors of the study collected 478,854 contract notices and attempted to merge them with award notices from a French database. Following this research trajectory, the contribution is enriched by various research areas, including Business Process Management (BPM), general PM, and a combination of Natural Language Processing (NLP) and ML within the legal domain, the main contributions of which are outlined in the following subsections.

4.2.1 BPM applications

Legal processes have traditionally been examined from a BPM perspective, focusing primarily on issues related to regulatory compliance. This approach ensures that business processes adhere to relevant laws and regulations, maintaining legal integrity and operational efficiency [209, 96]. Some approaches were proposed to check the conformance of processes automatically, but an effort of human supervision remains necessary [102, 8]. The discipline of PM [249] is gaining increased attention across various application areas, expanding beyond its traditional focus within BPM research. This growing interest highlights the versatility and potential of PM techniques in diverse fields, such as healthcare [104] and education [21].

4.2.2 PM applications

The conformance checking research area [149] seems promising to investigate legal compliance issues [34, 182]. Recently, some works investigated a process-oriented approach to legal cases. At the intersection of PM and law, Mannhardt et al. [135] introduced a framework to describe the GDPR impact on the design of PM systems, while [66] discusses approaches to ensure privacy in PM. Previous studies investigated real-world cases of process discovery in the context of public procurement. For example, one case study leveraged a heuristic algorithm to uncover a concept drift in the publication of contracts in the Philippines [212]. Following a similar approach, further research highlighted the specific challenges of procurement processes in Croatia [184]. A comprehensive application of process discovery in the legal field is exemplified by [246], where the authors provided insights for improvement to increase judicial productivity by applying discovery techniques to extract lawsuit processes from the information system of the Court of Justice of the State of São Paulo, Brazil.

4.2.3 NLP and PM applications

A relevant area of AI research involves combining ML techniques with NLP. Recent studies have proposed the innovative approach of exploiting unstructured information contained within natural language textual descriptions of processes [18, 205, 232]. The goal is to transform these descriptions into formal process models, thereby enhancing the accuracy and usability of process modelling in various applications [211]. Another research investigates the enhancement of event logs through the integration of unstructured text. In [83], the approach aims to enrich the event logs with additional context and details extracted from textual data, improving the depth and quality of process analysis. In this direction, the work includes the extraction of new events and their corresponding timestamps from public tenders.

4.3 TED Dataset Overview

The legal documentation and data concerning public procurement tenders, along with the processes associated with publication procedures, are available on the TED website[1]. TED constitutes the online version of the Supplement to the Official Journal of the European Union (EU). It applies to all public tenders above specific contract values to be published in the aforementioned supplement. TED is the official online platform for European public procurement, where approximately 735,000 tender notices are published annually. These

[1] https://ted.europa.eu/TED/browse/browseByMap.do

include around 258,000 calls for tenders, with an estimated value of approximately €670 billion. This wide repository provides important information for businesses and organisations interested in participating in procurement activities across the EU.

The following sections describe the TED dataset, its enrichment through texts linked to tenders, and its specialisation for Italian cases. A full discussion of the data model obtained can be found in Appendix B.

4.3.1 Dataset of procurement notices

Within TED dataset, each tender is uniquely identified by an alphanumeric value known as the *document-number*, which serves as the key identifier for the tender. Several important features are associated with each tender, such as the *sector* to which the tender belongs is categorised as either "Services" (S), "Works" (W), or "Supplies" (U). Another relevant information is the *NUTS* code (Nomenclature of Territorial Units for Statistics [12]), which specifies the geographical region of the Contracting Authority (CA) responsible for issuing the tender notice. Further, the dataset provides details about the *type* of the CA, which can vary from national entities, such as Ministries, to supranational bodies like European Institutions or regional and local authorities, and the *amount* of the tender (a key feature, offering insights into the financial scale of the procurement process). Finally, for each tender, there is a *url* field that links to the full-text version of the tender, available in PDF format.

The tender dataset is available in CSV format, and the full schema for data is publicly available[2], providing comprehensive technical details on its structure and contents for users who wish to work with this kind of data (see Appendix B). The attributes of tenders form the foundation for analysing procurement patterns and trends across different sectors, regions, and levels of government.

Legal Processes and Activities

Although the dataset includes tender data from 24 EU countries, the research focused on the regional authorities of five EU member states - France (FRA), Germany (DEU), Italy (ITA), Spain (ESP), and Portugal (PTG) - based on recommendations from legal experts. According to these experts, the legal review systems in these five countries indeed share significant similarities, making them ideal for comparative analysis.

The public procurement process comprises five key stages (hereafter written in capitals). It begins with the PUBLICATION of the tender, followed by the PARTICIPATION of interested entities submitting their bids. After the evaluation, the tender is awarded (AWARD)

[2]https://data.europa.eu/api/hub/store/data/ted-csv-data-information-v3-5.pdf

by an Economic Operator (EO), leading to the formalisation of the contract (CONTRACT-START). The process concludes with the contract's stipulated end date (CONTRACT-END). Each stage is associated with a *date* in the format *dd-mmm-yy* (e.g. 05-JAN-22 to indicate 5 January 2022). A sixth event often emerges from the tender textual documents, namely the BID-OPENING, which occurs before the AWARD stage and plays a crucial role in the evaluation process; in this case, since it is textual data, the format of the date may vary from one tender to another (e.g. 05.01.2022, 05 January 22, etc.). Figure 4.1 illustrates an example of the text used in the bid opening section for DEU and ESP cases.

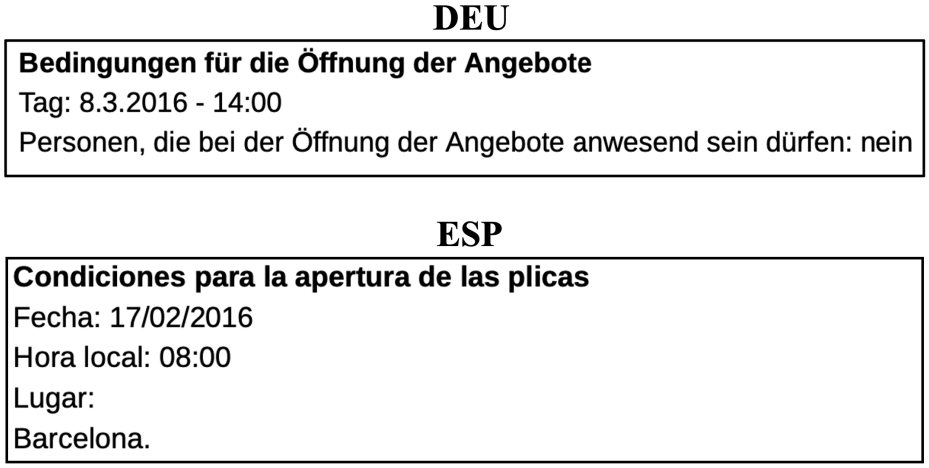


Fig. 4.1 TED - Example of texts in the bid opening section for DEU and ESP cases

4.3.2 Dataset specialisation with Italian Cases

To provide more detailed insight into the ITA cases, the focus shifts to ANAC dataset, as described in Section 3.2.2.

The data available in the ANAC dataset consists of twelve key stages (hereafter written in capitals): a tender is created (TENDER_NOTICE event) and published on the Italian[3] or TED communication channels of the public administration (respectively, PUBLICATION-IT or PUBLICATION-EU). A tender is then awarded (AWARD) by an EO, followed by the start of a contract execution (CONTRACT-START). After the start of the contract, various combinations of events may occur: an EO may sub-contract the tender (SUBCONTRACT); tender execution may be suspended due to problems or unforeseen events (SUSPENSION and REPRISE); the event to be performed may vary from the original bid (VARIANT). Depending on the tender, its status may be cyclically recorded (STATE), or checks related to

[3]https://www.gazzettaufficiale.it

the status of the tender may be performed (TESTING). The tender cycle usually concludes at the end of the contract's execution (CONTRACT-END) or following its 'concluded' status. Each stage is associated with a *date* in the format *yyyy-mm-dd* (e.g. 2022-01-05 to indicate 5 January 2022).

Finally, the period 2016-2022 is examined, as it closely corresponds to the period during which the Italian public procurement code was in effect. The code was introduced in April 2016 and remained in force until it was replaced on 31 March 2023[4].

4.4 Methodology

The methodology to analyse legal processes, such as public procurement, through automation using PM and AI follows a series of steps, which are summarized in Figure 4.2. A preliminary step in the research involves collecting procurement data, including the official TED dataset and the full-text version of each tender stored inside of it; this data has been gathered through *web scraping* [51].

Three key steps are explored in this workflow: first, constructing the process event log from the TED dataset by extracting relevant information through a parser to generate a file suitable for PM analysis. Second, the event log is enriched using an automatic feature extraction method from tender-related texts, leveraging innovative techniques with LLMs. Finally, a national dataset is used to track new events or link additional process features to the tenders outlined in the TED dataset. The remainder of this section elaborates on these topics. The source code supporting operations described hereafter is publicly available[5].

4.4.1 Scraping public procurement repositories

The preliminary step involves collecting the TED dataset, including tender data in CSV files and their full-text versions. This task was carried out performing web scraping on the main TED repository[6], to automate the process of downloading all the files.

4.4.2 Event log of legal process

The starting point of a PM research is an *event log*, i.e. a set of *traces* where each trace stores a sequence of *events*. Each trace represents the execution of a sequence of *activities*

[4]Legislative Decree n. 50 of 2016, replaced by Legislative Decree n. 36 of 2023
[5]https://bit.ly/3xQelsU
[6]https://data.europa.eu/data/datasets/ted-csv?locale=en

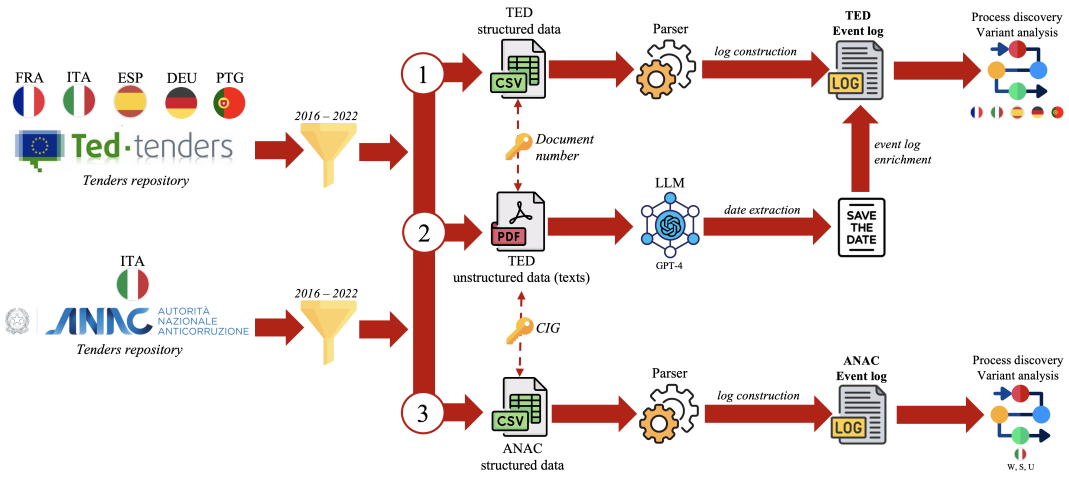


Fig. 4.2 Methodological steps from dataset features to event log enrichment (TED event log) and Italian specialization (ANAC event log)

that occurred during a single execution of the process. The event log captures such information, possibly together with some additional data (e.g., the resource who performed the activity) [249]. Each trace in the event log is identified using a so-called *Case ID* (case identifier).

In particular, three mandatory attributes are needed to generate an event log from a set of non-process-oriented samples, such as the TED dataset used as a starting point: first, the *Case ID* indicates which case of the process is responsible for each event. Second, the *event class* (or *activity name*) specifies which activity the event refers to. Third, the *timestamp* specifies when the event occurred [251]. An event log may carry additional attributes in its payload; these are called *event-specific attributes* (or *event attributes* for short) [249].

In this case study, the tender document number serves as the case identifier. The activities include the corresponding execution date at a day-level granularity, standardised in the *yyyy-mm-dd* format (e.g. 2022-01-05 to represent 5 January 2022). In addition, event attributes are added to the event log file, including information on the corresponding sector, the NUTS code, and the amount of each tender.

Finally, to maintain a consistent event log, features not relevant to the analysis were removed, e.g., the VAT number of the CAs, as well as cases too short (in terms of number of activities), which are not meaningful according to domain experts. These short cases are likely the result of input errors by data entry form fillers.

4.4.3 Process discovery and variant analysis

The event log can be explored with an initial analysis, which included filtering the data, exploring the processes through discovery techniques, and identifying potential bottlenecks in the workflow[7].

Different traces in the event log are called *process variants* (*variants*) since they represent alternative ways to execute the same process (cases may perform activities in a different order before the end). A variant analysis was conducted to uncover any significant differences between subgroups of process executions. Variant analysis refers to a set of techniques used to analyse event logs to detect and explain variations between two or more process flows. In this study, the process models generated for the identified variants were compared. This exploratory phase enables us to gather valuable insights into whether these variants exhibit meaningful differences and helps determine which specific properties warrant further investigation and deeper exploration.

4.4.4 Event log enrichment with LLM

Information extraction is a highly intricate task that requires the implementation of various techniques, including but not limited to named entity recognition, regular expressions, and text matching [143]. Each of these methods contributes to the broader challenge of accurately extracting relevant data from unstructured text. In recent years, advancements in LLMs have introduced groundbreaking innovations in this domain, significantly improving the accuracy and efficiency of information extraction processes. By leveraging wide and diverse datasets and algorithms, these models have achieved remarkable progress in understanding and processing natural language with a depth and precision previously unattainable [7]. This enabled a more nuanced and effective approach to handle complex text data across various contexts, further enhancing the potential of information extraction workflows.

Moreover, comprehensive reviews and empirical studies have detailed the integration of LLMs into various information extraction workflows, emphasizing their ability to generalise across diverse contexts and tasks [190]. As part of the research on date extraction, OpenAI models were specifically explored, with a particular focus on GPT-4 [3].

As described in Section 4.3.1, the specific and relevant sections related to bid opening from the tender full-text versions were extracted and used as input for processing by LLMs. The dates identified from these texts were associated with the newly identified BID-OPENING event, which was then added to the TED event log, ensuring that all relevant occurrences were accurately recorded and integrated into the timeline of tender activities. To

[7]Fluxicon DISCO (https://fluxicon.com/disco) has been adopted process analysis

verify the correctness of the LLM's responses, the expected results for every language were labelled, and the responses were evaluated against these labels. The *prompt* [30] was refined through tuning cycles to ensure all LLM's responses were 100% accurate.

From a technological perspective, the *PyPDF2* library was used[8] to extract tender texts, and OpenAI *APIs*[9] to send input for processing by LLMs. This workflow enabled the integration of PDF text extraction with the advanced capabilities of LLMs, enhancing the data processing pipeline.

4.4.5 Extraction of Italian cases

Given that a section of the TED texts contains links to national references, the CIG number (see Section 4.3.2) is utilized as the reference identifier for ITA cases. The CIG numbers were extracted from the relevant section within the TED tender texts. Following extraction, the corresponding tenders in the ANAC dataset formed the basis for creating the ANAC event log, specifically focused on ITA cases. As in previous sections, the resulting ANAC event log was subsequently analysed for process discovery and variant analysis.

4.5 Results

The following sections present the results of constructing the TED event log, discovering key workflows, and enriching the TED event log. Additionally, the ITA cases were analysed using the ANAC event log.

4.5.1 Legal dataset

The TED event log contains 9,228 FRA tenders, 86,674 DEU tenders, 1,474 ESP tenders, 2,220 ITA tenders, and 2,333 PTG tenders, for a total of 99,637 cases. Using this event log, the five country cases were compared using standard metrics, such as the number of traces, instances, and variants. Figure 4.3 provides an overview of the TED event log in CSV format. It displays two *tracks*, identified by the *Case ID*, *Activity* (e.g. PUBLICATION, PARTICIPATION, etc.), *Timestamp* (e.g. 2017-03-17), and the event attributes *Sector*, *Amount*, *Nuts* and *Country*.

[8] https://pypi.org/project/pypdf
[9] https://pypi.org/project/openai

```
Case ID;Activity;Timestamp;Sector;Amount;Nuts;Country
...
2017106814;PUBLICATION;2017-03-17;S;1035000.0;FR107;FRA
2017106814;PARTICIPATION;2017-05-09;S;1035000.0;FR107;FRA
2017106814;AWARD;2017-06-07;S;1035000.0;FR107;FRA
2017106814;CONTRACT-START;2017-09-01;S;1035000.0;FR107;FRA
2017106814;CONTRACT-END;2022-07-31;S;1035000.0;FR107;FRA
2017107959;PUBLICATION;2017-03-20;S;637622.4;ITE19;ITA
2017107959;PARTICIPATION;2017-05-03;S;637622.4;ITE19;ITA
2017107959;CONTRACT-START;2017-06-01;S;637622.4;ITE19;ITA
2017107959;AWARD;2017-06-15;S;637622.4;ITE19;ITA
2017107959;CONTRACT-END;2020-05-31;S;637622.4;ITE19;ITA
...
```

Fig. 4.3 An extract from the TED event log in CSV format including two cases: one from France (FRA) and another from Italy (ITA)

4.5.2 Event log of legal process

As stated above, a legal event log was obtained, consisting of 99,637 cases and 553 variants. The cases from all five countries have a median duration of 17.4 months and a mean duration of 25.2 months. The process model in Figure 4.4 highlights the most common process behaviours; rectangles represent process activities, while edges represent pair-wise ordering relations among the activities. The darker a rectangle is, the more frequently the corresponding activity occurs in the event log. Similarly, the thicker an edge is, the more frequently it is observed in the event log that the target activity eventually follows the source activity. The exploration of the diagram and the event log with domain experts show that some data need to be filtered out to get a better version of the process, excluding process variants not applicable; for instance, it was observed that some paths (e.g. from CONTRACT-END to PUBLICATION) are not possible according to domain experts and some cases with extremely high duration also, e.g., the tenders without an ending date (CONTRACT-END). This issue is due to data quality problems, as the tender data were entered manually by the operators of the individual CAs.

4.5.3 Process discovery of five states

Process variants were analyzed by considering the performance of each country, expressed in terms of duration. Table 4.1 summarizes the differences in event logs corresponding to FRA, DEU, ITA, ESP and PTG processes. Interestingly, significant differences in the number of procedures and the average duration were immediately noticed. Overall, the table highlights substantial differences in case duration and management across EU countries. DEU has a notably high number of cases with considerable variability; ITA stands out for the longest

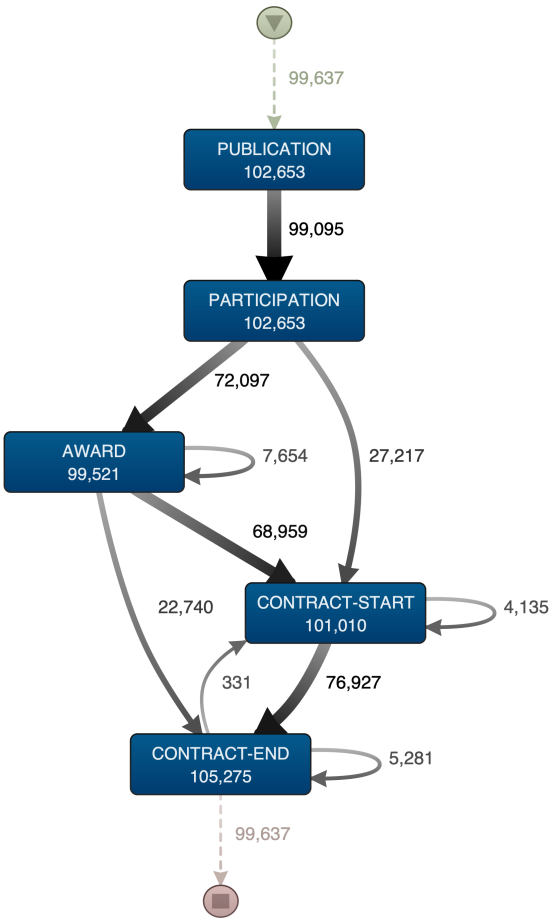


Fig. 4.4 TED - Process discovery from event log (all countries)

average duration. PTG shows potentially quicker processes but is based on a smaller sample size than DEU and FRA.

In line with Table 4.1, Figure 4.5 shows the median duration of cases across the five countries. The time required for a tender to be awarded to a winner (AWARD) ranges, on average, between 37 days (Figure 4.5.d) and 76 days (Figure 4.5.e). Similarly, the days elapsing between the awarding of a tender (AWARD) and the start of the contract with the winning entities (CONTRACT-START) vary from 13 days (Figure 4.5.e) to 39 days (Figure 4.5.a and Figure 4.5.c).

The results were discussed with domain experts, as reported in Section 4.6.

4.5.4 Event log enrichment

Starting from the complete TED dataset, the presence of the bid opening section was identified in 58,366 instances for FRA cases (representing 92% of all cases), 93,194 instances for cases

Table 4.1 TED event log - Mean, median, and standard deviation of case duration in months by country

Country	Cases	Mean duration	Median duration	SD
DEU	86,674	25.17	17.35	23.02
ESP	1,474	24.29	23.05	15.91
FRA	9,228	36.99	27.69	28.22
ITA	2,220	41.70	38.58	24.05
PTG	2,333	19.77	14.62	11.82

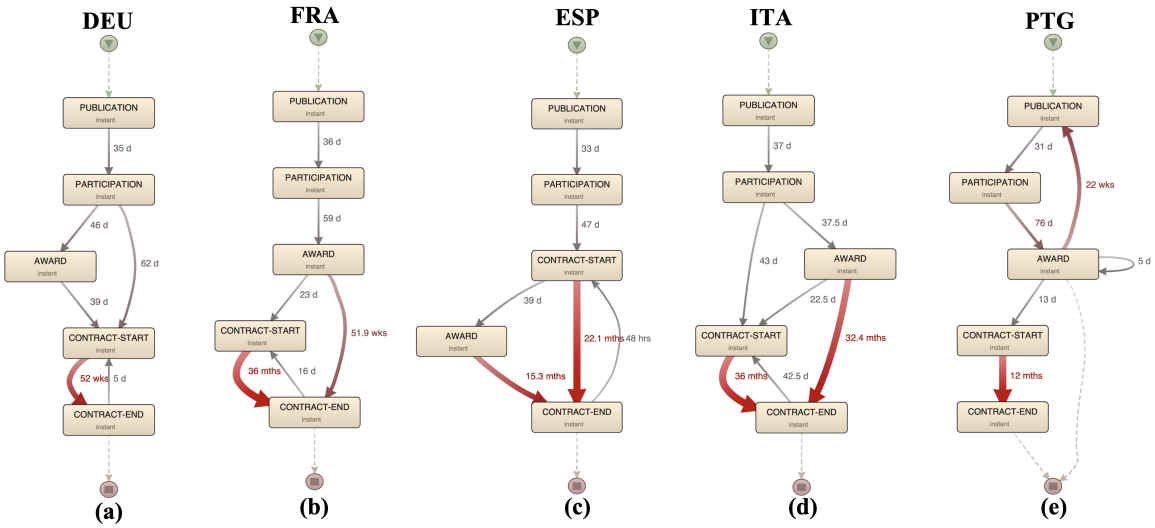


Fig. 4.5 TED event log - Median duration of the cases of five countries from

related to DEU (84% of cases), 15,725 instances for cases involving ITA (93% of cases), 23,038 instances for cases concerning ESP (75% of cases), and 2,164 instances for cases from PTG (93% of cases).

