Towards modelling beef cattle management with Genetic Programming

Type of the Paper (Research Article)

Towards Modelling Beef Cattle Management with Genetic Programming

Francesca Abbona [1,4,*], Leonardo Vanneschi [2,3], Marco Bona [4] and Mario Giacobini [1]

[1] University of Torino, Largo Paolo Braccini 2, 10095 Grugliasco, Turin, Italy;

[2] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal;

[3] LASIGE, Departamento de Informàtica, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

[4] Associazione Nazionale Allevatori Bovini Razza Piemontese, Carrù, Italy;

* Corresponding author.

E-mail addresses:

francesca.abbona@unito.it (F. Abbona), lvanneschi@novaims.unl.pt (L. Vanneschi), marco.bona@anaborapi.it (M. Bona), mario.giacobini@unito.it (M. Giacobini).

Abstract: Among the Italian Piemontese Beef Breedings, the yearly production of calves weaned per cow, that is the calves that survive during the period of 60 days following birth, is identified as the main target expressing the performance of a farm. Modelling farm dynamics in order to predict the value of this parameter is a possible solution to investigate and highlight breeding strengths, and to find alternatives to penalizing factors. The identification of such variables is a complex but solvable task, since the amount of recorded data among livestock is nowadays huge and manageable through Machine Learning techniques. Besides, the evaluation of the effectiveness of the type of management allows the breeder to consolidate the ongoing processes or, on the contrary, to adopt new management strategies. To solve this problem, we propose a Genetic Programming approach, a white-box technique suitable for big data management, and with an intrinsic ability to select important variables, providing simple models. The most frequent variables encapsulated in the models built by Genetic Programming are highlighted, and their zoological significance is investigated *a posteriori*, evaluating the performance of the prediction models. Moreover, two of the

final expressions selected only three variables among the 48 given in input, one of which is the best performing among GP models. The expressions were then analyzed in order to propose a zootechnical interpretation of the equations. Comparisons with other common techniques, including also black-box methods, are performed, in order to evaluate the performance of different type of methods in terms of accuracy and generalization ability. The approach entailed constructive and helpful considerations to the addressed task, confirming its key-role in the zootechnical field, especially in the beef breeding management.

Keywords: Precision Livestock Farming; Evolutionary Algorithms; Machine Learning; Cattle Breeding; Piemontese Bovines.

1. Introduction

The management of livestock by continuous automated real-time monitoring of production, reproduction, health and welfare of the herd, and its environmental impact is defined as Precision Livestock Farming (PLF) (Berckmans, 2017; Berckmans and Guarino, 2017). PLF supplements the skills of the farmer, the veterinarian, and the technician by a continuous collection of livestock information, with the support of information technologies. It can play a crucial role in the early detection of diseases and it objectively assesses animal condition and welfare in modern livestock production, representing a tool that supports many farmers as decision-makers. Despite the biological process is too complex to replace farmers by technology, it still offers more possibilities to save money and to change farmers' lives, as a more accurate management system can be achieved, leading to better approach of the genetic potential of today's livestock species. The breeder must generally deal with animals' problems, like their health conditions and social behavior, that affect the quality of the product, the life of the animal, and the performance of the farm. Indeed, PLF is an emerging field in which the main aspect involves the development of proper tools and novel technologies, suitable to monitor each animal and the whole breeding. The resulting increased knowledge, elaborated through mathematical models, may provide the offset of overall incurred costs of the farm, as these issues are identified in advance, allowing decisions to be made in time.