Figure 4.6 illustrates the process of the various states with the new BID-OPENING event. It can be seen that for ESP cases (Figure 4.6.c), the time elapsing from tender participation (PARTICIPATION) to bid opening (BID-OPENING) is 7 days, thus showing how the estimated time of 47 days in the not enriched event log (Figure 4.5.c) may reveal further time detail. Finally, this event extracted from the text does not appear to be 'blocking' as its occurrence begins on average immediately after the last PARTICIPATION in the tender. Enriching the event log identified a new event that enables deeper insights into the process and proves useful (as in the ESP case) for understanding how activities were allocated.

After refining the prompt, the LLM proved to be efficient, returning all dates of interest correctly. Overall, interactions with the API of the LLM remained largely stable, although there were occasional errors when submitting service requests. Extracting the relevant dates

from the text portions of the PDFs required approximately four hours. These runtimes should be considered for longer or more complex tasks on texts.

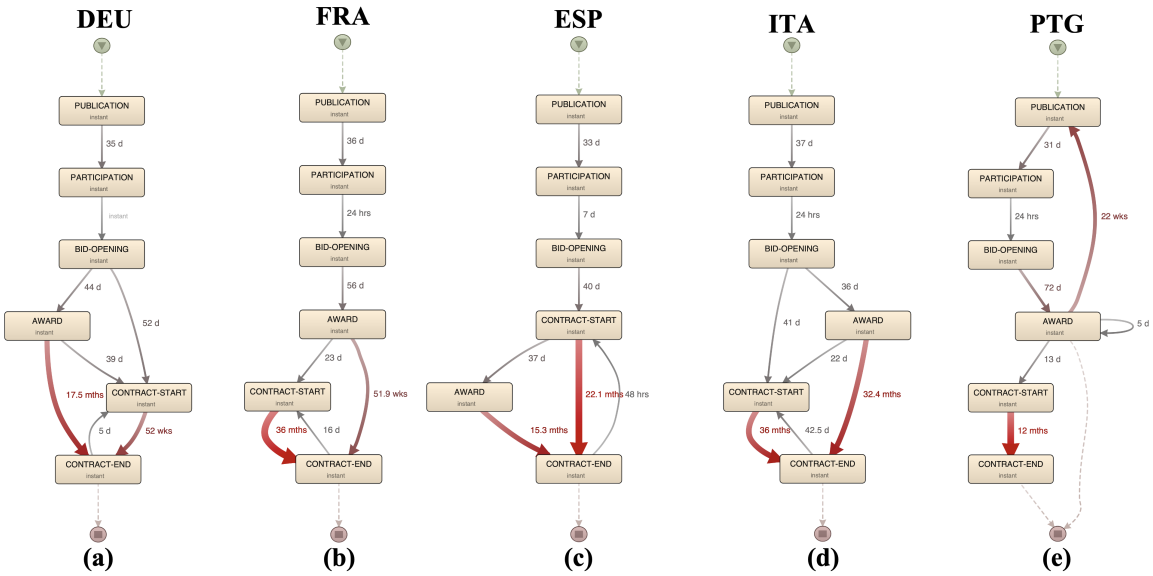


Fig. 4.6 TED event log - Process enhanced with new BID-OPENING event

4.5.5 ANAC Italian cases

Among the 2,220 cases in the TED event log, a significant 2,017 include direct references to the ANAC event log. From these instances, a subset of 861 cases has been identified as particularly useful and relevant for analyzing the ITA event log, offering valuable insights for further study. Similar to TED, cases with at least the events of TENDER_NOTICE, AWARDS, CONTRACT-START, and CONTRACT-END were considered useful. Figure 4.8.a displays the graph of the event log for *Supplies* cases, including 300 cases and 10 activities, with a median case duration of 39.7 weeks (approximately 9.1 months) and an average case duration of 55.1 weeks (approximately 12.7 months). Figure 4.8.b displays the graph of the event log for *Services* cases, including 780 cases and 12 activities, with a median case duration of 23.1 weeks (approximately 5.3 months) and an average case duration of 57.7 weeks (approximately 13.3 months). Figure 4.8.c displays the graph of the event log for *Works* cases, including 6 cases and 5 activities, with a median case duration of 164.3 weeks (approximately 37.9 months) and an average case duration of 134.3 weeks (approximately 30.9 months).

As for TED processes, the ANAC results were discussed with domain experts, as reported in Section 4.6.

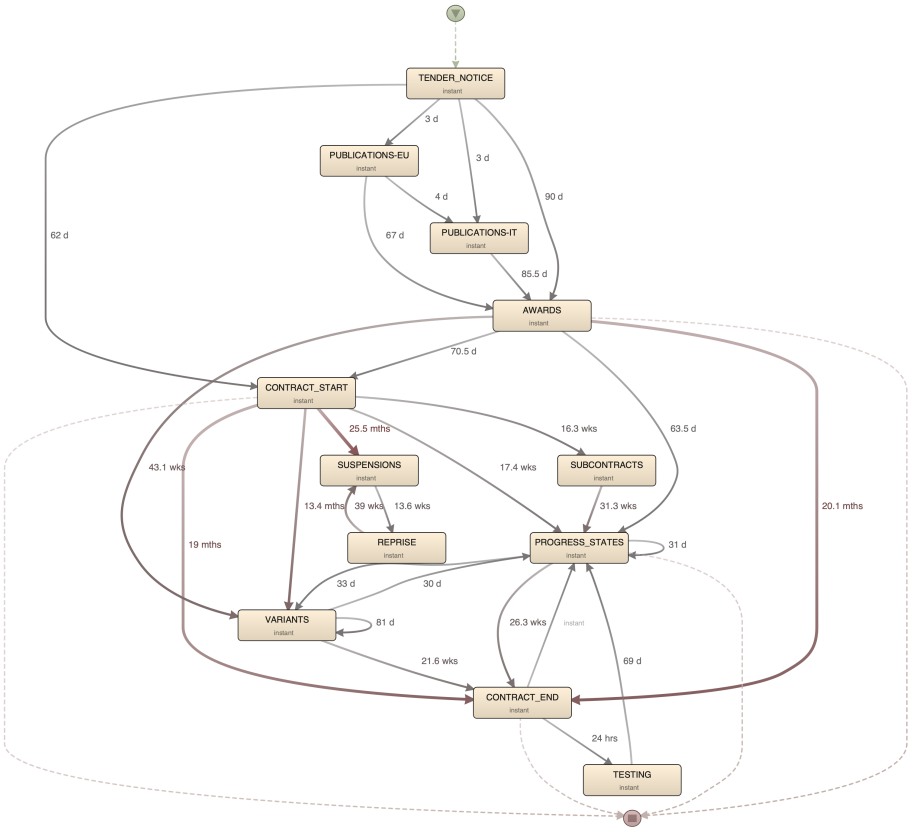


Fig. 4.7 ANAC event log with ITA cases specialization

4.6 Discussion

This section presents an overview of key factors and considerations influencing public procurement timelines and processes, as discussed by legal experts. It addresses data availability across countries, a comparative approach, and the impact of legislative frameworks on procurement procedures. Additionally, it outlines specific time-frames for various procurement methods and potential bottlenecks. In conclusion, having specialized the dataset with cases from the ANAC website, the final subsection offers a detailed analysis of the Italian case.

4.6.1 Data availability

The disparity in data availability across countries on the TED portal can be attributed to several factors. Some countries may lack the infrastructure or resources to collect and maintain accurate data, while others may place a lower priority on data sharing, as noted in an OECD report [165]. Furthermore, TED relies on contributions from countries that are more engaged with international institutions, while limited access to advanced technologies in

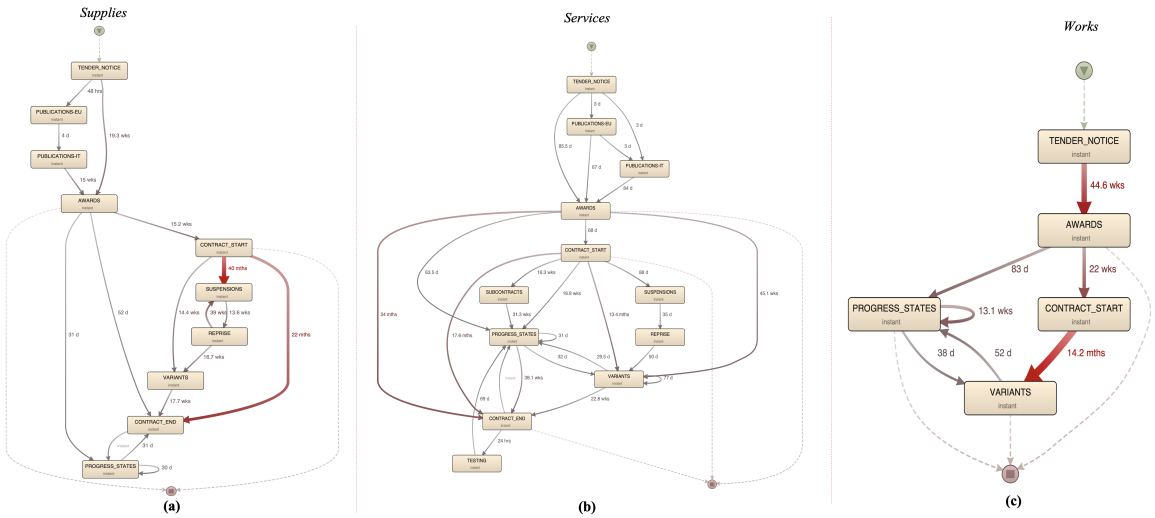


Fig. 4.8 ANAC event log with ITA cases specialization, divided by sector of tender: Supplies (a), Services (b), Works (c)

some regions affects data quality and quantity. As reflected in OECD findings, these factors contribute to uneven data availability across different states [164]. The same consideration can be applied to the national ANAC dataset.

4.6.2 General considerations

As previously mentioned, this research methodology enables country-based comparative analyses across different time frames. This facilitates an objective evaluation of the impact of legislative changes and promotes benchmarking among various contracting authorities, encouraging best practices and standardization in procedures. To suggest that procurement procedures and public tenders depend solely on the legislative framework is an oversimplification. Still, this technology certainly provides a general overview from which further investigations may begin, incorporating both qualitative and quantitative approaches that include the involvement of stakeholders, such as key officials in public contracting offices or public contract law experts.

Reversing a more traditional perspective, this work offers insights that are particularly suited to legal considerations, focusing primarily on the ITA case and, secondarily, on comparative analyses across different legal systems. PM enables effective monitoring of compliance with recent amendments in the Public Contracts Code. Through the analysis of event logs, it is possible to verify whether procurement procedures adhere to the newly imposed deadlines, allowing authorities to identify potential non-compliances swiftly.

Originally, the Italian Public Contracts Code did not mandate binding timelines for public administration tenders and bids. However, amendments introduced in 2020 to address the pandemic emergency through D.L. 76/2020 [194] subsequently converted by L. 120/2020 [196] (applicable until 31 December 2023) set specific timelines within which procedures should begin and conclude. This decree-law provides that in the case of direct awarding procedures, a period of no more than two months should elapse between the determination to contract and the awarding of the contract. In the case of negotiated procedures, a period of no more than four months should elapse, while in the case of contracts above the threshold, a period of no more than six months should elapse, regardless of whether the open or restricted procedure was implemented. This stipulation has now been incorporated into the current Public Contracts Code (D.Lgs. 36/2023[197]), which establishes terms for cases where the criterion of the economically most advantageous offer is followed.

In such cases, the time frame is set at no more than (i) nine months for the open procedure, (ii) ten months for the restricted procedure (with a maximum of seven months), (iii) four months for the competitive procedure with negotiation, (iv) seven months for the negotiated procedure without prior publication of a contract notice, (v) nine months for the competitive dialogue, and (vi) twelve months for the innovation partnership. These time frames are set out in Article 17 and Annex I.3. Therefore, failure to comply with the aforementioned maximum time frames constitutes a case of silence default, which can be addressed through legal action in accordance with Articles 31 and 117 of Legislative Decree 104/2010 [192].

4.6.3 Analysis of TED cases

Figure 4.6.c illustrates the main stages of the ITA public procurement process, including both average timelines (in days) and the overall contract duration (in months). The process starts with the publication of the tender notice, followed by 37 days for bid submission. After receiving proposals, the contract is awarded 43 days later. Some procedures take 37.5 days. A further 22.5 days may pass before the contract starts, allowing for final negotiations. The contract duration ranges from 32.4 to 36 months, depending on the procedure.

Multiple potential bottlenecks have been identified within this process. The first occurs between the PUBLICATION and AWARD phases, where the process takes approximately 80 days (37 days for participation and 43 days for awarding). This could be considered a bottleneck, especially in the case of complex or large-scale tenders. To address this, the evaluation process could be optimized through digital tools that automate pre-evaluation or by applying simplified selection methods for tenders with a limited number of participants. An additional delay arises between the award and contract start phases, where a 22.5-day gap is observed. This could also represent an avoidable delay. Improving the final

negotiation phase by digitizing the process and reducing the need for physical or paper-based interactions could be a solution. Furthermore, an internal monitoring system to track and reduce response times by the administration could be beneficial. The overall contract duration, ranging from 32.4 to 36 months, may depend on the complexity of the project. However, in some cases, it may be useful to introduce clauses that encourage early completion without sacrificing quality. Introducing bonuses for companies that complete work ahead of schedule or imposing penalties for unjustified delays could be effective. The use of project management technologies, such as project management software, could also facilitate the monitoring and management of timelines.

The identification of possible bottlenecks and inefficiencies in procurement procedures through PM allows for targeted improvements, which can lead to reduced waiting times and enhanced efficiency of the public procurement system, aligning with the objectives of legislative amendments. In line with the results, this methodology not only measures the duration of the entire procurement procedure but also allows for the precise visualization of various process phases, such as PUBLICATION, PARTICIPATION, AWARDING, CONTRACT-START, and CONTRACT-END. For instance, comparing data from ITA and Germany, it is evident that the timelines of DEU procedures are significantly longer than those in ITA. However, as explained, the number of procedures published on the TED site by the DEU public administration far exceeds that of ITA. Comparing data from the countries under analysis, the duration of timelines is directly proportional to the number of procedures involved. Similarly, ITA is notable for having the longest average duration, while PTG appears to have faster processes, although the sample size is smaller compared to that of Germany and FRA. For the purposes of this research, the Italian case warrants further analytical exploration.

4.6.4 Analysis of Italian cases

In public procurement, in ITA as elsewhere, distinguishing between “Services”, “Works”, and “Supplies” is crucial for several reasons. Each category is governed by different regulations that outline distinct requirements, procedures, and obligations. For example, “Works” contracts often involve stricter safety standards (so the amount increased for this) than Services or Supplies. Contract types also differ based on the nature of the procurement. Moreover, the separation allows for a clearer economic evaluation of each area, making it easier to identify specialised EOs. Overall, dividing tenders into Works, Services, and Supplies promotes transparency, ensures compliance with legal frameworks, and helps mitigate risks like fraud or irregularities. Furthermore, timelines and procedures are tailored to the specific nature of each category. For example, work contracts often require longer completion times and

stricter deadline management compared to services or supplies, which typically have shorter delivery times. In fact, as noted in Section 4.5.5, Works has an average time of about 31 months, much longer than Supplies (9.1 months) and Services (5.3 months).

The model can then be used to facilitate a comparison between the time frames that are concretely adopted in public procurement and those that are established by the legislator. Specifically, according to Legislative Decree 50/2016 [193], at least 36 days must elapse from the award phase to the contract start phase - as the standstill[10] period is 35 days - and no longer than 60 days.

The model developed in this study allows us to ascertain that, with regard to services, the mean time between the award of a contract and its start is 36.4 days; for works, it is 55.2 days; and for supplies, it is 24.2 days. The model thus enables the verification of whether the time-frames set by the regulatory framework adhere to common public procurement practice. It is important to note that the discrepancy between the ideal and actual time-frames, as illustrated in the realised model, is limited to the context of supply contracts. In particular, the figure in question deviates significantly from the expected standstill period of 35 days. This apparent inconsistency can be explained by the existence of exceptions to the standstill period identified by the legislator. In particular, Article 32(10) of [193] enshrines the non-operation of the aforementioned deadline for the awarding of sub-threshold services and supplies [36]. With respect to the time-frames established for the purpose of making awards, Law Decree 76/2020 [195] has established the following parameters: a maximum of two months for direct award procedures, four months for negotiated procedures, and six months for above-threshold contracts, in both open and restricted procedures.

4.7 Conclusion and future work

This work explored the adoption of PM techniques in a legal process to understand its execution. The case study demonstrated the feasibility of legal process analysis by combining process discovery, LLMs, and variant analysis techniques. The accompaniment of domain experts facilitated the conduct of the experiments presented in the case study.

By implementing a methodology based on objective data and automated analyses, PM enhances the transparency of procurement procedures. This helps prevent corruption and fosters greater trust in the public procurement system, crucial aspects in rapidly evolving legislative contexts. From this perspective, PM technologies could be combined with NLP techniques, allowing for an in-depth tender documentation analysis. PM, on one hand,

[10] In public procurement, the *standstill* period is a mandatory waiting period between the contract award notification and the final contract signature.

facilitates the visualization of procedures and their steps. At the same time, in-depth analysis of textual content can identify the occurrence, or absence, of reasons for derogation from standard procedures, which, if missing, could represent a non-compliance with the law potentially subject to appeal.

In response to RQ1, the analysis demonstrates how PM techniques were successfully applied to a legal dataset of public tender processes across five European countries. By analysing 99,637 cases, PM uncovered the main activities, event sequences, and procedural variants, effectively modelling and visualising legal processes through event log data. The analysis revealed process variants across the five countries, highlighting procedural differences. Finally, the results highlighted inefficiencies, including excessively long durations or process variants with impossible transitions between activities and underlining data quality issues.

In response to RQ2, the analysis shows that legal processes across different countries can be compared using performance metrics such as the number of procedures and their duration. The results reveal clear differences in case management and duration across the countries considered. For instance, ITA stands out for having the longest average duration, while PTG appears to have quicker processes. DEU cases, with a notably higher number of procedures, show significant variability in their timelines. These comparisons highlight important trends and can guide further analysis, particularly in understanding the factors influencing the duration of legal processes.

In response to RQ3, extracting key information and significant events, such as the bid opening section, from the TED dataset has been shown to enhance the effectiveness of the PM model. The data indicate that this information is present in most cases. Notably, considering the varying execution times across the dataset, using LLM has been crucial in efficiently extracting the BID-OPENING section, making it a key enabler in improving the PM models across different legal systems.

In the future, the investigation will be further developed by exploring different approaches (e.g., automated variant analysis) and applying them to other features in the event log (amount, NUTS, etc.). Furthermore, other LLM foundations, such as LLaMA [240], will be explored instead of proprietary tools (such as OpenAI) to extract more data from the texts and enrich the process event log. Future work will also focus on extending process discovery by incorporating context-aware approaches that include global variables and intercase features. This enhancement aims to provide a more comprehensive understanding of process dynamics, as highlighted in recent research advocating for context integration in process analysis [43, 218]. Finally, the application of predictive algorithms on trace prefixes

will be explored to identify in advance processes that are excessively long, or that will not conclude (e.g. for a complaint to the administrative justice by EOs).

Chapter 5

LLMs for Law: Exploring Applications in Public Tenders

5.1 Introduction

An area of great interest in legal informatics research concerns decision-support systems. Most activities in law work are usually carried out by domain experts, who are faced with an increasing amount of data and possible risks of incurring errors, such as not reaching all documents of interest.

Expert systems have been proposed by research on information retrieval (IR) and recommendation systems (RS) to facilitate legal practioners' work. Such systems facilitate the work of practitioners by leveraging computational capabilities, big data management, and data mining. In this context, the most used methods concern knowledge extraction and identification of documents from large legal datasets. In recent years, the impact of LLMs in Natural Language Processing (NLP) has also grown in the direction of exploiting their potential to improve RSs [269]. The adoption of RSs in the context of legal informatics represents a significant challenge to improve the efficiency of legal work organizations. This section explicitly focuses on extracting similar legal documents to facilitate the instruction of administrative processes. In particular, legal experts often have to manually track down the referenced legislation given a specific field, relying on experience and knowledge gained in prior work. This method, however, is relatively time-consuming and obviously not without risk of omissions and gaps. Increasing computational capabilities allows the volume of regulations to be managed to try to get to all similar documents quickly and accurately. A legal area of great interest for the impact on public administration is the public tender process. The creation of new tenders often requires consideration of previous tenders already

published in the specific target area. Domain experts confirm that they have to spend a lot of time to achieve the broadest and most correct knowledge possible without any support tool.

To address this gap, this research focuses on applying generative technologies to a real-world problem by proposing a general framework for suggesting similar public procurements. In particular, the adoption of Retrieval-Augmented Generation (RAG) [94] is proposed—an approach in NLP that combines information retrieval techniques with text generation. A RAG-based system leverages information retrieval results to enhance text generation, thereby enabling the production of more accurate and relevant responses to user queries.

A proof-of-concept study related to ANAC public tenders (Chapter 3.2.2) demonstrates the feasibility of the proposed approach by presenting the initial results. Additionally, an analysis of execution time, a relevant issue in LLMs, is included. All phases of the study and evaluation were supervised by domain experts in the legal field.

High-resolution images for this chapter are available at https://bit.ly/4eRONLT. The source code for the experiments conducted in this chapter is available in the repository: https://github.com/roberto-nai/PhD-THESIS.

5.2 Experimental Setup

Since the goal can be classified as a *question answering* task [40], the method was based on the RAG approach. RAG is a novel approach in the field of NLP that combines the strengths of IR and advanced language generation technologies. This technique is designed to enhance the capability of LLM in generating responses that are not only relevant but also contextually accurate [188]. The core of a RAG system is the integration of two components: a *retriever* and a *generator*; the retriever is responsible for fetching relevant pieces of information or documents from a database or corpus based on a given query, while the generator, typically an LLM, synthesizes the retrieved information and the original query to produce a coherent response [188]. For a deeper understanding of the RAG architecture, refer to [120]. By augmenting the generative capabilities of LLMs with the accuracy of information retrieval, RAG systems offer a valuable tool for a wide range of applications, including response systems, which is the focus of this research.

Dataset filtering

Following the suggestion of domain experts involved in the case study, the focus was directed toward a significant subset of the ANAC dataset already described in Chapter 3.2.2. Specifically, the public tender process from 2016 to 2022 in Northern Italy, encompassing

eight regions[1], was considered. This choice was motivated by qualitative and quantitative aspects related to the number of procurements and their heterogeneity, allowing for consistent data collection. The total number of procurements extracted from the ANAC catalogue after filtering for the eight regions of Northern Italy from 2016 to 2022 is 992, 137, forming the dataset on which this research is based.

General framework

The general framework of the present research includes several steps, which can be summarized by Figure 5.1. ANAC data filtering and processing allow us to obtain the texts of interest and the metadata necessary to set up the system. Subsequently, the embedding phase transforms the texts into vectors. The following steps concern contextual research, which includes the user formulating a question to query the reference data. In the example in the Figure, a public official (user) is looking for procurements related to purchasing syringes for public hospitals. Based on the query, the search in the database is carried out by the question system that transforms the textual query in an embedding and, following the best similarity score, retrieves and returns the texts (procurement object) to the user as an answer. Finally, the system proposes a ranking of the most relevant documents returned to the user, which will assess their relevance.

Processing steps

The initial step of the framework involves the collection and pre-processing of relevant data, which are appropriately filtered to serve as the basis for the RAG system. The refined dataset is then stored in the underlying database, a search engine serving as the retrieval component of the RAG system. The Open Data, filtered by region and year (see Section 5.2), were saved into this database. The process of indexing the dataset is enhanced by *embeddings*, which transform the texts into a high-dimensional space, improving the retrieval process's efficiency through semantic similarity.

Following the configuration of the retriever, attention shifts to the generator component, which relies on a LLM. This step is critical for generating coherent and contextually appropriate responses by leveraging both user input and the texts retrieved by the database. Integrating the retriever and generator represents a key phase in the RAG pipeline, combining the information retrieved by the database with the natural language processing capabilities of LLMs.

[1]The regions involved are: Aosta Valley, Piedmont, Liguria, Lombardy, Emilia-Romagna, Veneto, Friuli-Venezia Giulia, Trentino-Alto Adige

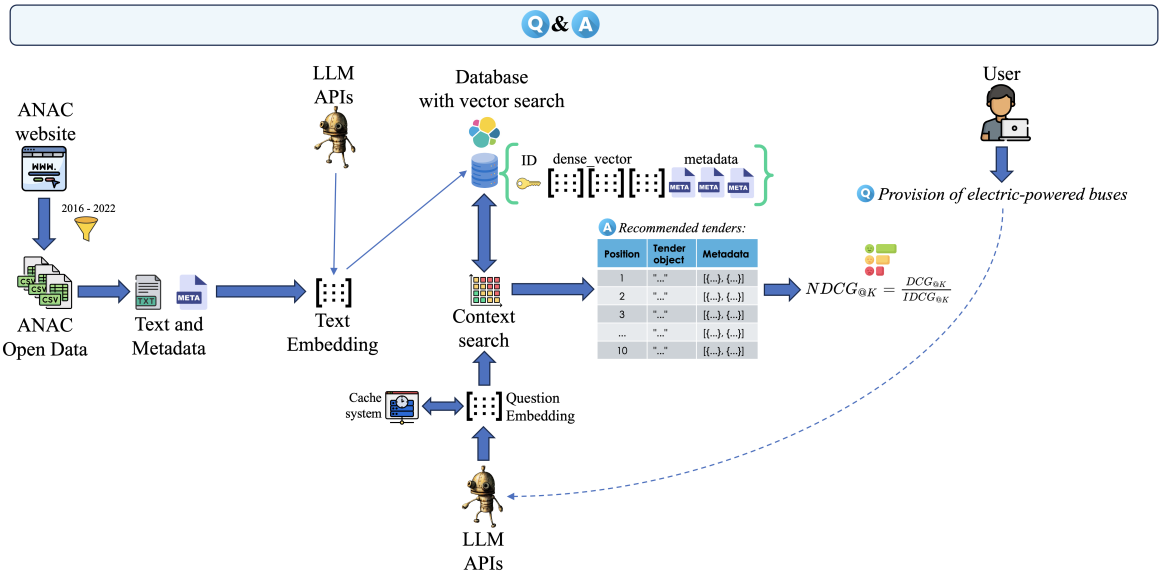


Fig. 5.1 Pipeline adopted for the question answering task with LLM; the workflow starts from the left with the collection of data, then the creation of the embeddings, the saving in the database and then the presence of a question (query) and the search for a contextual response

The final steps involve fine-tuning and evaluating the RAG to ensure it meets specific performance criteria. For this purpose, the proposed results were evaluated based on their position and a user relevance score derived from the metadata of the recommended tender (such as tender type, sector, CPV division, etc., as described in Section 3.2.1).

Evaluation

To ensure the effectiveness and relevance of the approach, the system was evaluated using the *Normalized Discounted Cumulative Gain* (NDCG) metric [265]. This evaluation method allows quantitatively assessing how well the system ranks recommended items in terms of their relevance to users. Two key pieces of information were used to evaluate the NDCG: the ordered list of recommended items and the relevance values for each item. The recommended items are ordered according to their relevance in the system (using the typical cosine similarity measure [106]), and a relevance value is entered for each item to reflect its importance or usefulness to the user.

To calculate NDCG, the *Discounted Cumulative Gain* (DCG) is first computed using the formula $DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2(i+1)}$, where rel_i is the relevance of the item at position i, and k is the number of top items considered. The *Ideal DCG* (IDCG) is the maximum possible DCG obtained by arranging the items in the perfect order of relevance. NDCG is then the

ratio of the actual DCG to the IDCG, normalized in the range between 0 and 1, calculated as $NDCG@k = \frac{DCG@k}{IDCG@k}$. This metric measures how effectively the recommendation system ranks the different items according to their relevance.

Technology

In terms of technology, the LangChain framework[2] was used to implement the RAG system. LangChain facilitates the connection between information retrieval capabilities and the generative capacity of LLMs by reducing coding time. Elasticsearch [87] was chosen for adoption, as it is the most widely used search engine software that also supports dense vectors [273] (a generalization of embeddings generated via LLMs), according to the DB-Engines Ranking of Search Engines[3] as of September 2024.

The calculation of the $NDCG@k$ index can be performed using the Python library scikit-learn[4]. As LLMs rapidly scale up to gigantic models [222], deploying them as the core of RAG systems has become challenging. This is primarily due to the computational expense of LLMs, which makes their implementation on standard hardware often unfeasible. To address this issue, many companies, such as Cohere (CAI)[5] and OpenAI (OAI)[6], have begun providing access to their proprietary LLMs through various APIs.

The framework proposed here adopts the Cohere and OpenAI APIs, which support bulk calls of up to 96 and 2048 texts per call, respectively. This capability is particularly relevant when a large amount of data needs to be transformed into embeddings. Cohere offers the ability to generate embeddings with two dimensions: `embed-multilingual-v3.0` (1024) and `embed-multilingual-light-v3.0` (384), while OpenAI offers three embeddings: `text-embedding-ada-002` (1536), `text-embedding-3-small` (1536), and `text-embedding-3-large` (3072).

During the indexing phase, each of the five models mentioned above was tested. Subsequently, 10 queries were executed for each model used in the RAG, and the top 10 responses were compared based on the NDCG. In addition to the RAG system, queries were also executed on a textual version of the database, which served as a performance baseline; once again, the top 10 responses obtained from the database algorithm (based on BM25 [201]) were evaluated against the NDCG.

To avoid computational waste, a small cache system was implemented to save user query embeddings and prevent recomputation after crashes or freezes. A native LangChain

[2]https://www.langchain.com
[3]https://db-engines.com/en/ranking/search+engine
[4]https://scikit-learn.org
[5]https://cohere.com
[6]https://openai.com

functionality stores embeddings on the local file system, thereby minimizing API calls. Exploring other cache system solutions is beyond the scope of this research.

5.3 Overall Results

5.3.1 LLM results

Table 5.1 presents a comparative analysis of retrieval model performance using the NDCG metric at a cutoff of 10. The columns in the table represent the results obtained in the various implementations of the RAG: BM25 using text mode; $\text{Cohere}_{\texttt{light}}$ and $\text{Cohere}_{\texttt{large}}$ using respectively `embed-multilingual-light-v3.0` and `embed-multilingual-v3.0`; $\text{OpenAI}_{\texttt{ada-002}}$ using `text-embedding-ada-002`; $\text{OpenAI}_{\texttt{3-small}}$ using `text-embedding-3-small`; $\text{OpenAI}_{\texttt{3-large}}$ using `text-embedding-3-large`. Table 5.2 shows the query number 10 and the results obtained in response from the RS with the OpenAI`3-large` model; the query is based on a "green" search based on CPV division number 34 (transport equipment and auxiliary transport products), which may include electric vehicles and other environmentally friendly transport solutions.

An initial observation indicates that $\text{OpenAI}_{\texttt{3-large}}$ model consistently outperforms the other models with the highest average NDCG@10 score of 0.848, highlighting its superior ability to rank relevant documents at the top of the search results. $\text{OpenAI}_{\texttt{ada-002}}$ model has the second-highest average score of 0.837, underlined in the table, suggesting that while it is less effective than the $\text{OpenAI}_{\texttt{3-large}}$ model, it still significantly improves upon the baseline BM25 and the $\text{OpenAI}_{\texttt{3-small}}$ model. Regarding Cohere, neither model outperformed the OpenAI models in the task; finally, it can be seen that $\text{Cohere}_{\texttt{light}}$ model (which has the smallest dimension) does not outperform BM25; this result was also highlighted in [108].

In any case, while historically robust, the BM25 model shows its limitations in this comparison, especially against the more sophisticated models provided by LLMs.

Overall, the results imply that more advanced and larger models tend to yield better relevance ranking performance, though the degree of improvement can vary by query and models adopted. This suggests a nuanced landscape where model selection for information retrieval tasks should consider the trade-offs between computational resources and the incremental gains in retrieval effectiveness.

Table 5.1 Results (NDCG10) for every experiment with Cohere (CAI) and OpenAI (OAI) embeddings. The best score is marked in **bold** and the second best is underlined

Query	**BM25**	**CAI**$_{\texttt{light}}$	**CAI**$_{\texttt{large}}$	**OAI**$_{\texttt{ada-002}}$	**OAI**$_{\texttt{3-small}}$	**OAI**$_{\texttt{3-large}}$
1	0.762	0.775	0.879	0.789	0.833	0.832
2	0.803	0.762	0.842	0.822	0.814	0.837
3	0.749	0.661	0.845	0.785	0.806	0.701
4	0.732	0.901	0.774	0.909	0.925	0.968
5	0.775	0.631	0.817	0.836	0.739	0.799
6	0.900	0.811	0.772	0.846	0.813	0.854
7	0.918	0.816	0.844	0.825	0.833	0.788
8	0.768	0.912	0.795	0.909	0.844	0.916
9	0.823	0.778	0.786	0.770	0.827	0.905
10	0.824	0.872	0.877	0.881	0.843	0.880
Avg. NDCG@10	0.805	0.792	0.823	0.837	0.827	**0.848**

5.3.2 Evaluation

The evaluation of the proposed approach was conducted qualitatively by legal domain experts who were involved in all stages of the design process. In the proof-of-concept, the tool was developed to address a specific need of legal practitioners: to provide comprehensive support in the preliminary stage of the legal process by offering a complete picture of the legislation related to a particular legal sector or topic. Specifically, the analysis of the results generated by the automated system indicated that the system functioned effectively, delivering relevant and accurate suggestions. The tool proved to be reliable and, notably, faster compared to the traditional process, which often involves manual searching, frequently relying on paper documents and risking incompleteness and partiality. Domain experts value the potential of such a tool to assist in their daily work while also emphasizing the importance of human oversight over the results provided.

A quantitative evaluation has not been conducted due to the limited number of cases in this proof-of-concept. An indicator-based evaluation is planned for a future, expanded version of this research.

5.3.3 Timing of the experiments

Observations show that API interactions are mostly stable, with occasional service request errors. Each API call consistently took around 300ms across both companies tested. Service errors and latency times can be influenced by factors such as input length and server workload,

Table 5.2 Query number 10 "Provision of electric-powered buses" results from $OAI_{3\text{-large}}$

Result	Answer	NDCG@10
1	Supply of plug-in electric buses	0.995
2	Supply of short plug-in electric buses	0.991
3	Supply of 6 electric buses and jest mini plug-in	0.977
4	Bus supply of long electric buses plug-in	0.958
5	Bus supply of regular electric buses plug-in	0.955
6	Supply of medium electric buses plug-in	0.953
7	Supply of 3 city buses; electrically powered	0.953
8	Supply of plug-in electric articulated buses	0.952
9	Supply of class II natural gas-powered buses	0.671
10	Construction of charging infrastructure for overnight plug-in electric buses and related maintenance at the depot	0.392
Avg.		**0.880**

which vary over time. The `text-embedding-3-large` model took the most time for both creating embeddings and computing results. The total size of the database with embeddings is about 120 GB. All the computations were carried out by an ARM architecture-based chip with 3.2 GHz speed (10-Core CPU / 24-core GPU) and 32 GB of RAM.

5.4 Conclusions

This chapter outlines the creation of a decision support system designed to help legal practitioners manage increasing data volumes. A real public procurement dataset was utilised to combine a recommendation system with generative model technologies in a proof-of-concept study. The results demonstrated the feasibility of applying this approach within the legal field. Although the implementation requires considerable computational effort, the performance improvements offer significant value to domain experts.

The conclusion addresses a specific request from domain experts for a user-friendly application that can run on commonly available equipment. The proposed solution is to develop a web-based application to manage the results, allowing for ongoing testing and feedback to support continuous improvement. Future work includes the implementation of a service using Flask[7] and ReactJS[8] to enhance usability through a web-based interface.

Regarding further developments, additional evaluation metrics beyond NDCG, are planned, along with a more detailed qualitative evaluation by domain experts, incorpo-

[7] https://flask.palletsprojects.com/en/3.0.x
[8] https://react.dev

rating more queries for system testing. From a technological perspective, a comparison between the proposed RAG framework and state-of-the-art solutions, such as Fast-RAG[9], is planned. The intention is to explore the use of open-source LLMs, like LLaMA [239], as an alternative to proprietary tools such as Cohere and OpenAI.

Finally, since the general framework and proof-of-concept results in the legal domain have shown the tool's effectiveness, testing in other domains, such as healthcare, is considered a future direction.

[9]https://github.com/intellabs/FastRAG

Chapter 6

Enhancing E-Learning: A Process Mining Approach for Short-Term Tutorials

6.1 Introduction

E-learning systems are digital platforms facilitating online education and training, delivering educational content and interactive experiences via the Internet. These systems support remote learning, often incorporating multimedia elements, assessments, and collaboration tools to enhance the educational experience. Moreover, they can record data on the activities carried out by students during their learning process.

Several previous studies have investigated the adoption of automated techniques to extract knowledge from information system (IS) data and improve learning knowledge ([107]). For instance, it has been demonstrated how helpful information can be extracted to identify learning patterns and provide recommendations on what to study next ([72]). This body of research has primarily focused on learning activities having a relatively long duration, typically several months or years. However, short tutorials on specific topics are becoming increasingly popular via the Web. These short tutorials make it possible to introduce complex topics in learning units of only a few hours. These flexible tools can adapt themselves to individual students' schedules and learning preferences.

Despite their increasing popularity, their analysis and assessment have received little attention in the literature. In particular, based on the available research, no previous study has investigated the feasibility of applying evidence-based educational data mining techniques to data generated from these tutorials. When considering complete courses or, anyway, long-term learning activities, the analysis can usually rely on a large volume of data, encompassing multiple modules, assignments, and exams, which support the extraction of detailed and

nuanced insights into student behavior and learning patterns. When analyzing short tutorials, the data volume is usually much smaller, often limited to a single session or a few interactions. This can make it easier to analyze but may provide less comprehensive insight. This research aims to fill this gap. More precisely, it aims to investigate the feasibility of leveraging EDM techniques to provide valuable insights into the learning processes underlying short tutorials. In particular, it proposes a methodology for collecting data in an IS by tracking user behavior to evaluate short tutorials. The proposed methodology is capable of collecting timed events, which can then be analyzed using techniques from the PM discipline, namely process discovery, variant analysis, and PPM. This aligns with the broader goal introduced in Chapter 1 of exploring the application of PM and PPM techniques in different domains, specifically focusing on educational environments. The research highlights how these methods, previously employed in legal data contexts, can be generalised to enhance the understanding of teaching and educational processes.

Such process-aware analysis techniques have already shown their benefits in a variety of different learning processes and analysis objectives in the Educational PM discipline [21]. To showcase the capabilities of the methodology, a case study was conducted through the creation and administration of an instructional tutorial employing Web-based technologies capable of collecting data during the learning process for a short period of time (approximately one to two hours) addressing the learning process of about 250 Italian students in an introductory programming course.

The first research objective involves the automated extraction of useful information in short-term learning process. Special attention has been paid to analyse learning outcomes to identify success factors based on the educational journey. In particular, it aims to uncover students' activity flows, considering both aggregated process indicators and process variants. A second research objective concerns the ability to predict the tutorial's outcome starting from the initial steps.

The methodology concretely supports the above-mentioned research objectives. First, it allows us to mine students' learning processes and uncover potential relations between different variants and learning outcomes. The corresponding research question concerns extracting students' activity flows in a short-term learning process and assessing whether there is any relation to the outcome (RQ1). Second, focusing mostly on prediction, the data from the case study have been analysed to estimate the relationship between behavior in performing the tutorial and learning. The research question concerns the ability to predict students' outcomes at different stages of the learning process in short-term tutorials (RQ2).

To validate the reliability of the obtained insights, the teachers involved in administering the tutorial, as domain experts, contributed to the discussion of the results and added their comments directly to various parts of the case study's results analysis.

In the remainder of this chapter, the case study is introduced in Section 6.2 and the methodology is presented in Section 6.3. The results are then examined in Section 6.4, followed by a discussion in Section 6.5. Subsequently, the related work is reviewed and compared in Section 6.6. Finally, Section 6.7 concludes the chapter.

High-resolution images for this chapter are available at https://bit.ly/4eRONLT. The source code for the experiments conducted in this chapter is available in the repository: https://github.com/roberto-nai/PhD-THESIS.

6.2 Case study

A case study based on a web tutorial to track students' behavior during their learning process is discussed. The tutorial has been administered to two groups of undergraduate students enrolled in the second year of their degree programs in management and economy. The educational context is that of an introductory computer science course, with a class of students homogeneous in age group and fairly evenly distributed by gender. The students are prompted by the teachers administering the tutorial to take the course individually and, in general, are left free to possibly interact, with no control over their individual behavior. This work expands on the idea presented in a previous paper that illustrated the framework and early results of the application to 70 students while also proposing the possibility of a qualitative follow-up with ethnographic research ([158]).

Tutorial web The tutorial consists of 10 web pages on topics related to learning Python programming. Each page is a self-contained introductory lesson that can be performed without prior knowledge. A multiple-choice question from the 10 pages tests whether students follow and learn the topics covered. In the end, the sum of the correct answers suggests whether they are learning well or poorly (see Section 6.3.3).

To investigate the learning sequence of the subject taught, the tutorial proposes three different paths. Figure 6.1 represents the sequence of lessons proposed in the tutorial according to the Business Process Model and Notation (BPMN) [259]. The following three lessons out of ten are the same for all students: the final one (FUNCTIONS), as well as the first three (INTRODUCTION, FIRST PROGRAM, and VARIABLES). After the third lesson about variables, the student can choose one of three learning paths in which the order of the three topics presented changes.

In particular, the following six lessons are included in three topics (i.e., three topics of two lessons each):

- **DATA TYPES**: a lesson to introduce different data types (TYPES) and a lesson on the conversion between different data types (CONV);
- **DATA STRUCTURES**: a lesson on the collection of similar data items (LISTS) and a lesson about unordered collection of keys and values (DICTS);
- **CONTROL STRUCTURES**: a lesson to execute a block of code according to specific conditions (IF_ELSE) and a lesson to repeat a block of code until a specific condition (FOR).

Based on current understanding, the *Track1* (DATA TYPES, DATA STRUCTURES, CONTROL STRUCTURES) is the most typical order according to contemporary computer science manuals. Nevertheless, in the tutorial, learners can follow one of two other tracks with the same contents but in a different order: *Track2* (DATA STRUCTURES, DATA TYPES, CONTROL STRUCTURES) or *Track3* (CONTROL STRUCTURES, DATA TYPES, DATA STRUCTURES). For this specific analysis objective, *Track1* is the benchmark for comparison. *Track3* is the furthest from the ideal path, according to the domain experts involved.

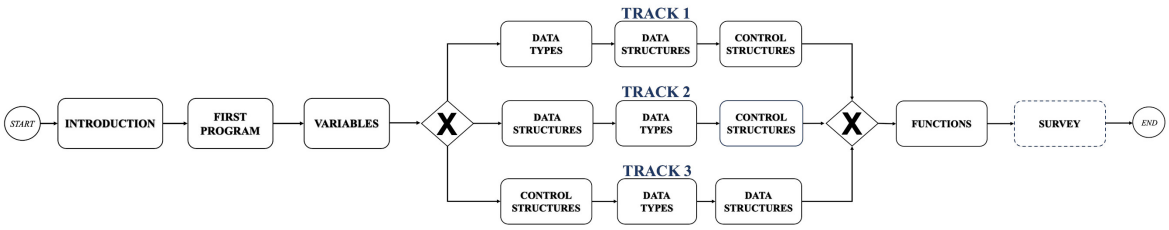


Fig. 6.1 The three different learning paths; on the top, the standard learning flow (*Track1*)

6.3 Methods

The methodology adopted in the present work includes several phases, as summarized in Figure 6.2. The first phase concerns the *design* stage, which involves defining the tutorial's content and quizzes. In the second phase, the *data collection* involved the development and administration of a web tutorial that incorporated the previously defined content and quizzes. The student activity monitoring system ensures that all interactions with the tutorial are recorded accurately. In the *event log construction* phase, the collected tracking data into an event log. This process involved structuring the raw data into a format suitable for analysis,

allowing us to trace each student's learning through the tutorial. In the *variant analysis* phase, analysed the event log to identify how students interacted with the tutorial and identify patterns that might influence learning outcomes. Finally, in the *outcome prediction* phase, applied predictive analytics to the event log data. This involved using the insights gained from the variant analysis to predict future outcomes, such as student performance and areas where students might struggle. The predictions made in this stage aimed to inform educators and improve the tutorial's effectiveness.