In a recent study conducted by CREA of Lodi, a Research Centre for animal production and aquaculture, a survey was addressed to cattle farmers in one of the most intense dairy breeding provinces in Italy, to investigate the diffusion of precision farming tools (Abeni et al., 2019). Most breeders reported the use of electronic equipment, and the owners of larger farms showed a greater propensity for PLF technology, stating that, considering costs and benefits, the biggest

problem in purchasing monitoring systems is the time required to manage the generated data. The major consequence of continuous monitoring of animals is a huge amount of data, the so-called “Big Data” (Cole et al., 2012; Lokhorst et al., 2019). Given the numerosity and complexity of these data, the databases cannot be visually inspected. If on one side PLF approach aims for a greater "accuracy" on the quantity and quality of information, entailing the development of monitoring systems, on the other side it must deal with the transformation of big data into meaningful information. For this reason, the use of Machine Learning (ML) techniques is becoming increasingly common. ML is a subfield of artificial intelligence, addressed to the study of algorithms for prediction and inference. Learning from data is at the core of ML, and hence this field of research is suitable for the management of large data sets and used to predict livestock issues, such as time of disease events, risk factors for health conditions, failure to complete a production cycle, as well as the genome of complex traits (González-Recio et al., 2014;Morota et al., 2018).

Many are the research studies in the cattle sector based on the application of ML techniques in farm management, to model for example the individual intake of cow feed, to optimize health and fertility, and to identify potential disease predictors for several pathologies, such as Bovine Viral Diarrhoea Virus (BVDV), Infectious Bovines Rhinotracheitis (IBR), Bovine Tuberculosis (TB), lameness, and mastitis (Amrine et al., 2014; Bovine Diseases and Resources; Guzhva et al., 2016; Machado et al., 2015; Nasirahmadi et al., 2017; Ortiz-Pelaez and Pfeiffer, 2008; Rodero et al., 2012; Williams et al., 2016; Yao et al., 2016). Beef cattle management is usually an extensive breeding system: the bovines are bred completely on pastures and they are not subject to massive checks such as dairy cattle, since they are more resistant and less exposed to stress factors (Derner et al., 2017). Therefore, a precision farming approach is rarely used in these cases. However, nowadays the number of new intensive farms is also increasing in the meat sector (Cozzi et al. 2009). The Piemontese, mostly concentrated in the Italian region of Piedmont, belongs to this category, mainly because available pastures are not sufficient for the total number of animals (Bona et al., 2005; Savoia et al., 2019). Consequently, in order to optimize the management of this type of breeding, it is necessary to constantly monitor the animals, introducing and adapting to beef cattle the necessary tools implemented for the dairy sector. The latter already offers a wide range of devices, because dairy bovines generally have a shorter average life compared to the lifespan of beef bovines. They are often affected by diseases and metabolic problems and are crossed often with beef cattle for better performance and higher yield (Rutten e al., 2013; Hessle et al., 2019). The breeding cycle is reduced compared to other income-producing species, and there is no daily movement of the animals (e.g. milking). The aspects of greatest interest are the composition of the ration and the consumption of food and water, behavioral remarks, the quality of the structures that host the cattle (temperatures, humidity, lighting), growth, slaughter yield, and carcass quality. The

unit of measurement is not individual in the case of beef cattle, but each animal affects the performance of the farm after all. The lower impact of critical points in the meat sector entails that the adoption of sensors, not yet specific for this type of animal and with a high cost, is probably not worth the economic investment. The lower impact of critical points in the meat sector entails that the adoption of sensors, not yet specific for this type of animal and with a high cost, is probably not worth the economic investment.

Among the major beef cattle breeds, Italian 'Piemontese' represents a characteristic element of the territory of Piedmont, a region in Northwestern Italy. Organoleptic and zootechnical remarkable qualities result above all in a greater tenderness of the meat and exceptional character skills, such as meekness, maternal attitude, resistance to diseases, little stress, and great adaptation to pasture. It, therefore, allows easy management and the development of the local area ("Associazione Nazionale Allevatori Bovini Razza Piemontese"; Bona et al., 2005). The Piemontese cattle derives its name from this region, its cradle of origin, even if today it is spreading in several foreign countries. The bovines are bred in beef intensive farms, which are therefore provided with the installation of stables to control the animals, a grazing for feeding purposes, the addition of different artificial fodder on feed and curative intents, and particular attention to the reproduction of the livestock.