Each phase - design, data collection, event log construction, variant analysis, and outcome prediction - was integral to the comprehensive evaluation and enhancement of the web tutorial.

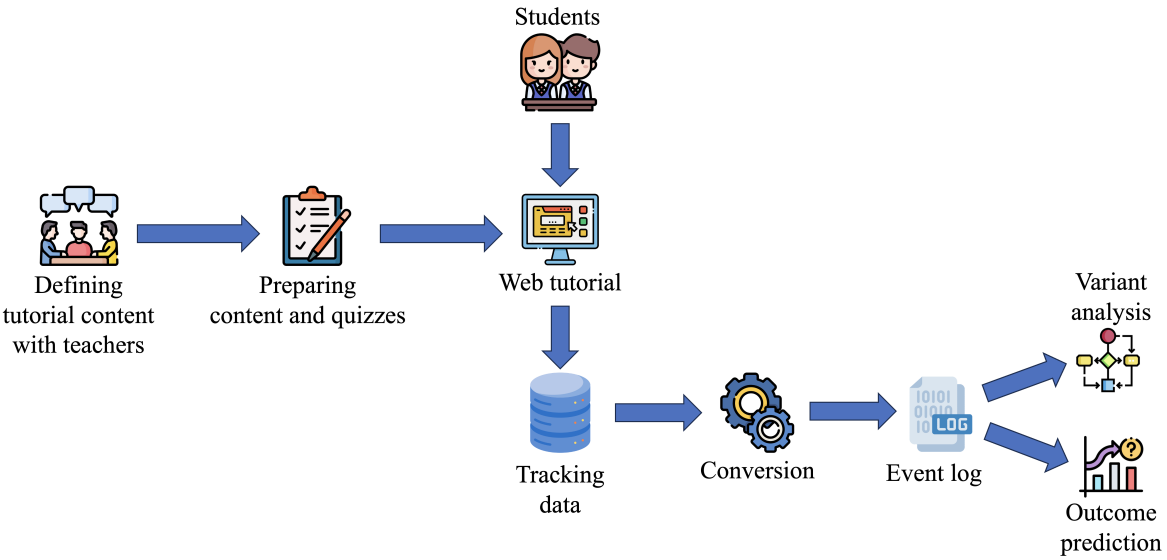


Fig. 6.2 Summary of the case study's phases: definition of the tutorial content, web portal for the administration of the content, activity tracking, convert tracking into an event log, variant analysis, and prediction on the event log

6.3.1 Web-tracking and technologies

Web behavior tracking. Each page/lesson is divided into 3-4 paragraphs describing the topics. While browsing the tutorial, the following events were tracked within the web pages: PAGE-IN (enter the page), PAGE-OUT (exit the page to access the following lesson or quiz), MOUSE-IN (the mouse pointer moves onto a new paragraph element), MOUSE-OUT (the mouse pointer moves out of a paragraph element), CLICK (a student clicks on a paragraph element), DBCLICK (a student double-clicks on a paragraph element). Each page also tracks movement between paragraphs (numbered 0 to 3). E.g., INTRO_MOUSE-IN_2 means

that the student has entered with the mouse in the second paragraph of the introductory lesson. In addition to the events, the following data are recorded: the session-id (to trace the activities back to a specific browser session), the name of the page (INTRODUCTION, FIRST PROGRAM, VARIABLES, etc.) on which the events occurred, the result of the quiz on that page, the *Track* on which the events occurred (1, 2 or 3), and optional data provided by the students at the time of the final survey (e.g., tutorial evaluation).

Technologies. The tutorial described in Section 6.2 has been implemented by adopting the following technologies to track behavior in web pages during the course of the tutorial. The *front-end* (the graphical user interface) and the contents of each tutorial page are written in HyperText Markup Language (HTML), using the open-source frameworks Bootstrap[1] for the layout and jQuery[2] for event tracking, based on ([64, 32]). The *back-end* programming language for track functionality called up via jQuery is PHP[3] while the Database Management System (DBMS) for storing and retrieving tracking data is MySQL.[4] For web tracking, the following steps were performed: a session-id was created when a student first accessed the web tutorial; all sections (paragraphs) of the tutorial pages were labelled in HTML and, based on the student's interaction with the pages, a jQuery function performed an asynchronous call to the server to track the interaction, saving the DBMS table using the session-id as the key entry. The web tutorial is available online.[5]

6.3.2 Event log construction

The starting point for PM research is an *event log*, representing an extraction from the IS on the execution of activities in a process ([249]). An event log includes a set of *traces*, whereas each trace stores a sequence of *events*, each representing the execution of an *activity* occurred at a given *timestamp* during a process, possibly together with some additional data (e.g., the resource who performed the activity) ([249]). Every trace is identified using a so-called *Case_ID*, which is the session-id assigned automatically by the web browser to a student when navigating with the browser within the tutorial. The fields Activity and Timestamp are the data traced by jQuery during the tutorial execution. Additional information, such as data provided by students, is added for each track (e.g., track choice). To focus on the learning flow, four kinds of activities have been studied: the entrance and exit on a web page (PAGE-IN, PAGE-OUT) and the entrance and exit in each paragraph of the page (MOUSE-

[1] https://getbootstrap.com
[2] https://jquery.com
[3] https://www.php.net
[4] https://www.mysql.com
[5] http://webtutorial.altervista.org/python

IN, MOUSE-OUT). The number of CLICKs and DBCLICKs were instead used as trace features.

Different groups of traces in the event log are called *process variants* (*variants*) since they represent alternative ways to execute the same process (i.e., users may perform activities in a different order before the end). After exploring the dataset, some traces configured as outliers can be removed if they appear to be wrong or not harmonious, such as in the case of processes that are too short according to domain experts (e.g., students who perhaps only opened the initial pages without proceeding with the tutorial). To focus on the most significant cases, consider students having completed the tutorial in between 5 minutes and 2 hours and a half.

6.3.3 Outcome analysis

A relevant analysis in the present work concerns the distinction between students who performed well and poorly. To have an indicator of whether students have understood the tutorial content, rely on the 10 answers given to the quizzes between each page, counting one point for each correct answer. The distribution of the results can be divided into two parts through the median class as a threshold; in the case study, the median value obtained from the quiz results used as a threshold to distinguish the two parts is 0.7. As discrete data/classes, two groups of almost equal size. Finally, in the current proof-of-concept, processes with a *negative outcome* (OUT-NEG) are defined as all cases below the threshold, while processes with a *positive outcome* (OUT-POS) are defined as those above the threshold. The value of each student's outcome has been saved in the event log.

6.3.4 Process mining techniques

Variant analysis (RQ1) A first exploration involves the analysis of event logs and diagrams obtained from process discovery. According to RQ1, both the complete log and the individual processes of interest have been inspected. First, investigate the students' performance concerning the corresponding learning outcomes (Section 6.4.2). In particular, intend to analyse the log according to several dimensions to identify interesting behaviors. The variants will be considered in relation to the overall completion time of the tutorial, the time spent on each page, and the student's movements between paragraphs and pages of the tutorial. also proceeded with an *automated* comparison aimed at quantifying the existing differences. More precisely, this work, apply the approach proposed by [23], which takes into account both *behavioral* and *context* process similarity. The first one considers how activities are executed in the compared executions. The second one takes into account the context in which

the executions occur, defined using the data attributes stored in the event log. The approach takes as input two event logs corresponding to the set of executions to compare. Then, it computes the differences among them in terms of behavior or context and builds a transition system representing the behavior of both variants, where states or edges showing relevant differences between the two variants are annotated accordingly. Note that these annotations are visualized using different colours and thicknesses of the transition system elements.

Second, compared the three discovered processes about *Track1*, *Track2*, and *Track3* to investigate any differences of interest (Section 6.4.3). The first analysis concerns tracking timing and outcome. Second, focus on the times between individual lessons in the three tracks. Finally, examine the backward jumps between paragraphs and pages in each track (intended as the action of returning to a previous section or page of the tutorial). intend to examine this behavior as it may indicate a desire for better learning or distraction, according to the domain experts.

To analyse the event log, used academically licensed DISCO from Fluxicon[6], as well as ProM[7] to perform the automated variant analysis (*process-comparator* module).

Predictive Process Monitoring (RQ2) PPM ([133]) is a branch of PM research that aims to predict the future development of ongoing process cases given their uncompleted traces. According to the RQ2, aim to predict students' performance based on the learning process taken by the students in earlier stages (an *outcome-based* prediction). Figure 6.5 summarises the phases of the PPM exploration. First, from the complete event log, refer to the sequences of events recorded up to a certain point in time during the execution of a process. These partial event sequences are called *prefixes* to be used for predicting the future behavior of the process. In the training phase on machine learning models, *prefixes* extracted from the traces of the event log ([54]) become vectors according to different encoding techniques. The research used *Index Encoding* (IE), *Boolean Encoding* (BE) and *Frequency Encoding* (FE) ([54]) methods to verify which one leads to better results with the available data. In particular, in IE, each feature corresponds to a position order in the sequence, and the possible values for each feature are the event classes; BE represents a sequence through a feature where an event is indicated by 1 if it occurred in the prefix, 0 otherwise; FE represents the control flow in a case with the frequency of each event class in the case. Figure 6.3 describes an example of the three encodings; the IE (Figure 6.3.a) includes the sequence of events that occurred for each Case ID (e.g., for Case 'ID01', the first event occurred IF_ELSE_PageIN_0), and the third is IF_ELSE_MouseIN_1). BE (Figure 6.3.b) assigns 1

[6]https://fluxicon.com/disco
[7]https://promtools.org

for events that occurred and 0 for those that did not occur for each Case ID (e.g., for Case 'ID01', the event IF_ELSE_PageIN_0 occurred while IF_ELSE_MouseOUT_1 did not). Finally, FE (Figure 6.3.c) includes the frequency with which the events occurred (e.g., for case 'ID01', the event IF_ELSE_PageIN_0 occurred 3 times).

Case ID 🔑	**Event 1**	**Event 2**	**Event 3**	**...**	**Event *n***
ID01	IF_ELSE_PageIN_0	IF_ELSE_MouseIN_1	IF_ELSE_MouseOUT_1	…	…
ID02	IF_ELSE_PageIN_0	IF_ELSE_MouseIN_2	IF_ELSE_MouseOUT_3	…	…
…	…	…	…	…	…

(a) – Index Encoding (IE)

Case ID 🔑	**IF_ELSE_PageIN_0**	**IF_ELSE_MouseIN_1**	**IF_ELSE_MouseOUT_1**	**...**	**Event *n***
ID01	1	1	0	…	0
ID02	1	1	0	…	1
…	…	…	…	…	…

(b) – Boolean Encoding (BE)

Case ID 🔑	**IF_ELSE_PageIN_0**	**IF_ELSE_MouseIN_1**	**IF_ELSE_MouseOUT_1**	**...**	**Event *n***
ID01	3	4	5	…	0
ID02	2	1	6	…	1
…	…	…	…	…	…

(c) – Frequency Encoding (FE)

Fig. 6.3 Examples of prefixes encoded with simple-index (a), boolean (b) and frequency (c)

Finally, supervised experiments are applied to these trace representations to obtain a predictive model. Such a model can then be applied to new partial traces. At runtime, predictions are made on incomplete traces. Since the research aims to make predictions as early as possible, focused on the subset of the prefix log with the initial part of the process, i.e. a length of 40 (which corresponds to the first page/lesson of the tutorial), 80 (which corresponds to the second page/lesson of the tutorial), or 160 (which corresponds to the third page/lesson of the tutorial). trained two single classifiers: Random Forest (RF) and eXtreme Gradient Boosting (XGB). The traces in input to classifiers are zero-padded to have a fixed length.

Figure 6.4 graphically shows a complete trace of length *n* (Figure 6.4.a) as well as the trace prefixes of length 1 (Figure 6.4.b), length 2 (Figure 6.4.c), and length 3 (Figure 6.4.d) with zero-padding.

Outcome prediction In terms of technology, used the open-source toolkit Nirdizati ([200]), which supports the various phases of the PPM just described.[8] Table 6.1 summarises the trace features used as input for the prediction models. The output is the binary classification between *positive outcome* (OUT-POS) or *negative outcome* (OUT-NEG).

[8]http://research.nirdizati.org

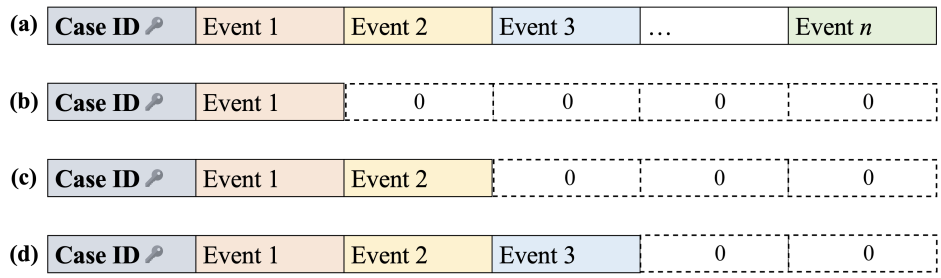


Fig. 6.4 Examples of prefixes starting from a complete trace (a) and related trace prefix of length 1 (b), length 2 (c), and length 3 (d), all with zero-padding to have the same length

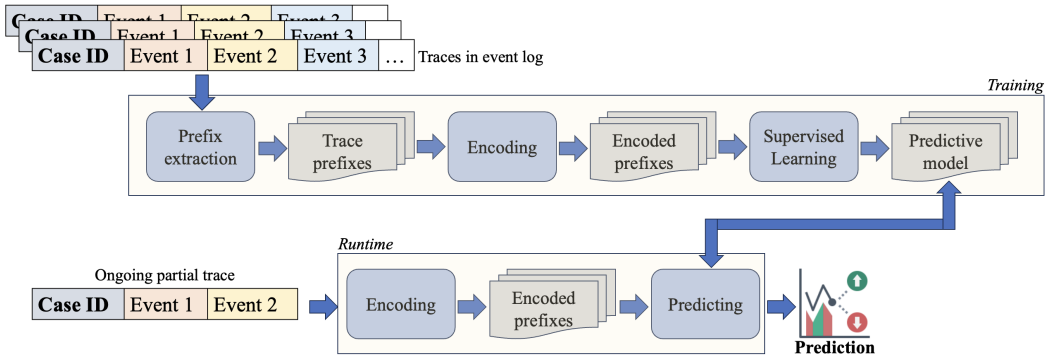


Fig. 6.5 Overview of the PPM exploration based on machine learning models

The prediction results are evaluated with *K-fold cross-validation* and *F1-Score*, i.e. the harmonic mean between *recall* and *precision* ([84]), and the *Area Under the Curve* (AUC) ([67]). The F1-Score metric is a unique measure of models' prediction performance with an imbalanced dataset ([31]), while the AUC metric is calculated by assessing a classification model's ability to distinguish between classes ([67]). The hyperparameters optimisation used by Nirdizati is Hyperopt ([16]).

The computations were carried out by an ARM architecture-based chip with 3.2 GHz speed (10-Core CPU / 24-core GPU) and 32 GB of RAM.

6.4 Results

6.4.1 Event log analysis

The complete dataset includes sessions from the tutorial administrations. Table 6.2 shows the main statistics of the event log: most students concluded the tutorial with a median duration of 37.8 minutes as well as an average duration of 41.3 minutes. The standard deviation (STD) of 28.9 minutes is quite relevant; in fact, there is considerable variability. In the most extreme

Table 6.1 Trace features used as input for the prediction models

Trace feature	**Description**
`event`	Encoded version of the event
`trace_total_time`	Total trace time
`track`	Tutorial track {1, 2, 3}
`student_age`	Age of the student
`student_gender`	Gender of the student {M, F, other}
`click_num`	Number of clicks
`dclick_num`	Number of double clicks
`page_jump_num`	Number of backward jump per page
`para_jump_num`	Number of backward jump per page paragraph

cases, some students completed the tutorial very quickly (5 minutes) while, on the contrary, a few students needed 2 hours to complete it.

Table 6.3 shows a snapshot of the resulting event log, with the three main properties in the event log (Case ID, Activity, Timestamp) and an example of the other features added as attributes of the traces, namely the type of track that the student travelled. According to the final survey, 82% of students expressed high appreciation for the tutorial. This seems relevant both to ensure the effectiveness of the proposed approach and to proceed with the examination of the results.

Table 6.2 Main statistics on the event log obtained from the tutorial: the first line shows the statistics of all cases, the second line those of cases with a positive outcome (OUT-POS), the third line shows those with a negative outcome (OUT-NEG). Times are expressed in hours (h), minutes (m), and seconds (s)

Cases	**Total**	**Median duration**	**Mean duration**	**STD**	**Min duration**	**Max duration**	**Backward jumps (avg)**
Complete	242	37.8 m	41.3 m	28.9 m	5 m, 8 s	2 h, 19 m	1.64
OUT-POS	130	42.7 m	45.8 m	32.8 m	5 m, 8 s	2 h, 2 m	1.79
OUT-NEG	112	32 m	36.1 m	21.7 m	5 m, 51 s	2 h, 19 m	1.56

6.4.2 Learning processes and outcome analysis

Analysis of the learning process' timing In terms of time analysis, focus on the overall duration of the learning process and the time spent on individual tutorial pages. The duration of the learning processes (median and average duration) clearly indicates that students with positive outcomes took longer. As summarized in Table 6.2, the median duration of the tutorial is about 46 minutes for students with positive outcomes, while students with negative

Table 6.3 A sample example of the event log including the activities of a single student identified with Case ID 'ID01', navigating to the IF-ELSE, FOR, TYPES, and LISTS web pages in *Track3* learning path ('Track' and 'Quiz' are trace features)

Case ID	Activity	Timestamp	Track	Quiz
ID01	IF_ELSE_PageIN_0	2023-03-29 17:38:50	3	0.8
	IF_ELSE_MouseIN_1	2023-03-29 17:39:13		
	IF_ELSE_MouseOUT_1	2023-03-29 17:39:23		
	IF_ELSE-Q_PageIN_0	2023-03-29 17:40:02		
	IF_ELSE-Q_PageOUT_0	2023-03-29 17:40:15		
	FOR_PageIN_0	2023-03-29 17:40:16		
	FOR_MouseIN_1	2023-03-29 17:40:20		
	FOR_MouseOUT_1	2023-03-29 17:40:24		
	FOR_MouseIN_2	2023-03-29 17:40:31		
	TYPES_PageIN_0	2023-03-29 17:46:44		
	TYPES_MouseIN_2	2023-03-29 17:46:57		
	TYPES_MouseOUT_2	2023-03-29 17:47:01		
	LISTS_PageIN_0	2023-03-29 17:55:41		
	LISTS_MouseIN_1	2023-03-29 17:55:42		
	LISTS_MouseOUT_1	2023-03-29 17:55:48		
	LISTS_MouseIN_2	2023-03-29 17:57:56		

Table 6.4 The duration (in seconds) on individual pages for the group of cases with a positive outcome (OUT-POS) and negative outcome (OUT-NEG)

	INTRO	PROG	VARS	IFELSE	FOR	TYPES	CONV	LIST	DICT	FUNCT
OUT-NEG	66	79,5	71	79	59	126	81	61	58.5	28
OUT-POS	120	120	138	138	116	294	150	150	82	58

outcomes, took a median time of 32 minutes. Concerning students with poor performance spending less time consulting computer tutorials could be attributed to several factors. A hypothesis is that these students may lack the necessary foundational knowledge to engage effectively with the tutorial content. As a result, they may rush through the material without fully understanding it, leading to lower performance outcomes [229].

The behavior of the students on individual pages with positive/negative outcomes is also analyzed. note how the top-performing students were slower for each of the ten pages, and identify a significantly longer median duration than those who performed poorly. As Table 6.4 highlights, times on pages are always broadly higher for the group that will get a successful outcome. The stay on the pages can often be more than twice as long. Interestingly, this behavior appears already in the first pages, suggesting a student's attitude that can thus be intercepted as early as the first part of the tutorial execution.

The movements between pages or paragraphs By observing the jumps between different pages or activities (i.e. paragraphs) during the course of the tutorial (Table 6.2), a meaningful difference in students' behavior in carrying out the tutorial. The average number of backward jumps in relation to the learning outcome has been computed. The group of students with positive outcomes appears to go back more frequently (1.79 jumps backwards on average), with less linear behavior, than those with a negative outcome (1.56 jumps backwards on average). A behavior perhaps aimed at improving content understanding, corresponding to a more reflective attitude.

Together with the previous observation about timing, the result seems to indicate that students with positive outcome focus more carefully on the content and return to topics already covered, while those with negative outcome proceed quickly towards the next paragraph, without going back very often and making sure they have understood the tutorial content.

Automated variant analysis results Finally, performs a statistical comparison of sub-groups' traces with positive and negative outcomes (as mentioned in Section 3.3, exploit *Process Comparator* plugin in ProM tool). Such a comparison allows to identify which parts of the tutorial appear relatively more significant. Figure 6.6 reports the obtained results, whereas the darker the colour tone, the stronger the statistical relevance of the difference between activities.

To provide an idea of this type of analysis, describe three cases that are of interest. First, the focus has been set on the frequency of activities. In Figure 6.6.a, the central paragraphs in the pages concerning conversions (CONV) appear relatively more frequent among students with positive outcome.

Second, in Figure 6.6.b, log comparison indicates that there are statistically significant differences in terms of duration for performing the activities in the section LISTS. Cases with positive outcomes spent more time, compared to negative ones, on the paragraphs related to LISTS learning.

Third, regarding the differences between activities with respect to the corresponding remaining times, the diagram in Figure 6.6.c shows that there are differences in the INTRO and PROG sections. Being the initial activities of the tutorial, this observation confirms what had already been found in the analysis about timing, namely that students with positive outcome take longer from the tutorial beginning to finish the activities.

These results illustrate the possibilities offered by this type of automatic analysis. Overall, these suggestions may indicate the parts of the tutorial to focus on to propose possible improvements.

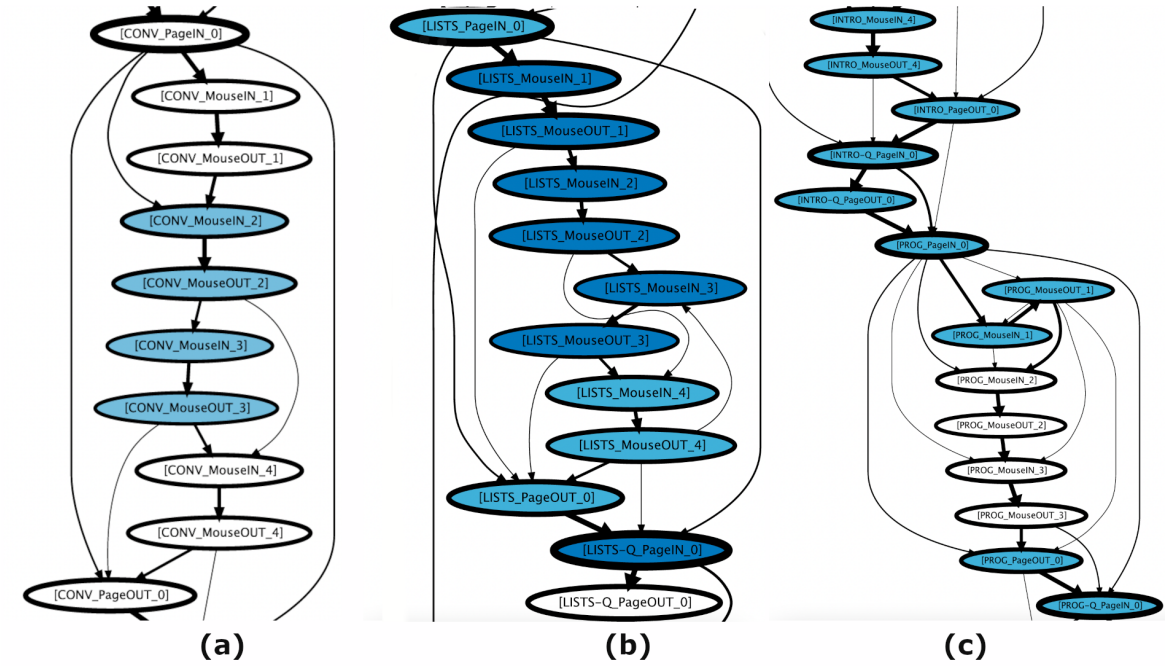


Fig. 6.6 The automated processes comparator analysis output of negative or positive outcomes, with respect to (a) trace frequency, (b) elapsed time, and (c) remaining time. Image generated with ProM

6.4.3 Analysis of learning tracks

The three learning tracks To explore the three learning paths, consider the main measures on time and performance. As summarized in Table 6.5, *Track1* is longer than the other two (median duration of 42.1 minutes), achieving better results (71.2% of correct answers). On the opposite, *Track3* is relatively shorter (29.9 minutes), with a lower performance (64.9% of correct answers). These results suggest the existence of some differences, to be examined in more detail in the next paragraphs by focusing on time and outcome.

Table 6.5 Students' performances in the three tutorial tracks, i.e. the number of cases, the mean, the median, and the STD in terms of minutes for each track

Track	Cases	Median (min)	STD (min)	Correct answers
1	33%	42.1	22.4	71.2%
2	28%	32.4	26.6	70.1%
3	39%	29.9	35.2	64.9%

Time analysis of learning tracks A further insight concerns the analysis of times between individual pages. focus on the central activities of the tutorial concerning the three topics

(of two lessons each) into which the flow described in Figure 6.1 is divided. As depicted by Figure 6.7, examine the time between pages of the three tracks, i.e. the pairs TYPES and CONV (DATA TYPES topic), IF_ELSE and FOR (CONTROL STRUCTURES topic), LISTS and DICTS (DATA STRUCTURES topic).

Two interesting regularities appear relatively evident. First, notice the regularity of a quickening towards the concluding activities in all tracks, regardless of track type. In fact, in each track, the initial topic always took longer than the others that followed in the exercise. Similarly, when the topic appears at the end of the track, it is always carried out faster. This phenomenon can be interpreted as a familiarity gained with the content of the tutorial or an indicator that the student gets bored and tries to go faster in the second part, regardless of the lessons he or she has to go through.

A second observation is that the order in which topics are presented affects the duration of the execution. Specifically, for the same content, the duration is different depending on whether it is presented earlier or later. For example, CONTROL STRUCTURES topics are performed more slowly if presented at the beginning (median duration of 5.4 minutes, in *Track3*) and much faster if presented later (2.8 minutes in *Track1* and 3 in *Track2*).

These recurrent activity flows, therefore, suggest presenting attention to the order of the activities, as the most important ones should be offered at the beginning of the short tutorial when attention appears highest.

Outcome analysis of learning tracks A joint examination of the three tracks' median duration and the outcome provides additional insights. Positive cases are always longer than negatives for each track, as mentioned. More interestingly, the median duration of *Track1* is always higher than *Track2* or *Track3*, both for cases with positive (44.2 instead of 39 or 43.2) and negative outcomes (39.6 instead of 31.6 or 26.8). This seems to imply that *Track1* favours a greater depth of contents.

Focusing on *Track3*, students with positive outcome had a very long average duration, almost equal to *Track1*, while in contrast, those with negative outcome were the group that went the fastest of all. A possible interpretation is that *Track3* forced those who wanted to achieve good results to pay more attention, while it accelerated the progress of the tutorial for those who were not motivated to achieve a good result.

The analysis of backward jumps (Table 6.6) confirms how *Track3* was the one that forced students to go deeper into the topics, regardless of whether the learning outcome is positive or negative. While the numerosity of the subgroups does not allow for generalization, these findings deserve to be further investigated, as they show the importance of focusing on the paths taken. A qualitative investigation would be necessary to understand the differences

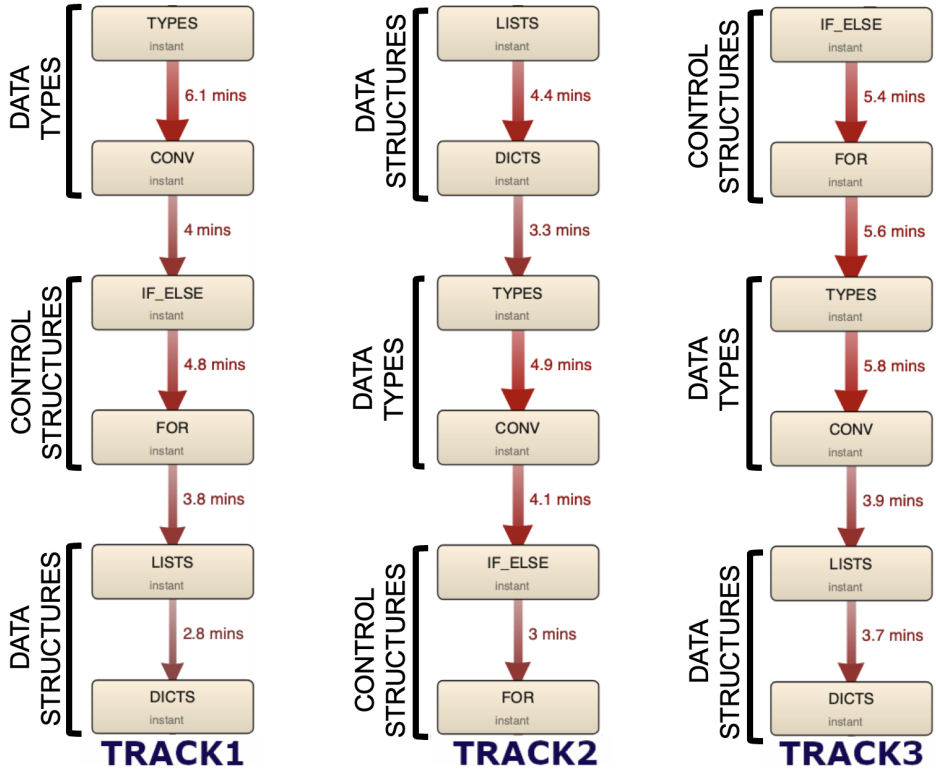


Fig. 6.7 Performance analysis (median duration between the activities) of the three tracks' central activities (the first part common to all tracks is not present). Image generated with Fluxicon DISCO

between the paths and evaluate the contents proposed by the learning track, which is out of the scope of the current work.

Table 6.6 Average number of jumps per page based on quiz result for the group of cases with positive outcome (OUT-POS) and negative outcome (OUT-NEG)

Cases	**Track 1**	**Track 2**	**Track 3**
OUT-POS	1.91	3.21	2.03
OUT-NEG	1.82	2.12	2.18

6.4.4 Outcome predictions

This Section describes the predictive model results to investigate the outcome of the pathway after the first part of the tutorial, according to the RQ2. As mentioned in Section 6.3.4, prefixes of lengths 40, 80, and 160 have been extracted to investigate the first half of the process whose outcome has to be predicted. Table 6.7 describes the results obtained from the XGB and RF models. The prediction results improve as the prefix size increases. Apart from the shortest prefix length (40), which gets poor results, already with a length of 80, the XGB model (better than RF) gets results of some interest. Interestingly, XGB with IE is somehow always better than RF. In the best case, before the midpoint of the student's online course, the final trajectory was predicted with about 70% accuracy using XGB with IE coding (F1-Score of 0.6721, Accuracy of 0.6741, Precision of 0.6846, Recall 0.6781, and AUC 0.7221); both AUC and F1 are consistent in defining the best classifier for each prefix.

Even though the algorithms are both ensemble types [56], it can be observed that RF performs better with FE encoding while XGB with IE encoding. These prediction results are not only quite satisfactory in themselves, but more importantly, they show a good possibility in the proof-of-concept, a sign that such an analysis can be done and at the same time provides a baseline from which to start and to compare with. In terms of time, the computation for training the machine learning models took about 30 minutes for the 40 prefixes to about 4.5 hours for the 160 prefixes (the most time-consuming optimization is that of the XGB).

Domain experts can analyse the prediction model's results and make other considerations. By using a prediction model, teachers can more timely identify students who might encounter difficulties in the tutorial. This allows them to intervene early and provide targeted support to improve students' performance. Knowing at-risk students allows teachers to adapt their teaching to meet the specific needs of these students. They can provide additional resources, offer individual tutoring sessions or change the pace of the course to ensure that at-risk students have a better chance of success. Focusing instructional resources on

Table 6.7 Prediction results: for each prefix and its relative encoding (BE, IE or FE), it is possible to compute the performance (F1-score, Accuracy, Precision, Recall) of each algorithm (RF or XGB). The best performance is achieved by XGB in the configuration with prefixes encoded in IE mode, while the best RF results are obtained from prefixes encoded in FE mode. **Bold** values are the cases with the best results for each prefix length, while the second best ones are underlined

Algorithm	Prefix len	F1-score	Accuracy	Precision	Recall	AUC
RF (IE)	40	**0.5677**	**0.5683**	**0.5702**	**0.5696**	**0.5788**
RF (BE)	40	0.5522	0.5547	0.5586	0.5569	0.5777
RF (FE)	40	0.4223	0.4737	0.47	0.4816	0.5409
XGB (IE)	40	0.5633	0.5652	0.5688	0.5672	0.594
XGB (BE)	40	0.5305	0.5318	0.5341	0.5335	0.5465
XGB (FE)	40	0.559	0.5596	0.5615	0.5609	0.5878
RF (IE)	80	0.549	0.5823	0.6458	0.5956	0.6695
RF (BE)	80	0.546	0.5496	0.5574	0.5547	0.5886
RF (FE)	80	0.5598	0.5643	0.5743	0.5699	0.6443
XGB (IE)	80	**0.6168**	**0.6184**	**0.626**	**0.6225**	**0.668**
XGB (BE)	80	0.5534	0.5538	0.5565	0.5561	0.5515
XGB (FE)	80	0.5895	0.5895	0.5907	0.5907	0.627
RF (IE)	160	0.5479	0.5509	0.5571	0.5549	0.5875
RF (BE)	160	0.5699	0.5727	0.5798	0.5767	0.6026
RF (FE)	160	0.6182	0.6191	0.624	0.6219	0.6619
XGB (IE)	160	**0.6721**	**0.6741**	**0.6846**	**0.6781**	**0.7221**
XGB (BE)	160	0.49	0.4928	0.4964	0.4965	0.5267
XGB (FE)	160	0.5041	0.5225	0.5376	0.5308	0.5585

students needing additional support can optimise teaching efficiency. Teachers can allocate more time and resources to these students, enabling them to maximise their educational impact. Using the model as a continuous assessment tool, teachers can continuously monitor student performance throughout the tutorial. They can timely identify changes in students' performance over time and adapt teaching strategies accordingly. Moreover, by analysing the predictive data provided by the model, teachers can assess the effectiveness of their tutorial and identify areas requiring improvement. They can then modify course content, teaching methods, or assessments to maximise students' success.

6.5 Discussion

This section will discuss the strengths and weaknesses of the approach, some reflections on the capability of PM, and the generalizability of the work.

Strengths and limitations of this work The teachers involved in administering the tutorial evaluated the results positively. From an instructional point of view, information about the learning pathways allows teachers to understand what is happening within the specific lessons. Sequential learning has been recognized as the winning strategy in most cases. In addition, speed of execution and a lack of desire to go deeper were recognized as key factors in learning failure. It is acknowledged that the study has some weaknesses. First, there is a lack of contextual knowledge, e.g., previous knowledge of programming skills from students involved in the tutorial (being non-computer courses, it assumed that almost everyone was ignorant on the topic - in any case, are interested in an aggregate/average measure, so outliers are smoothed). Second, do not discriminate between users with difficulties in using computers or interacting with technology. While assuming they are a minority in the study, this research tries to take this into account for future work. Third, the approach is focused on data that can be tracked by the information system. This means that the investigation of the cognitive dimension is not immediate. The qualitative analysis of the learner's educational context at the moment of the tutorial's administration is a common problem in other studies in the educational PM field. Finally, can improve the survey by increasing data requested by the students, e.g. demographic data.

Finally, this research has some concluding remarks on the technology's capacity adopted in this work. As the variation of events is relatively low, this has resulted in a limitation to the full utilisation of the PM's potential. Due to the lack of a wide variation of events, the insights generated in this work may not fully reflect the dynamics of the underlying processes. It's already been pointed out that the work focuses on control flow analysis and the automatic extraction of events recorded in the computer system. This analysis may result in a narrow view of the process, potentially leading to incomplete or distorted conclusions, and must be incorporated with contextual knowledge, as mentioned above, within the study's limitations. To address this issue, identify three main strategies that should be considered by future work aimed at leveraging PM techniques for this kind of analysis. First, there is a need for a diverse and comprehensive dataset. Future studies should aim to include a wider range of event types and instances to capture a broader spectrum of process variations; second, complement other analytical methods to provide a more holistic understanding of the process, e.g. through qualitative analysis; third, case selection should be carefully considered, including a diverse sample of cases in order to improve the applicability of PM and lead to more robust results; fourth, an iterative and integrated approach with domain experts (as suggested by studies on interactive PM) starting from the preliminary data collection and analysis stages to improve the richness of the dataset gradually.

Generalizability of the results Regarding the approach's generalizability, highlight that the methodology proposed to generate the event log can be easily applied to leverage PM on other web-based tutorials under the condition that they track similar kinds of data. argue that such conditions are easy to satisfy. The work is based on web technologies, which became a common way to offer self-learning tutorials. In addition, the results, intended as data, techniques and instruments are publicly accessible, thus the results can be replicated. Second, the proposed solution can be easily applied to a broader context, both as a type of user of the tutorial and as content. Short-term tutorials, in fact, can be adopted for various types of audiences, not only university students, as in the case. Furthermore, the contents can also vary, defining in a congruent way an adequate linguistic register for the description of the proposed contents.

6.6 Related work

This section provides an overview of related work, highlighting the main differences in the work to position it with respect to the state of the art.

The work falls within the stream of studies on learning with computer-based methods, which typically involve the measurement, collection, analysis, and reporting of data about learners and the context in which they occur. Such studies investigate students' actions through traces detected by e-learning systems in the context of LMSs ([243]). Courses based on Learning Management Systems (LMSs), such as Massive Open Online Courses (MOOCs). In the following, focus on work leveraging PM and machine learning techniques to model learning processes and predict their outcomes.

Learning processes and process mining. The recent discipline of PM concerns ideas, methods, and tools to extract knowledge from a time series of activities, i.e. event logs ([249]). The student's behavior can be explored in three directions: comparison of students' behavior, performance prediction based on students' behavior, and learning strategy evaluation ([261]). Several previous studies have already explored PM to improve educational processes ([85](. Process discovery techniques were also used to investigate students' different web behavior strategies in tackling quizzes in online tests ([105]). Similar to the work, the authors investigated the adoption of PM to analyse students' quiz-taking behavior patterns, but they focused on an LMS. In [146], the authors promoted a correlation study between the behavior of the learner (i.e. the number of connections between the sections of a course followed) during the learning process and their mark obtained on the final exam, starting from an event log obtained from the LMS. The study in [39] aims to discover the self-regulated learning processes of students in an e-learning course using PM techniques by applying the

Inductive Miner algorithm to interaction traces from 101 university students on the Moodle platform. The algorithm revealed optimal models for both passing and failing students, offering insights into successful self-regulated learning processes. Another study used PM from a university LMS to analyse learners' behavior ([215]), while the study in ([217]) analyses 20 cases to study patterns linked to learning performance, enhancing teaching guidance with process-oriented feedback.

Predicting the learning outcome. In ([271]), web-log data from a Moodle-based LMS were used to investigate 84 students' academic achievements. A multi-regression analysis showed a significant correlation with the final learning grade. Finally, the authors suggest that "educators should pay more attention to improve the process of learners' achievement". In another study, student's behavior was monitored for evaluation purposes during a semester by constructing an event log of their activities in a specific LMS ([37]). The authors stated that teachers must design teaching strategies that provide early or real-time detection of students who do not follow the learning path. Predictive models have been implemented using students' behavior based on an edX-based LMS ([50]), to identify underperforming students early ([48]), as well as students' abilities before and after problem-solving tasks ([125]) by using Gradient Boosting Decision Trees on historical event logs. Other predictive studies involve the automated analysis of traces left by students in MOOCs ([204]), also by differentiating various subgroups of learners ([129]), demonstrating how to predict the performance of students at an early stage ([245]), as well as to predict student's outcome in a course by exploiting information on LMSs ([244]). A previous research identified three main types of outcome prediction: the exact final grade (e.g., the range can correspond to a scale from 0 to 10), mapping into a limited number of categories, usually 4 or 5, or a discretization into two categories, i.e. negative/positive ([101]). The work focused on the last categorization.

Learning styles. Learning styles have been the subject of many studies that recognized the existence of multiple factors, often attributable to the learner's personal characteristics or the used technologies. A recent literature review summarized the existing theories on learning styles ([241]). As they generally suffer from validity and reliability issues ([42]), no theory outweighs the others. Nevertheless, one of the most popular theories that have been applied in e-learning systems is the Felder-Silverman one ([69]). Their theory includes the categorization between sequential versus global learning styles: sequential learning style concerns the acquisition of understanding in a linear fashion, with a logical progression of ordered steps; on the contrary, a global learning style involves absorbing material more disorderedly, including non-linear connections and jumps between the various parts ([70]).

In [147], process discovery has been applied to investigate learning styles in a MOOC course, finding a positive correlation between sequential learning and students' performance.

Process analysis revealed that successful students followed the learning path while less successful students did not ([37]). A relevant issue concerns the consideration of the learners' goals and their regulatory mechanisms. A conceptual model and a practical case example have been proposed with the adoption of a feedback-driven dashboard, i.e. a dashboard designed on the basis of empirical evidence to enhance learning regulation by providing both cognitive and behavioral feedback ([217]). In their work, process discovery has been adopted to investigate the interactions between user participants. Previously, process discovery has been used to analyze the detailed logs of novice users' interactions within a specific tool in [216]. This kind of research connects process-mining enabled analysis of learning processes and behaviors with learning theories, aligning data collection and analysis with underlying learning processes from the learning sciences. By examining 20 cases with over 10,000 logged events, process discovery helped identify patterns and sequences in the learning process. The work contributes to this cross-domain direction by studying learning behavior in a real-world situation.

The study in [121] explores how customizing web-based learning to match individual styles -distinguishing between "Explorers" (who prefer self-navigation) and "Observers" (who follow structured paths)- can enhance learning effectiveness. With 58 participants, findings suggest that learning outcomes improve when the system's navigation style aligns with the user's learning preference, emphasizing the potential of adaptive learning platforms. "Explorers" performed better when jumping between content, while "Observers" excelled with linear navigation. This indicates that customized learning platforms, responsive to individual preferences, can enhance learning outcomes. investigated these modes of behavior in a short tutorial. Finally, a relevant feature of a learning style concerns its duration. The assumption is that the learning style remains fixed for the duration of the tutorial, according to [241].

Learning design. Studies on learning styles demonstrated how hypermedia technologies benefit learners with different needs ([126]). As in this work, the application of automated process analysis in education has also been shown to have an impact on the field of Learning Design, which can be defined as "a methodology for enabling teachers/designers to make more informed decisions in how they go about designing learning activities and interventions, which is pedagogically informed and makes effective use of appropriate resources and technologies" ([132]).

According to a recent review, the most frequent kind of learning concerns 'assimilative activity', such as reading module materials, which corresponds to the one addressed in the work ([199]).

The study leverages PM and machine learning techniques with a similar purpose to previous studies, namely, to determine learning processes describing behaviors of successful and less successful students and to predict students' performance before the end of the learning trajectory. Compared with the state-of-the-art, the distinguishing features and improvements of the work include the following main points:

- the focus of the analysis concerns a learning path of short duration (two hours at most) and not months or years as in most studies;
- the exploitation of web technologies to track behavior within tutorial paragraphs on web pages, and do not use data from pre-existing systems such as MOOCs used by most studies in this area;
- the application of PM analysis on short tutorials in such a tracking system.

According to available knowledge, no previous work has addressed this type of analysis on relatively short learning paths, exploiting web-based technologies with PM techniques.

6.7 Conclusion

The research proposed a methodology for studying the learning of short tutorials using the combination of a web tracking system and the application of PM techniques at descriptive level. A practical case study demonstrated how this methodology could investigate the learning path and activity flows of students who did well and poorly (RQ1). The analysis suggests differences in students' learning and satisfaction adopting a specific order among topics.

Finally, the proposed methodology can be applied to identify possible bottlenecks and other hints in relatively short learning paths. The fact that the student who performs poorly goes fast from the start as well as behaves with a more linear path instead of jumping back to previous paragraphs, may suggest that the system can make appropriate slowdowns or alerts when it detects potentially dangerous behavior in learning. The prediction results (RQ2) encourage the adoption of a prediction system in the tutorial's initial part (ideally at the end of the third lesson) to investigate students who are at risk of insufficient learning after the first part of the course.

Future work. This research aims to increase the number of tutorial administrations to obtain more statistically significant results. In addition, plan to extend the survey with more variables, e.g., demographic data and previous knowledge. From a learning design perspective, would like to gather more suggestions on the usability front and address bottleneck

analysis of the present tutorial to identify valuable suggestions for implementing an improved version. The new version of the tutorial can then be resubmitted to another similar set of students to investigate the improvements, as part of prescriptive process monitoring ([114]). For instance, the PM analysis can identify paragraphs of the actual version of the tutorial where most students spend too much and be grounds for restructuring for a new, improved version. This research aims to extend the work by implementing appropriate feedback to students in order to investigate aspects of the cognitive thinking process or regulation.

Moreover, as domain experts have suggested, a survey of student's initial knowledge of the subject (programming in Python) before the tutorial may be included to assess its benefits at the end of the learning path. As far as the prediction phase is concerned, in future research, explainability issues ([139]) will be explored as well as deep-learning models such as long short-term memory, generative adversarial networks, and transformers, which require more significant amounts of data to be trained effectively ([103]).

Chapter 7

Conclusion and Future Work

This thesis explored data science techniques such as ML, PM, PPM, and LLM applied to legal data, obtaining interesting results. Regarding sustainability research topics, the analysis began by investigating public procurement for "green" services. The study has been extended to all procurements, both on national and European datasets. Finally, it shows how these techniques can also be applied in domains other than the legal, such as the educational sector. In summary, the contributions of the thesis covered the following topics:

- ML for Law. Chapter 3 explored the application of ML algorithms to predict the likelihood of legal complaints in public tenders. The research used various ML models, including Logistic Regression, Decision Trees, Random Forest, XGBoost, and Naive Bayes, to build predictive models that aided in the early identification of potential legal disputes. Implementing SHAP (SHapley Additive exPlanations) provided transparency and interpretability to the models, which is essential for building trust in automated decision systems within the legal domain. The results showed that ML techniques are highly effective for predicting legal risks, offering significant improvements in tender management and decision-making processes.

- PM for Law. Chapter 4 examined the use of PM techniques to assess transparency and efficiency in European public tenders. The study transformed public procurement data into event logs, allowing for process discovery, variant analysis, and the prediction of procedural inefficiencies. The results provided insights into how PM can optimize tendering processes by identifying bottlenecks and forecasting remaining process time. These findings improve operational efficiency and enhance transparency in the public procurement lifecycle, demonstrating the value of PM in legal and public administration contexts.

- LLM for Law. Chapter 5 describes the creation of a decision support system for legal practitioners to manage increasing volumes of data. Using a real public procurement dataset, the approach integrated a recommendation system with generative model technologies in a proof-of-concept study. The results demonstrated the feasibility of applying this system in the legal field. Although the implementation requires substantial computational resources, domain experts highly value performance benefits.

  The conclusion emphasizes a specific request from domain experts for an application that is easy to manage using standard equipment. The proposed solution is to develop a web browser-based application for managing results, allowing for ongoing testing and feedback to drive further improvements. Future plans include implementing a service using Flask and ReactJS to facilitate the use of the tool in a web-based environment.

  Further work is planned to introduce additional evaluation metrics beyond NDCG and conduct a more detailed qualitative evaluation by domain experts, incorporating more queries for submission to the system. On the technological front, there are plans to compare the proposed RAG framework with state-of-the-art alternatives, such as Fast-RAG. Regarding LLMs, there is an intention to explore the use of *open* LLMs, such as LLaMA, as alternatives to proprietary tools like Cohere and OpenAI. Given that LLMs are often black-box models, it is not possible to determine the extent to which they have been trained on legal data, which presents a limitation for their application in specialized fields like law.

  Finally, given that the general framework presented here in the legal field and the corresponding proof-of-concept results have demonstrated the tool's usefulness, the application has the potential to be tested in other domains, such as healthcare.

- PPM for EDU. Chapter 6 extended the application of PM and PPM techniques to educational EDU environments, demonstrating that the benefits of PM can go beyond traditional legal and public procurement settings. The chapter presented a case study using PPM techniques to analyse short-term e-learning tutorials. Event logs generated from student interactions in these tutorials were used to discover patterns, predict learning outcomes, and identify inefficiencies in the learning process. By leveraging PM techniques, the study provided actionable insights into how learning paths could be optimized to enhance educational outcomes.

  The results demonstrated that PPM can effectively model and predict student performance in e-learning environments, offering educators a powerful tool for improving tutorial design and tracking student progress. This extension of PM techniques to the EDU domain illustrates the versatility and adaptability of the methods developed

in this thesis, suggesting further opportunities for applying PPM in various sectors beyond legal and public procurement contexts.

Final Remarks. Overall, the thesis presents a comprehensive investigation into using advanced data science techniques -ML, PM, PPM, and LLMs- in legal and educational contexts. The research has shown that these methods effectively address real-world challenges in public procurement, legal risk prediction, and optimisation of the educational process. By integrating these cutting-edge technologies into the workflows of legal practitioners and educators, this thesis contributes to a more efficient, transparent, and data-driven approach to manage complex processes in these domains.

Future work will continue to refine and expand the tools developed further to enhance their applicability and impact across multiple fields. For example, this research aims to explore the potential of NN-based methods in future work to complement the findings presented here. NN, such as DL models, are known for their ability to handle complex modeling tasks, but their application often involves significant computational requirements and large datasets to achieve reliable results [86]. Given the scope of this thesis, the focus was placed on interpretable and resource-efficient models, such as LR and RF, which offer transparency and perform well with smaller datasets [206]. Moreover, in structured data contexts, ensemble methods can sometimes provide comparable performance to neural networks without the added complexity [220]. This approach allows for a balance between interpretability, feasibility, and performance, while leaving room for future exploration of more complex models.

Furthermore, while LLMs have been effectively utilized in this research, their black-box nature and the potential biases in their responses are acknowledged as critical challenges, particularly in applications with social or legal impact [206, 25]. LLMs, despite their impressive capabilities, can propagate or even amplify biases present in training data, raising concerns about fairness and reliability [13]. Addressing these biases requires systematic evaluations and mitigation strategies, which are planned as part of future work. For this thesis, the focus remained on leveraging the strengths of LLMs for knowledge extraction while adhering to the scope and time constraints of the research. Future directions will explore methods to enhance transparency and fairness in LLM-based systems, including explainability techniques and bias mitigation frameworks [124, 78].

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

[2] ACFE (2020). 2020 report to the nations—the acfe's 11th study on the costs and effects of occupational fraud. Accessed: 2024-10-03.

[3] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[4] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[5] Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., and Vrahatis, M. N. (2019). No free lunch theorem: A review. *Approximation and Optimization*, pages 57–82.

[6] Adnan, K. and Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6(1):1–38.

[7] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[8] Amantea, I. A., Robaldo, L., Sulis, E., Boella, G., and Governatori, G. (2021). Semi-automated checking for regulatory compliance in e-health. In *25th International Enterprise Distributed Object Computing Workshop, EDOC Workshop 2021, Gold Coast, Australia, October 25-29, 2021*, pages 318–325. IEEE.

[9] Arrieta, A. B. et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

[10] Azzopardi, L., Moshfeghi, Y., Halvey, M., Alkhawaldeh, R. S., Balog, K., Di Buccio, E., Ceccarelli, D., Fernández-Luna, J. M., Hull, C., Mannix, J., et al. (2017). Lucene4ir: Developing information retrieval evaluation resources using lucene. In *ACM SIGIR Forum*, volume 50, pages 58–75. ACM New York, NY, USA.

[11] Banker, K., Bakkum, P., Verch, S., Garrett, D., and Hawkins, T. (2011). *MongoDB in Action*. Manning Publications.

[12] Becker, S. O., Egger, P. H., and von Ehrlich, M. (2010). Going nuts: The effect of eu structural funds on regional performance. *Journal of Public Economics*, 94(9):578–590.

[13] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

[14] Bennetot, A., Donadello, I., El Qadi, A., Dragoni, M., Frossard, T., Wagner, B., Saranti, A., Tulli, S., Trocan, M., Chatila, R., et al. (2021). A practical tutorial on explainable ai techniques. *arXiv preprint arXiv:2111.14260*.

[15] Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015a). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008.

[16] Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015b). Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008.

[17] Bernd W. Wirtz, J. C. W. and Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated ai governance framework for public administration. *International Journal of Public Administration*, 43(9):818–829.

[18] Berti, A., Kourani, H., Häfke, H., Li, C., and Schuster, D. (2024). Evaluating large language models in process mining: Capabilities, benchmarks, and evaluation strategies. In van der Aa, H., Bork, D., Schmidt, R., and Sturm, A., editors, *Enterprise, Business-Process and Information Systems Modeling - 25th International Conference, BPMDS 2024, and 29th International Conference, EMMSAD 2024, Limassol, Cyprus, June 3-4, 2024, Proceedings*, volume 511 of *Lecture Notes in Business Information Processing*, pages 13–21. Springer.

[19] Bex, F. J. (2024). Ai, law and beyond: A transdisciplinary ecosystem for the future of ai & law. *Artificial Intelligence and Law*, pages 1–18.

[20] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

[21] Bogarín, A., Cerezo, R., and Romero, C. (2018). A survey on educational process mining. *WIREs Data Mining Knowl. Discov.*, 8(1).

[22] Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., and Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*.

[23] Bolt, A., de Leoni, M., and van der Aalst, W. M. P. (2018). Process variant comparison: Using event logs to detect differences in behavior and business rules. *Information Systems*, 74(Part):53–66.

[24] Bolt, A. and van Zelst, S. (2020). Insights from variant analysis in process mining. *Business Process Management Journal*, 26(4):857–872.

[25] Bommasani, R. et al. (2021). Opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[26] Bosio, E., Djankov, S., Glaeser, E. L., and Shleifer, A. (2021). Transparency in public procurement: A review of evidence from the field. *Journal of Public Economics*, 200:104442.

[27] Breiman, L. (2001a). Machine learning, volume 45, number 1 - springerlink. *Machine Learning*, 45:5–32.

[28] Breiman, L. (2001b). Random forests. *Machine Learning*, 45(1):5–32.

[29] Brown, T. et al. (2020a). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[30] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020b). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[31] Buckland, M. K. and Gey, F. C. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19.

[32] Bujlow, T., Carela-Español, V., Solé-Pareta, J., and Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, 105(8):1476–1510.

[33] Carlson, J. L. (2013). *Redis in Action*. Manning Publications Co.

[34] Carmona, J., van Dongen, B. F., Solti, A., and Weidlich, M. (2018). *Conformance Checking - Relating Processes and Models*. Springer.

[35] Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., and Costa, J. (2020). *Network Analysis for Fraud Detection in Portuguese Public Procurement*, pages 390–401. Springer, Cham.

[36] Cartei, G. F. and Iaria, D., editors (2023). *Commentary on the New Public Procurement Code*. Naples. pp. 125 et seq.

[37] Cenka, B. A. N., Santoso, H. B., and Junus, K. (2022). Analysing student behaviour in a learning management system using a process mining approach. *Knowledge Management & E-Learning*, 14(1):62–80.

[38] Ceravolo, P., Comuzzi, M., De Weerdt, J., Di Francescomarino, C., and Maggi, F. M. (2024). Predictive process monitoring: concepts, challenges, and future research directions. *Process Science*, 1(1):2.

[39] Cerezo, R., Bogarín, A., Esteban, M., and Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, 32(1):74–88.

[40] Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint*, arXiv:1704.00051. Available at arXiv:1704.00051 [cs.CL].

[41] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). XGBoost: Extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

[42] Coffield, F., Ecclestone, K., Hall, E., and Moseley, D. (2004). Learning styles and pedagogy in post-16 learning: A systematic and critical review. Learning and Skills Research Council, London.

[43] Conforti, R., La Rosa, M., ter Hofstede, A. H., and van der Aalst, W. M. (2015). Recommendation-based trace clustering. *Data & Knowledge Engineering*, 98:47–64.

[44] Cord, M. and Cunningham, P. (2008). *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Springer Science & Business Media, Berlin, Germany.

[45] Dandl, S. and Molnar, C. (2020). Counterfactual explanations.

[46] Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications.

[47] De Carolis, L., Palumbo, F., and Bartolomeo, M. (2022). Corruption in public procurement: How to mitigate risks? *Journal of Public Administration Research and Theory*, 32(1):58–76.

[48] De Smedt, J., Deeva, G., and De Weerdt, J. (2019). Mining behavioral sequence constraints for classification. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1130–1142.

[49] Decarolis, F. and Giorgiantonio, C. (2020). Corruption red flags in public procurement: New evidence from italian calls for tenders. *SSRN Electronic Journal*.

[50] Deeva, G., Smedt, J. D., Saint-Pierre, C., Weber, R., and Weerdt, J. D. (2022). Predicting student performance using sequence classification with time-based windows. *Expert Systems with Applications*, 209:118182.

[51] DeVito, N. J., Richards, G., and Inglesby, P. (2020). How we learnt to stop worrying and love web scraping. *Nature*, 585:621–623.

[52] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, pages 4171–4186.

[53] Dhurandhar, A., Graves, B., Ravi, R., Maniachari, G., and Ettl, M. (2015). Big data system for analyzing risky procurement entities. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1741–1750, New York, NY, USA. Association for Computing Machinery.

[54] Di Francescomarino, C. and Ghidini, C. (2022). Predictive process monitoring. In van der Aalst, W. M. P. and Carmona, J., editors, *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*, pages 320–346. Springer.

[55] Dietterich, T. G. (1995). Overfitting and underfitting in machine learning. *ACM Computing Surveys*, 27(3):326–327.

[56] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer.

[57] Dignum, V. (2017). Responsible artificial intelligence: designing ai for human values. *Journal of Ethics and Information Technology*.

[58] Dinan, E. et al. (2021). Anticipating safety issues in ai deployment. *arXiv preprint arXiv:2109.09794*.

[59] Dixit, B. (2016). *Elasticsearch Essentials*. Packt Publishing Ltd, Birmingham, UK.

[60] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[61] Du, M., Liu, N., and Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

[62] Eberendu, A. C. (2016). Unstructured data: an overview of the data of big data. *International Journal of Computer Trends and Technology (IJCTT)*, 38(1):46–50.

[63] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

[64] Ermakova, T., Fabian, B., Bender, B., and Klimek, K. (2018). Web tracking - A literature review on the state of research. In Bui, T., editor, *51st Hawaii International Conference on System Sciences, HICSS 2018, Hilton Waikoloa Village, Hawaii, USA, January 3-6, 2018*, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL).

[65] European Organization For Nuclear Research and OpenAIRE (2013). Zenodo.

[66] Fahrenkrog-Petersen, S. A. (2019). Providing privacy guarantees in process mining. In Rosa, M. L., Plebani, P., and Reichert, M., editors, *Proceedings of the Doctoral Consortium Papers Presented at the 31st International Conference on Advanced Information Systems Engineering (CAiSE 2019), Rome, Italy, June 3-7, 2019*, volume 2370 of *CEUR Workshop Proceedings*, pages 23–30. CEUR-WS.org.

[67] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

[68] Fazekas, M. and Kocsis, G. (2021). Extra-legal influences in public procurement: Evidence from administrative data. *Government Information Quarterly*, 38(2):101568.

[69] Felder, R. and Silverman, L. (1988). Learning and teaching styles in engineering education. *Journal of Engineering Education*, 78:674–681.

[70] Felder, R. M. and Brent, R. (2016). *Teaching and learning STEM: A practical guide*. John Wiley & Sons.

[71] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 873–881.

[72] Feng, G. and Fan, M. (2024). Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization. *Expert Systems with Applications*, 237:121555.

[73] Ferguson, R. (2019). *Beginning JavaScript: The Ultimate Guide to Modern JavaScript Development*. Apress, Berkeley, CA.

[74] Friedl, J. E. (2006). *Mastering Regular Expressions*. O'Reilly Media, Inc., 3rd edition.

[75] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[76] Gallego, J., Rivero, G., and Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 37(1):360–377.

[77] Gallego, R. and Ibarlucea, G. (2021). Preventing fraud, corruption, and collusion in public procurement: An application of machine learning models. *Journal of Public Procurement*, 21(2):149–169.

[78] Garcia, A. et al. (2022). Measuring bias in llms: A framework for analysing social and legal impacts. *Journal of AI Research*, 75:123–146.

[79] Garcia, J. M., Romero, H., and Gonzalez, P. (2019). Public procurement award price estimator based on machine learning techniques. *Computers in Industry*, 108:19–27.

[80] García Rodríguez, M. J., Montequín, V., Ortega-Fernández, F., and Balsera, J. (2019). Public procurement announcements in spain: Regulations, data analysis, and award price estimator using machine learning. *Complexity*, 2019:1–20.

[81] García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., and Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047.

[82] Gaudreault, J.-G., Branco, P., and Gama, J. (2021). An analysis of performance metrics for imbalanced classification. In Soares, C. and Torgo, L., editors, *Discovery Science*, volume 12986, pages 67–77. Springer, Cham.

[83] Geeganage, D. T. K., Wynn, M. T., and ter Hofstede, A. H. (2022). Text2el: Exploiting unstructured text for event log enrichment. In *2022 16th Int. Conf. on SITIS*, pages 1–8.

[84] Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 3rd edition.

[85] Ghazal, M. A., Ibrahim, O., and Salama, M. A. (2017). Educational process mining: A systematic literature review. In *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, pages 198–203.

[86] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[87] Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. O'Reilly Media, Inc., Sebastopol, CA, USA.

[88] Gourley, D. and Totty, B. (2002). *HTTP: The Definitive Guide*. O'Reilly Media, Inc.

[89] Grainger, T. and Potter, T. (2014). *Solr in Action*. Manning Publications Co., Shelter Island, NY, USA.

[90] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

[91] Greve, C. and Lægreid, P. (2018). Public procurement and the risks of legal challenges: A policy perspective. *International Journal of Public Sector Management*, 31(5):589–606.

[92] Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58.

[93] Gupta, S., Kaiser, G., Neistadt, D., and Grimm, P. (2003). Dom-based content extraction of html documents. In *Proceedings of the 12th international conference on World Wide Web*, pages 207–214. ACM.

[94] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938, Virtual Conference. PMLR.

[95] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2nd edition.

[96] Hashmi, M., Governatori, G., Lam, H., and Wynn, M. T. (2018). Are we done with business process compliance: state of the art and challenges ahead. *Knowl. Inf. Syst.*, 57(1):79–133.

[97] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition.

[98] Hewitt, E. (2010). *Cassandra: The Definitive Guide*. O'Reilly Media, Inc.

[99] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[100] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

[101] Hu, X., Cheong, C. W. L., Ding, W., and Woo, M. (2017). A systematic review of studies on predicting student learning outcomes using learning analytics. In Hatala, M., Wise, A. F., Winne, P., Lynch, G., Ochoa, X., Molenaar, I., Dawson, S., Shehata, S., and Tan, J. P., editors, *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*, pages 528–529. ACM.

[102] Jiang, J., Aldewereld, H., Dignum, V., Wang, S., and Baida, Z. (2015). Regulatory compliance of business processes. *AI Soc.*, 30(3):393–402.

[103] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[104] Jorge Munoz-Gama et al. (2022). Process mining for healthcare: Characteristics and challenges. *J. Biomed. Informatics*, 127:103994.

[105] Juhaňák, L., Zounek, J., and Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior*, 92:496–506.

[106] Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Pearson Education International, Upper Saddle River, NJ, USA, 2nd edition.

[107] Kabudi, T., Pappas, I. O., and Olsen, D. H. (2021). Ai-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2:100017.

[108] Kamalloo, E., Zhang, X., Ogundepo, O., Thakur, N., Alfonso-Hermelo, D., Rezagholizadeh, M., and Lin, J. (2023). Evaluating embedding apis for information retrieval. *arXiv preprint arXiv:2305.06300*. Available at arXiv:2305.06300 [cs.IR].

[109] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

[110] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1137–1145.

[111] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1137–1145. Montreal, Canada.

[112] Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, 39(4):261–283.

[113] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

[114] Kubrak, K., Milani, F., Nolte, A., and Dumas, M. (2022). Prescriptive process monitoring: *Quo vadis*? *PeerJ Comput. Sci.*, 8:e1097.

[115] Kumar, S. et al. (2023). Application of explainable artificial intelligence to analyze basic features of a tender. In *2023 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 1–6. IEEE.

[116] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[117] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[118] Leemans, S. J. J., Fahland, D., and van der Aalst, W. M. P. (2014). Discovering block-structured process models from event logs. *Lecture Notes in Computer Science*, 8518:311–329.

[119] Leemans, S. J. J., Fahland, D., and van der Aalst, W. M. P. (2018). Process discovery algorithms in practice. *Information Systems*, 74:30–52.

[120] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9459–9474.

[121] Liegle, J. O. and Janicki, T. N. (2006). The effect of learning styles on the navigation needs of web-based learners. *Computers in Human Behavior*, 22(5):885–898.

[122] Lima, M., Silva, R., Lopes de Souza Mendes, F., R. de Carvalho, L., Araujo, A., and de Barros Vidal, F. (2020). Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1580–1588, Online. Association for Computational Linguistics.

[123] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

[124] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

[125] Liu, F., Zhao, L., Zhao, J., Dai, Q., Fan, C., and Shen, J. (2022). Educational process mining for discovering students' problem-solving ability in computer programming education. *IEEE Transactions on Learning Technologies*, 15(6):709–719.

[126] Liu, M. and Reed, W. (1994). The relationship between the learning strategies and learning styles in a hypermedia environment. *Computers in Human Behavior*, 10(4):419–434.

[127] Liu, Q. and Wu, Y. (2012). Supervised learning. In Seel, N. M., editor, *Encyclopedia of the Sciences of Learning*, pages 3243–3245. Springer, Boston, MA.

[128] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[129] Luna, J. M., Fardoun, H. M., Padillo, F., Romero, C., and Ventura, S. (2022). Subgroup discovery in moocs: a big data application for describing different types of learners. *Interactive Learning Environments*, 30(1):127–145.

[130] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774.

[131] Lyra, M. S., Pinheiro, F. L., and Bacao, F. (2022). Public procurement fraud detection: A review using network analysis. In Benito, R. M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L. M., and Sales-Pardo, M., editors, *Complex Networks & Their Applications X*, pages 116–129, Cham. Springer International Publishing.

[132] Macfadyen, L. P., Lockyer, L., and Rienties, B. (2020). Learning design and learning analytics: Snapshot 2020. *Journal of Learning Analytics*, 7(3):6–12.

[133] Maggi, F. M., Francescomarino, C. D., Dumas, M., and Ghidini, C. (2014). Predictive monitoring of business processes. In Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H., and Horkoff, J., editors, *Advanced Information Systems Engineering - 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings*, volume 8484 of *Lecture Notes in Computer Science*, pages 457–472. Springer.

[134] Mannhardt, F. and de Leoni, M. (2019). Event log quality: Methods and techniques. *ACM Transactions on Management Information Systems*, 10(2):1–23.

[135] Mannhardt, F., Petersen, S. A., and Oliveira, M. F. (2018). Privacy challenges for process mining in human-centered industrial environments. In *2018 14th International Conference on Intelligent Environments (IE)*, pages 64–71.

[136] Marczynski, A. and Marín, J. M. (2018). *Compendium of Good Practices on Anti-Corruption for OGP Action Plans*. Transparency International, Berlin, Germany.

[137] McCandless, M., Hatcher, E., and Gospodnetic, O. (2010). *Lucene in Action, Second Edition*. Manning Publications.

[138] Mehta, A. (2017). Comparison between json and bson for data interchange formats. *Journal of Data Science and Engineering*, 2(1):12–21.

[139] Meo, R., Nai, R., and Sulis, E. (2022). Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In Chiusano, S., Cerquitelli, T., and Wrembel, R., editors, *Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, volume 13389 of *Lecture Notes in Computer Science*, pages 25–34. Springer.

[140] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.

[141] Modrusan, N., Rabuzin, K., and Mrsic, L. (2021). Review of public procurement fraud detection techniques powered by emerging technologies. *International Journal of Advanced Computer Science and Applications*, 12.

[142] Modrušan, N., Rabuzin, K., and Mrsic, L. (2020). Improving public sector efficiency using advanced text mining in the procurement process. In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, pages 200–206.

[143] Mohit, B. (2014). Named entity recognition. In Zitouni, I., editor, *Natural language processing of semitic languages*, pages 221–245. Springer.

[144] Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45.

[145] Molnar, C. (2019). *Interpretable Machine Learning*. Lulu. com.

[146] Moreno, M., Exposito, E., and Gueye, M. (2021). Process mining model to visualize and analyze the learning process. In *REES AAEE 2021 conference: Engineering Education Research Capability Development: Engineering Education Research Capability Development*, pages 698–706. Engineers Australia Perth, WA.

[147] Mukala, P., Buijs, J. C. A. M., Leemans, M., and van der Aalst, W. M. P. (2015). Learning analytics on coursera event data: A process mining approach. In Ceravolo, P. and Rinderle-Ma, S., editors, *Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2015), Vienna, Austria, December 9-11, 2015*, volume 1527 of *CEUR Workshop Proceedings*, pages 18–32. CEUR-WS.org.

[148] Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

[149] Munoz-Gama, J. (2016). *Conformance Checking and Diagnosis in Process Mining - Comparing Observed and Modeled Processes*, volume 270 of *Lecture Notes in Business Information Processing*. Springer.

[150] Murphy, K. P. (2012a). *Machine Learning: A Probabilistic Perspective*. MIT Press.

[151] Murphy, K. P. (2012b). *Machine Learning: A Probabilistic Perspective*. MIT Press.

[152] Musciano, C. and Kennedy, B. (2002). *HTML & XHTML: The Definitive Guide*. O'Reilly Media, Inc., 5th edition.

[153] Nai, R., Fatima, I., Morina, G., Sulis, E., Genga, L., Meo, R., and Pasteris, P. (2023a). AI applied to the analysis of the contracts of the italian public administrations. In Falchi, F., Giannotti, F., Monreale, A., Boldrini, C., Rinzivillo, S., and Colantonio, S., editors, *Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), Pisa, Italy, May 29-30, 2023*, volume 3486 of *CEUR Workshop Proceedings*, pages 255–260. CEUR-WS.org.

[154] Nai, R., Meo, R., Morina, G., and Pasteris, P. (2023b). Public tenders, complaints, machine learning and recommender systems: a case study in public administration. *Comput. Law Secur. Rev.*, 51:105887.

[155] Nai, R., Sulis, E., Fatima, I., and Meo, R. (2024a). Large language models and recommendation systems: A proof-of-concept study on public procurements. In Rapp, A., Caro, L. D., Meziane, F., and Sugumaran, V., editors, *Natural Language Processing and Information Systems - 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings, Part II*, volume 14763 of *Lecture Notes in Computer Science*, pages 280–290. Springer.

[156] Nai, R., Sulis, E., and Genga, L. (2023c). Automated analysis with event log enrichment of the european public procurement processes. In Sales, T. P., Araújo, J., Borbinha, J., and Guizzardi, G., editors, *Advances in Conceptual Modeling - ER 2023 Workshops, CMLS, CMOMM4FAIR, EmpER, JUSMOD, OntoCom, QUAMES, and SmartFood, Lisbon, Portugal, November 6-9, 2023, Proceedings*, volume 14319 of *Lecture Notes in Computer Science*, pages 178–188. Springer.

[157] Nai, R., Sulis, E., and Genga, L. (2024b). Enhancing e-learning effectiveness: a process mining approach for short-term tutorials. *Journal of Intelligent Information Systems*.

[158] Nai, R., Sulis, E., Marengo, E., Vinai, M., and Capecchi, S. (2023d). Process mining on students' web learning traces: A case study with an ethnographic analysis. In Viberg, O., Jivet, I., Muñoz-Merino, P. J., Perifanou, M. A., and Papathoma, T., editors, *Responsive and Sustainable Educational Futures - 18th European Conference on Technology Enhanced Learning, EC-TEL 2023, Aveiro, Portugal, September 4-8, 2023, Proceedings*, volume 14200 of *Lecture Notes in Computer Science*, pages 599–604. Springer.

[159] Nai, R., Sulis, E., and Meo, R. (2022a). Public procurement fraud detection and artificial intelligence techniques: a literature review. In Danai Symeonidou et al., editor, *23rd EKAW Int. Conf. proc.*, volume 3256 of *CEUR*. CEUR-WS.org.

[160] Nai, R., Sulis, E., and Meo, R. (2024c). Ith: an open database on italian tenders 2016–2023. *Scientific Data*, 11(1):1452.

[161] Nai, R., Sulis, E., Pasteris, P., Giunta, M., and Meo, R. (2022b). Exploitation and merge of information sources for public procurement improvement. In Irena Koprinska et al., editor, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 1752 of *Commun. Comput. Inf. Sci.*, pages 89–102. Springer.

[162] Niessen, M., Paciello, J., and Fernandez, J. (2020). Anomaly detection in public procurements using the open contracting data standard. In *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 127–134.

[163] Nti, I. K., Nyarko-Boateng, O., and Aning, J. (2021). Performance of machine learning algorithms with different k values in k-fold cross-validation. *Journal of Information Technology and Computer Science*, 6:61–71.

[164] OECD (2021). Cross-border data flows: Taking stock of key policies and initiatives. Accessed: 2024-09-17.

[165] OECD (2022). G20 compendium on data access and sharing across the public sector and with the private sector for public interest. Accessed: 2024-09-17.

[166] OECD (2023). *Government at a Glance 2023*. OECD Publishing, Paris, France.

[167] Ofoeda, J., Boateng, R., and Effah, J. (2019). Application programming interface (api) research: A review of the past to inform the future. *International Journal of Enterprise Information Systems (IJEIS)*, 15(3):76–95.

[168] Ovsyannikova, M. and Tkachenko, V. (2020). Identification of public procurement contracts with high risk of non-performance using machine learning. *International Journal of Public Sector Management*, 33(2/3):222–238.

[169] Palma, F., Estrada, A., Martin-Lopez, A., Segura, S., and Ruiz-Cortés, A. (2021). Online testing of restful apis: Promises and challenges. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 408–420.

[170] Pamučar, D., Božanić, D., and Ćosić, I. (2022). Application of a neuro-fuzzy neural network to predict public procurement success in construction companies. *Journal of Civil Engineering and Management*, 28(3):201–215.

[171] Passas, N. (2007). Corruption in the procurement process/outsourcing government functions: Issues, case studies, implications. *Report to the Institute for Fraud Prevention*.

[172] Patro, S. G. K. and Sahu, K. K. (2015). Normalization: A preprocessing stage.

[173] Pazienza, M. T., editor (1997). *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg.

[174] Perianes-Rodriguez, A., Waltman, L., and van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4):1178–1195.

[175] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

[176] Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., and Vrgoč, D. (2016). Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, pages 263–273. International World Wide Web Conferences Steering Committee.

[177] Plumed, F., Casamayor, J., Ferri, C., Gómez, J., and Vendrell Vidal, E. (2019). *SALER: A Data Science Solution to Detect and Prevent Corruption in Public Administration*, pages 103–117. Springer, Cham.

[178] Popa, A. and Mungiu-Pippidi, A. (2019). Uncovering close connections in public procurement: A machine learning approach. *Governance*, 32(1):91–109.

[179] Popa, M. (2019). Uncovering the structure of public procurement transactions. *Business and Politics*, 21(3):351–384.

[180] Potdar, K., Pardawala, T., and Pai, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175(4):7–9.

[181] Potin, L., Labatut, V., Morand, P.-H., and Largeron, C. (2023). Foppa: an open database of french public procurement award notices from 2010–2020. *Scientific Data*, 10(1):303.

[182] Pufahl, L. and Rehse, J. (2021). Conformance checking with regulations - A research agenda. In Koschmider, A. and Michael, J., editors, *11th Int. Workshop on EMISA*, volume 2867, pages 24–29. CEUR-WS.org.

[183] Que, X., Checconi, F., Petrini, F., and Gunnels, J. A. (2015). Scalable community detection with the Louvain algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium*, pages 28–37. IEEE.

[184] Rabuzin, K., Modrušan, N., Križanić, S., and Kelemen, R. (2022). Process mining in public procurement in croatia. In Lalic, B., Gracanin, D., Tasic, N., and Simeunović, N., editors, *Proceedings on 18th Int. Conf. on IS'20*, pages 473–480, Cham. Springer.

[185] Rabuzin., K. and Modrušan., N. (2019). Prediction of public procurement corruption indices using machine learning methods. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KMIS,*, pages 333–340. INSTICC, SciTePress.

[186] Rabuzin, T., Babić, D., and Horvat, A. (2019). Prediction of public procurement outcomes using advanced text mining techniques. *Journal of Public Procurement*, 19(3):233–251.

[187] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

[188] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. Technical report, OpenAI. OpenAI Technical Report.

[189] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[190] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

[191] Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. In *Encyclopedia of Database Systems*. Springer, Boston, MA.

[192] Republic, I. (2010). Legislative decree no. 104 of 2 july 2010, code of administrative procedure. Adoption of the Code of Administrative Procedure, reorganizing the rules governing administrative justice in Italy.

[193] Republic, I. (2016). Legislative decree no. 50 of 18 april 2016, public procurement code. Implementation of Directives 2014/23/EU, 2014/24/EU, and 2014/25/EU on public procurement, concessions, and procurement by entities operating in water, energy, transport, and postal services sectors.

[194] Republic, I. (2020a). Decree-law no. 76 of 16 july 2020, converted with amendments by law no. 120 of 11 september 2020. Urgent measures for simplification and digital innovation. Applicable until 31/12/2023.

[195] Republic, I. (2020b). Law decree no. 76 of 16 july 2020, simplification decree. Urgent measures for simplification and digital innovation. Applicable until 31 December 2023.

[196] Republic, I. (2020c). Law no. 120 of 11 september 2020, converting with amendments decree-law no. 76 of 16 july 2020. Conversion into law, with amendments, of Decree-Law No. 76 of 16 July 2020, introducing urgent measures for simplification and digital innovation. Applicable until 31/12/2023.

[197] Republic, I. (2023). Legislative decree no. 36 of 31 march 2023, new public contracts code. Reorganization of regulations on public contracts for works, services, and supplies. Effective from 1 July 2023.

[198] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

[199] Rienties, B., Toetenel, L., and Bryan, A. (2015). "scaling up" learning design: impact of learning design activities on LMS behavior and performance. In Baron, J., Lynch, G., Maziarz, N., Blikstein, P., Merceron, A., and Siemens, G., editors, *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, LAK '15, Poughkeepsie, NY, USA, March 16-20, 2015*, pages 315–319. ACM.

[200] Rizzi, W., Francescomarino, C. D., Ghidini, C., and Maggi, F. M. (2022). Nirdizati: an advanced predictive process monitoring toolkit. *CoRR*, abs/2210.09688.

[201] Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

[202] Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc.

[203] Rodriguez, H., McAfee, P., and Wesson, C. (2022). Collusion detection in public procurement: A machine learning approach using global data. *Journal of Competition Law and Economics*, 18(2):335–371.

[204] Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

[205] Ronzani, M., Ferrod, R., Francescomarino, C. D., Sulis, E., Aringhieri, R., Boella, G., Brunetti, E., Caro, L. D., Dragoni, M., Ghidini, C., and Marinello, R. (2021). Unstructured data in predictive process monitoring: Lexicographic and semantic mapping to ICD-9-CM codes for the home hospitalization service. In Bandini, S., Gasparini, F., Mascardi, V., Palmonari, M., and Vizzari, G., editors, *AIxIA 2021 - Advances in Artificial Intelligence - 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1-3, 2021, Revised Selected Papers*, volume 13196 of *Lecture Notes in Computer Science*, pages 700–715. Springer.

[206] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

[207] Ruhl, J. and Katz, D. (2015). Measuring, monitoring, and managing legal complexity. *Iowa law review*, 101:191–244.

[208] Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Pearson, 3 edition.

[209] Sadiq, S. W. and Governatori, G. (2015). Managing regulatory compliance in business processes. In vom Brocke, J. and Rosemann, M., editors, *Handbook on Business Process Management 2, Strategic Alignment, Governance, People and Culture, 2nd Ed*, International Handbooks on Information Systems, pages 265–288. Springer.

[210] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

[211] Sànchez-Ferreres, J., Burattin, A., Carmona, J., Montali, M., Padró, L., and Quishpi, L. (2021). Unleashing textual descriptions of business processes. *Softw. Syst. Model.*, 20(6):2131–2153.

[212] Sangil, M. J. (2020). Heuristics-based process mining on extracted philippine public procurement event logs. In *2020 7th Int. Conf. on BESC*, pages 1–4.

[213] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[214] Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

[215] Sedrakyan, G., De Weerdt, J., and Snoeck, M. (2016). Process mining enabled feedback: "tell me what i did wrong" vs."tell me how to do it right". *Computers in Human Behavior*, 57:352–376.

[216] Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., and Kirschner, P. A. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior*, 107:105512.

[217] Sedrakyan, G., Snoeck, M., and De Weerdt, J. (2014). Process mining analysis of conceptual modeling behavior of novices–empirical study using jmermaid modeling and experimental logging environment. *Computers in Human Behavior*, 41:486–503.

[218] Senderovich, A., Gal, A., Mandelbaum, A., and Weidlich, M. (2020). Context-aware resource analysis in business processes. *Computers in Industry*, 115:103173.

[219] Shafranovich, Y. (2005). Common format and mime type for comma-separated values (csv) files. Request for Comments.

[220] Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

[221] Smith, J. (2019). Ethical web scraping in the age of big data. *Journal of Information Ethics*, 28(1):45–57.

[222] Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Yazdani Aminabadi, R., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., and Catanzaro, B. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. Available at arXiv:2201.11990 [cs.CL].

[223] Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

[224] Srinivasan, R. and Chander, A. (2021). Biases in ai systems. *Communications of the ACM*, 64(8):44–49.

[225] Stonebraker, M. (2010). Sql databases v. nosql databases. *Communications of the ACM*, 53(4):10–11.

[226] Sulis, E., Caro, L. D., and Nanda, R. (2024). Introduction for computer law and security review: special issue "knowledge management for law". *Computer Law & Security Review*, 52:105949.

[227] Sulis, E., Humphreys, L., Vernero, F., Amantea, I. A., Audrito, D., and Caro, L. D. (2022a). Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts. *Inf. Syst.*, 106:101821.

[228] Sulis, E., Humphreys, L. B., Audrito, D., and Di Caro, L. (2022b). Exploiting textual similarity techniques in harmonization of laws. In Bandini, S., Gasparini, F., Mascardi, V., Palmonari, M., and Vizzari, G., editors, *AIxIA 2021 – Advances in Artificial Intelligence*, pages 185–197, Cham. Springer International Publishing.

[229] Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.

[230] Tata, S. and Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–14.

[231] Tax, N., Verenich, I., La Rosa, M., and Dumas, M. (2017). Predictive process monitoring: A survey. *ACM Computing Surveys*, 50(6):1–34.

[232] Teinemaa, I., Dumas, M., Maggi, F. M., and Francescomarino, C. D. (2016). Predictive business process monitoring with structured and unstructured data. In Rosa, M. L., Loos, P., and Pastor, O., editors, *Business Process Management - 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings*, volume 9850 of *Lecture Notes in Computer Science*, pages 401–417. Springer.

[233] Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[234] Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[235] Torres Berru, Y., López Batista, V. F., Torres-Carrión, P., and Jimenez, M. G. (2020). Artificial intelligence techniques to detect and prevent corruption in procurement: A systematic literature review. In Botto-Tobar, M., Zambrano Vizuete, M., Torres-Carrión, P., Montes León, S., Pizarro Vásquez, G., and Durakovic, B., editors, *Applied Technologies*, pages 254–268, Cham. Springer International Publishing.

[236] Torres-Berru, Y. and López Batista, V. F. (2021). Data mining to identify anomalies in public procurement rating parameters. *Electronics*, 10(22).

[237] Torres Berrú, Y., Batista, V., and Torres-Carrion, P. (2020). Data mining to detect and prevent corruption in contracts: Systematic mapping review. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, pages 13–25.

[238] Torres-Carrión, P. V., González-González, C. S., Aciar, S., and Rodríguez-Morales, G. (2018). Methodology for systematic literature review applied to engineering and education. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1364–1373.

[239] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[240] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023b). Llama: Open and efficient foundation language models.

[241] Truong, H. M. (2016). Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in Human Behavior*, 55:1185–1193.

[242] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

[243] Turnbull, D., Chugh, R., and Luck, J. (2020). Learning management systems: An overview. In Tatnall, A., editor, *Encyclopedia of Education and Information Technologies*, pages 1–7. Springer International Publishing.

[244] Umer, R., Mathrani, A., Susnjak, T., and Lim, S. (2019). Mining activity log data to predict student's outcome in a course. In *Proceedings of the 2019 International Conference on Big Data and Education*, ICBDE '19, page 52–58, New York, NY, USA. Association for Computing Machinery.

[245] Umer, R., Susnjak, T., Mathrani, A., and Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning*, 10(2):160–176.

[246] Unger, A. J., Neto, J. F. d. S., Fantinato, M., Peres, S. M., Trecenti, J., and Hirota, R. (2021). Process mining-enabled jurimetrics: Analysis of a brazilian court's judicial performance in the business law processing. In *Proc. of 18th ICAIL*, page 240–244, NY, USA. ACM.

[247] van der Aalst, W. (2013). Decomposing petri nets for process mining: A generic approach. *Distributed and Parallel Databases*, 31(4):471–507.

[248] van der Aalst, W. M. (2019). Process mining in public administration: Principles and challenges. *IEEE Transactions on Services Computing*, 12(4):561–572.

[249] van der Aalst, W. M. P. (2016a). *Process Mining - Data Science in Action*. Springer.

[250] van der Aalst, W. M. P. (2016b). *Process Mining: Data Science in Action*. Springer, Berlin, Germany.

[251] van der Aalst, W. M. P. and Carmona, J., editors (2022). *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*. Springer.

[252] van der Aalst, W. M. P., Weijters, A. J. M. M., and Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142.

[253] Van Eck, N. J. and Waltman, L. (2014). Visualizing bibliometric networks. In *Measuring scholarly impact*, pages 285–320. Springer.

[254] Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

[255] Varoquaux, G. (2021). Using and understanding cross-validation strategies. *Giga-Science*.

[256] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[257] Velasco, R. B., Carpanese, I., Interian, R., Neto, O. C. G. P., and Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *Int. Trans. Oper. Res.*, 28(1):27–47.

[258] Villata, S., Araszkiewicz, M., Ashley, K. D., Bench-Capon, T. J. M., Branting, L. K., Conrad, J. G., and Wyner, A. (2022). Thirty years of artificial intelligence and law: the third decade. *Artif. Intell. Law*, 30(4):561–591.

[259] von Rosing, M., White, S., Cummins, F., and de Man, H. (2015). Business process model and notation—BPMN.

[260] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):841–887.

[261] Wafda, F., Usagawa, T., and Mahendrawathi, E. (2022). Systematic literature review on process mining in learning management system. In *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 160–166.

[262] Waltman, L., Van Eck, N. J., and Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4):629–635.

[263] Wang, L. (2005). *Support Vector Machines: Theory and Applications*, volume 177 of *Studies in Fuzziness and Soft Computing*. Springer Science & Business Media.

[264] Wang, X., Zhang, H., and Zhang, H. (2016). Detecting fraud in procurement contracts using machine learning techniques. *Journal of Government Information Quarterly*, 33(1):94–106.

[265] Wang, Y., Wang, L., Li, Y., He, D., Liu, T., and Chen, W. (2013). A theoretical analysis of NDCG type ranking measures. *CoRR*, abs/1304.6480.

[266] Weber, M. and Burkart, M. (2019). Governance in public procurement: Balancing openness, integrity, and efficiency. *Public Administration Review*, 79(1):73–81.

[267] Weijters, A. J. M. M., van der Aalst, W. M. P., and de Medeiros, A. K. A. (2006). Process mining with the heuristics miner algorithm. *Beta Working Paper Series*.

[268] Wright, R. E. (1995). Logistic regression. *Reading and Understanding Multivariate Statistics*, pages 217–244.

[269] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., et al. (2023). A survey on large language models for recommendation. *arXiv preprint*, arXiv:2305.19860. Available at arXiv:2305.19860 [cs.IR].

[270] Xu, W. (2019). Toward human-centered ai: A perspective from human-computer interaction. *Interactions*, 26(4):42–46.

[271] Yu, T. and Jo, I. (2014). Educational technology approach toward learning analytics: relationship between student online behavior and learning performance in higher education. In Pistilli, M. D., Willis, J., Koch, D., Arnold, K. E., Teasley, S. D., and Pardo, A., editors, *Learning Analytics and Knowledge Conference 2014, LAK '14, Indianapolis, IN, USA, March 24-28, 2014*, pages 269–270. ACM.

[272] Zhao, B. (2022). Web scraping. In Schintler, L. A. and McNeely, C. L., editors, *Encyclopedia of Big Data*, pages 951–953. Springer International Publishing, Cham.

[273] Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

[274] Zhou, Z. (2021). *Machine Learning*. Springer.

Appendix A

ANAC Data Model

This Appendix provides a comprehensive overview of the ANAC Data Model, focusing on its structure, data availability, and practical applications. It includes detailed guidance on querying the database and examples illustrating its use, particularly in analysing public tenders and related processes.

High-resolution images corresponding to this chapter are available at https://bit.ly/4eRONLT. The source code for the experiments conducted in this appendix can be accessed at the GitHub repository: https://github.com/roberto-nai/PhD-THESIS. The complete database described below is available on the Zenodo [65] platform at https://zenodo.org/records/12179651.

A.1 Data Records

This section describes the ANAC dataset merged with BDAP and ISTAT features based on the processing steps described in the previous section. Figure A.1 provides a general overview of the database tables, with primary and foreign keys, categorised in four *sections* according to their contents, i.e. *Main*, *Registries*, *Award*, and *Activities after award*.

The *Main* tables are `TENDER_NOTICE` (Figure A.1.a) which contains basic data on tenders and its related tables `PUBLICATIONS`, `WORK_CATEGORY` and `PNNR_REWARD_MEASURES`, all linked by the primary/foreign key *CIG*; the table is linked to the registry tables (Figure A.1.b) via various foreign keys (e.g., `CONTRACTING_AUTHORITIES` with *tax code* of the CA), that complement it.

The *Registries* tables (Figure A.1.b) contain precisely the registries of the CAs, linked via the foreign key *istat code* and *tax code* to the ISTAT and BDAP tables; this section also

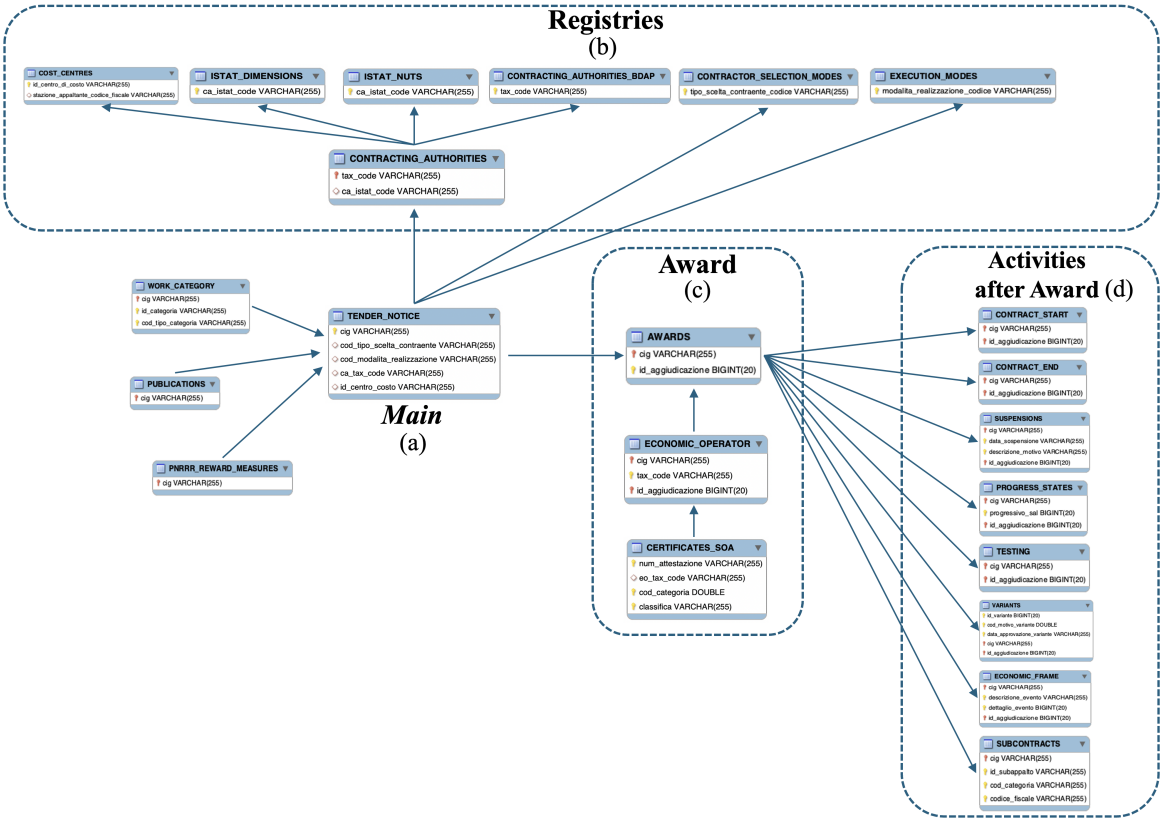


Fig. A.1 Data model of the ANAC dataset merged with BDAP and ISTAT. Schema generated with the freeware tool *MySQL Workbench* (https://dev.mysql.com/downloads/workbench)

contains the tables with the codes useful to classify a tender (EO selection codes and tender execution codes).

The *Award* tables contain data on tenders awarded by an EO: `AWARDS`, `ECONOMIC_OPERATOR` and `CERTIFICATES_SOA`, linked to the tender table via foreign key *CIG* (see `TENDER_NOTICE`).

The *Activities after the award* tables (`CONTRACT_START`, `CONTRACT_END`, `VARIANTS`, etc.) contain data on activities carried out through the tender after the award phase; these tables are linked to the `AWARDS` table via foreign keys *CIG* and *ID_AWARD*.

We provide an idea of the main tables' dimensions. In particular, table `TENDER_NOTICE` contains $3,336,360$ tenders, `CONTRACTING_AUTHORITIES` contains $435,18$ distinct CAs that created the tender notice, `AWARD` contains $1,830,388$ awarded tenders, and `ECONOMIC_OPERATOR` contains $1,828,831$ distinct EOs that awarded the tenders.

In the following, we describe each of the 22 tables included in ANAC dataset merged with BDAP and ISTAT (in alphabetical order):

1. `AWARDS`: Data on tender awarding. For each tender, it's possible to have multiple awards identified by a different award identifier (*id_award*); multiple awards occur in the event of the early ending of an award's revocation.

2. `CERTIFICATES_SOA`: Data on the SOA attestation, i.e. a document issued by a Certification Company following an investigation in which the possession of the requirements based on work carried out in the previous period. The certificate serves the company to prove, during the tender, its capacity to perform works belonging to a certain category of work and up to a certain amount.

3. `CONTRACT_END`: Data on the end of the contract between CA and EO referring to a specific tender;

4. `CONTRACT_START`: Data on the start of the contract between CA and EO referring to a specific tender;

5. `CONTRACTING_AUTHORITIES`: List of CA who created the tender; the two main primary keys (*tax_code* and *istat_code*) are used to connect with BDAP and ISTAT;

6. `CONTRACTING_AUTHORITIES_BDAP`: BDAP data organizational structure of CA and geographical coverage, identified by the primary key *tax_code*;

7. `CONTRACTOR_SELECTION_MODES`: List of criterion code - description for awarding a tender to an EO (e.g., classic tender, low budget, multi-year agreement, etc.);

8. `COST_CENTRES`: List of cost centre code - description on which a CA associates tender costs (e.g., public transport, local police, green);

9. `ECONOMIC_FRAME` List of the final costs of each tender (e.g. advertisement, consulting, material purchase, etc.);

10. `ECONOMIC_OPERATORS`: List of EOs who awarded tenders over time, after some verifications `CERTIFICATES_SOA`, identified by the primary key *tax_code*;

11. `EXECUTION_MODES`: Cost centre codes - description on which a CA associates tender costs;

12. `ISTAT_DIMENSIONS`: Population data by municipality identified via *istat_code*;

13. `ISTAT_NUTS`: List of NUTS data by municipality identified via *istat_code*;

14. `PNRRR_REWARD_MEASURES`: List of measure code - description listing the extra scoring criteria that can be attributed in the PNRR[1] tender awarding rankings (e.g. recruitment of staff with special needs, number of gender-equal employees, etc.);

[1] https://www.italiadomani.gov.it/content/sogei-ng/it/en/home.html

15. PROGRESS_STATES: Data on the progress (intermediate steps) of a tender;
16. PUBLICATIONS: Date of publication of the tender notice on official CA communication channels (e.g. GURI[2], TED[3], etc.);
17. SUBCONTRACTS: Data on eventual subcontracting between an EO and other suppliers to realize a part of the tender;
18. SUSPENSIONS: Data on eventual suspension of work on a tender (e.g.: weather problems, project problems, etc.);
19. TENDER_NOTICE: *Main* data, list of the tender created by CA in the various Italian regions from 2016 to 2023; starting from this table, most of the other tables are linked via the primary key *CIG*.
20. TESTING: Data about final checks of the Work, Supplies or Services of a tender;
21. VARIANTS: Data on variants (changes) to the originally awarded tender (e.g.: variants for urgent project requirements);
22. WORK_CATEGORY: List of category codes - description with the categorisation of each tender (e.g.: motorway work, bridge, viaduct, etc.).

A.2 Usage notes

A.2.1 Data Records availability

The ANAC dataset merged with BDAP and ISTAT database content is easily accessible through the Zenodo repository (https://doi.org/10.5281/zenodo.12179651) that contains an SQL schema of the database, as well as a CSV version of each table constituting this database. As a first step, execute the SQL script `ITH_db_catalogue.sql` in MySQL to create the database and its tables with primary and foreign keys. Next, import the CSV files into the various tables via the MySQL functionality (the names of the tables to be populated correspond with the names of the CSV files). Figure A.2 represents the command to be executed in MySQL to import each CSV file into the corresponding database table. Note the following order of data import (to avoid errors between primary and foreign keys): CONTRACTING_AUTHORITIES_BDAP, ISTAT_DIMENSIONS, CONTRACTING_AUTHORITIES, CONTRACTOR_SELECTION_MODES, EXECUTION_MODES, COST_CENTRES, TENDER_NOTICE, all other tables.

[2]https://www.gazzettaufficiale.it
[3]https://ted.europa.eu/en

Please note that some CSV files (e.g.: TENDER_NOTICE) are about 4.5 GB in size, so depending on the configuration of the workstation, importing data into the database may take several minutes, and it could be necessary to set some parameter such as execution timeout, importable file size, etc.

```
LOAD DATA INFILE 'TENDER_NOTICE.csv'
INTO TABLE TENDER_NOTICE
FIELDS TERMINATED BY ';'
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

Fig. A.2 Example of an SQL query to import a CSV file into the corresponding database table; the text in blue is related to the standard SQL syntax, while the text in orange (between the quotation marks) and black (without quotation marks) is the customisation of the query parameters

A.2.2 Queries on the database

The example in Figure A.3 refers to an example of a SQL query that extracts the tenders of the year 2016 by merging them with CA, award data, EO who awarded the contract, ISTAT and BDAP data on the region/province/municipality of the CA. The query selects all columns from the tables TENDER_NOTICE, CONTRACTING_AUTHORITIES, AWARDS, ECONOMIC_OPERATOR, ISTAT_NUTS, ISTAT_DIMENSIONS, and CONTRACTING_AUTHORITIES_BDAP. It performs a series of *left joins* to combine these tables based on matching primary/foreign keys.

```
SELECT
 TENDER_NOTICE.*, CONTRACTING_AUTHORITIES.*, AWARDS.*, ECONOMIC_OPERATOR.*,
 ISTAT_NUTS.*, ISTAT_DIMENSIONS.*, CONTRACTING_AUTHORITIES_BDAP.*
FROM
 TENDER_NOTICE
LEFT JOIN
 CONTRACTING_AUTHORITIES ON TENDER_NOTICE.ca_tax_code = CONTRACTING_AUTHORITIES.tax_code
LEFT JOIN
 AWARDS ON TENDER_NOTICE.cig = AWARDS.cig
LEFT JOIN
 ECONOMIC_OPERATOR ON TENDER_NOTICE.cig = ECONOMIC_OPERATOR.cig
LEFT JOIN
 ISTAT_NUTS ON CONTRACTING_AUTHORITIES.ca_istat_code = ISTAT_NUTS.ca_istat_code
LEFT JOIN
 ISTAT_DIMENSIONS ON CONTRACTING_AUTHORITIES.ca_istat_code = ISTAT_DIMENSIONS.ca_istat_code
LEFT JOIN
 CONTRACTING_AUTHORITIES_BDAP ON CONTRACTING_AUTHORITIES.tax_code = CONTRACTING_AUTHORITIES_BDAP.tax_code
WHERE
 TENDER_NOTICE.anno_pubblicazione = 2016;
```

Fig. A.3 Example of an SQL query to extract the tenders of the year 2016 by merging various tables; the text in blue is related to the standard SQL syntax, while the text in black is the customisation of the query parameters

A.2.3 Practical use of the dataset

To describe a possible analysis from the dataset, we provide an example of use: the analysis of public tenders for "green" services (i.e., CPV division number 90[4]) for CA types. First, the `TENDER_NOTICE` table has been filtered for *cpv_division* value 90; then, the `CONTRACTING_AUTHORITIES` table was cross-referenced with ISTAT tables (via *istat_code*) and BDAP table (via *tax_code*) to obtain information on the type and geographical location of the CAs themselves. Following, an analysis with georeferencing of tenders is proposed for the distribution of municipal expenditure by region and inhabitants. As a result, it was possible to generate a map of Italy with darker colours for regions that invested more and lighter colours for regions that spent less, as described in Figure A.4.

Fig. A.4 Map of Italy with darker colours for regions with higher investments and lighter colours for regions with lower expenditures for selected CPV division 90. The Italian map has been generated with the external library *Openpolis* (https://github.com/openpolis)

[4]The CPV codes starting with 90 encompass environmental services such as waste management, recycling, and consultancy. Many of the activities within this category are closely related to environmental protection.

Appendix B

TED Data Model

This Appendix focuses on the data records available from the TED public repository on the official EU website. It outlines the schema used to structure the dataset, highlighting the key variables that facilitate the analysis of public procurement processes.

B.1 Data Records

This section outlines the data schema for TED-related records[1], focusing on the two primary types of procurement data: *Common Framework Contracts* (CFCs) and *Common Award Notices* (CANs). The following subsections provide a detailed overview of the dataset's structure, key variables, and context.

B.1.1 Overview of the Dataset Schema

The TED data schema captures a wide range of information relevant to public procurement, enabling detailed analysis of contracting practices. It includes variables categorised by their application level, such as notice, lot, or contract award. Each variable is mapped to its corresponding schema version (XSD)[2], indicating its availability over time.

[1] https://ted.europa.eu/TED/main/HomePage.do

[2] The XSD version refers to the XML Schema Definition version used by the Publications Office of the EU, indicating the structure and quality of the data. Higher versions reflect enhancements in detail and consistency.

B.1.2 Accessibility and Use

The full dataset is publicly accessible[3], ensuring that researchers and practitioners can leverage it for transparency, accountability, and analytical purposes. The data's structured schema supports diverse applications, from compliance checks to process efficiency analyses.

B.1.3 Key Variables for CFCs and CANs

Table B.1 and Table B.2 summarise the core variables in the TED data schema for CFCs and CANs. Each variable is accompanied by its description, with shared descriptions applicable to both notice types. The tables feature four columns: variable names, descriptions for CFCs and CANs, and the level (e.g., notice, lot, or contract award), along with the XSD version from which the variable has been available.

[3] https://data.europa.eu/data/datasets/ted-csv?locale=en

Table B.1 Description of variables for CFCs and CANs (part 1/2)

Variable name	CFC Description	CAN Description	Level (since which version)
ID_NOTICE_CN	Unique identifier of the call for competition (usually contract notice)	N.A.	Notice
ID_NOTICE_CAN	N.A.	Unique identifier of the contract award notice / voluntary ex-ante transparency notice	Notice
TED_NOTICE_URL	Webpage of the notice on the TED website	Same as CFC	Notice
YEAR	Year of publication of the notice	Same as CFC	Notice
ID_TYPE	Standard form number	Same as CFC	Notice
DIRECTIVE	N.A.	The VEAT standard form can be used under several directives. This variable specifies the directive.	Notice
DT_DISPATCH	The date when the buyer dispatched (sent) the notice for publication to TED	Same as CFC	Notice
XSD_VERSION	Version of the XML schema definition used by the Publications Office of the EU to publish the data	Same as CFC	Notice
CANCELLED	1 = This notice was later cancelled	Same as CFC	Notice
CORRECTIONS	Number of later notices which corrected or added information to this notice	Same as CFC	Notice
FUTURE_CAN_ID	The publication ID of the CAN which followed this notice. This ID is used to link the CFC and CAN datasets.	N.A.	Notice
B_MULTIPLE_CAE	There is more than one contracting authority or entity	Same as CFC	Notice
CAE_NAME	Official name	Same as CFC	Notice
CAE_NATIONALID	National registration number	Same as CFC	Notice
CAE_ADDRESS	Postal address	Same as CFC	Notice
CAE_TOWN	Town	Same as CFC	Notice
CAE_POSTAL_CODE	Postal code	Same as CFC	Notice
ISO_COUNTRY_CODE	Country for the first listed authority	Same as CFC	Notice
B_MULTIPLE_COUNTRY	There are contracting authorities or entities from at least two different countries.	Same as CFC	Notice
CAE_TYPE	Type of contracting authority	Same as CFC	Notice
TYPE_OF_CONTRACT	Type of contract (Works, Supplies, Services)	Same as CFC	Notice
B_FRA_AGREEMENT	Y if the notice involves the establishment of a framework agreement	Same as CFC	Notice
FRA_ESTIMATED	Whether there are indications that this notice is actually about a framework agreement, even though it has not been marked as such by the buyer	Same as CFC	Notice
B_GPA	The contract is covered by the Government Procurement Agreement	Same as CFC	Notice
CPV	The main Common Procurement Vocabulary code of the main object of the contract	Same as CFC	Notice
B_OPTIONS	Options	Same as CFC	Notice
VALUE_EURO	CFC value, in EUR, without VAT	CAN value, in EUR, without VAT	Notice
VALUE_EURO_FIN_1	CFC value, in EUR, without VAT. If the value variable is missing, the framework value is used instead	CAN value, in EUR, without VAT. If the value variable is missing, this variable looks for it in all other fields	Notice
DURATION	Duration of the contract, framework agreement, or dynamic purchasing system in months	Same as CFC	Notice
CONTRACT_START	Starting	Same as CFC	Notice
CONTRACT_COMPLETION	Completion	Same as CFC	Notice
TOP_TYPE	Type of procedure (Competitive Dialogue, Negotiated, Open, etc.)	Type of procedure (Award without prior publication, Competitive Dialogue, Negotiated, Open, etc.)	Notice
B_ACCELERATED	The option to accelerate the procedure has been used	Same as CFC	Notice

Table B.2 Description of variables for CFCs and CANs (part 2/2)

Variable name	CFC Description	CAN Description	Level (since which version)
OUT_OF_DIRECTIVES	N.A.	The procurement falls outside the scope of application of the directive but a CAN was published anyway.	Notice
ENV_OPERATORS	Envisaged number of operators	Same as CFC	Notice
ENV_MIN_OPERATORS	Envisaged minimum number	Same as CFC	Notice
ENV_MAX_OPERATORS	Envisaged maximum number	Same as CFC	Notice
CRIT_CODE	Award criteria (Lowest price, Most economically advantageous tender)	Same as CFC	Notice
CRIT_PRICE_WEIGHT	Weight given to price	Same as CFC	Lot (since XSD 2.0.9)
CRIT_CRITERIA	Information on award criteria. For XSD < 2.0.9, the variable is usually unstructured text.	Same as CFC	Lot (since XSD 2.0.9)
CRIT_WEIGHTS	Information on award criteria weighing	Same as CFC	Lot (since XSD 2.0.9)
B_ELECTRONIC_AUCTION	An electronic auction has been used	Same as CFC	Notice
NUMBER_AWARDS	N.A.	The number of CAs for a given CAN	Notice
DT_APPLICATIONS	Time limit for receipt of tenders or requests to participate	Same as CFC	Notice
B_LANGUAGE_ANY_EC	Language(s) in which tenders or requests to participate may be drawn up - Any EU official language	Same as CFC	Notice
ADMIN_LANGUAGES_TENDER	Language(s) in which tenders or requests to participate may be drawn up - Official EU language(s):	Same as CFC	Notice
ADMIN_OTHER_LANGUAGES_TENDER	Language(s) in which tenders or requests to participate may be drawn up – Other:	Same as CFC	Notice
B_RECURRENT_PROCUREMENT	This is a recurrent procurement	Same as CFC	Notice
ID_AWARD	Unique Award identifier	Same as CFC	Award
ID_LOT_AWARDED	Lot No, an identifier of a lot within this Award	Same as CFC	Award
INFO_ON_NON_AWARD	N.A.	If the variable is empty, then a contract was awarded	Award
WIN_NAME	Official name	Same as CFC	Award
WIN_NATIONALID	National registration number	Same as CFC	Award
WIN_ADDRESS	Postal address	Same as CFC	Award
WIN_TOWN	Town	Same as CFC	Award
WIN_POSTAL_CODE	Postal code	Same as CFC	Award
WIN_COUNTRY_CODE	Country	Same as CFC	Award
B_CONTRACTOR_SME	N.A.	The contractor is an SME (as defined in Commission Recommendation 2003361EC)	Award (only in XSD 2.0.9)
CONTRACT_NUMBER	Contract number	Same as CFC	Award
TITLE	Title of the contract	Same as CFC	Award
NUMBER_OFFERS	Number of tenders received	Same as CFC	Award
NUMBER_TENDERS_SME	Number of tenders received from SMEs (as defined in Commission Recommendation 2003/361/EC)	Same as CFC	Contract award (since XSD 2.0.9)
NUMBER_TENDERS_OTHER_EU	Number of tenders received from tenderers from other EU Member States	Same as CFC	Contract award (since XSD 2.0.9)
NUMBER_TENDERS_NON_EU	Number of tenders received from tenderers from non-EU countries	Same as CFC	Contract award (since XSD 2.0.9)
NUMBER_OFFERS_ELECTR	Number of offers received by electronic means	Same as CFC	Contract award
AWARD_EST_VALUE_EURO	Estimated contract award value, in EUR, without VAT	Same as CFC	Award
AWARD_VALUE_EURO	Total final contract award value, in EUR, without VAT. If the value was not present, the lowest bid is included.	Same as CFC	Contract award
AWARD_VALUE_EURO_FIN_1	Final contract award value, in EUR, without VAT. If the value variable is missing, this variable looks for it in all other fields.	Same as CFC	Award
B_SUBCONTRACTED	The contract is likely to be subcontracted	Same as CFC	Award
DT_AWARD	Date of contract award	Same as CFC	Award