Since the National Association of Piemontese Cattle Breeders (ANABORAPI) is responsible for promoting the breed through the study of all processes of the Piemontese breeding (Lo svezzamento del vitello Piemontese, 2012; Relazione tecnica, 2018), information is stored in a complex database, i.e. the Herd Book of the Race. Besides animals' pedigrees, morphological details, and genetic values, a wide section of statistics among the economic efficiency of the farm are available. In particular, a constant monitoring of the average situation of breedings due to the main fertility parameters are provided, summarized by the average number of calves per cow produced in the last year. This is then translated into a brief economic summary, which compares the gross revenue with the mortality losses, providing the farmer with an indicator of breeding performance. However, this index does not include the effects of the weaning period. The physiological development process of the animal reaches completion in 60 days after birth. Calf mortality is an important cause of economic damages in Piemontese cattle farms: it represents for the farmer the loss of the economic value of the calf, and the reduction of both the herd's genetic potential and the size of the breeding. It is straightforward that the gestational phase alone is not exhaustive: it is crucial to consider neonatal mortality, outlining the calf's ability to survive, and the source of stress such as congenital calf's defects (i.e. arthrogryposis and macroglossia), compromising eventually the immune response and the growth rate, environmental and food conditions, that affect the quality of life of the newborn (Prince et al., 2003; Tao et al., 2018). Therefore, these zoological influential variables must be identified

among the numerous parameters within the complex database. Given the size of the databases, it is extremely difficult to recognize many of the substantial factors, and to be able to hypothesize a prediction model for the number of calves weaned per cow per year, that is a more accurate measure for the yield of the farm.

In order to investigate the production of Piemontese calves and its modelling, this study aims to examine which variables available in the dataset influence the performance of a breeding. In contrast with previous studies conducted by ANABORAPI, in which models are based on traditional statistical identification approaches, a priori assumptions about data or the relationship between the response and independent variables are avoided.

In ML approaches, the choice of which techniques lend best to the problem depends mostly on the objective to be achieved. There are many methods that can produce excellent results, by building accurate prediction models. However, there are different characteristics, intrinsic to the techniques, capable of better address the question that arisen, and which tend to be privileged in the studies. Therefore, since the task and the data are full of zoological meaningful features, a Genetic Programming (GP) approach is adopted as a baseline, as models are resumed in simple and interpretable expressions (Abraham et al., 2006; Koza, 1994). Thanks to its structure, the algorithm, belonging to the white-box methods category, can automatically create models "learning" from the data, process accurate results, and even extract critical information. Among other ML techniques, some common methods are selected, including black-box ones as Neural Networks and Random Forest to compare the results obtained with GP. Black -box model are generally outperforming, since their structure is able to capture the high non-linearity lying on data. However, as their definition suggests, they can be very unclear and do not explain the links between input and output variables, as well as the internal mechanisms leading to the results (Loyola-González, 2019).

The article is organized as follows: Section 2 is dedicated to the description of the background: the current applied method, modeling the breeding performance, is presented and the aim of the study is highlighted. The dataset is analyzed and the basic steps to prepare the benchmark are also described. Emphasis is placed on the division of the dataset into different partitions, illustrating the need for the techniques to learn on a portion of the dataset and to test the prediction models on new instances. Afterwards, the GP baseline and other ML methods are enlightened. Results are examined and discussed in Section 3, with regard to the readability of the expressions. The performance of GP is illustrated, and the comparison with other techniques is discussed. The research article concludes with some considerations and further developments, highlighted in Section 4.

2. Materials and Methods

2.1 The reference model

The model used to monitor the farm performance estimates the number of live calves produced per cow per year. It is a classic statistical model, formulated among zootechnical hypotheses, and it incorporates two variables extracted from the information of the single farm: the average calving interval (*intp*), that is the time span measured in days between a birth and the previous one, and the average calves mortality at birth:

$$Y_a = \frac{365}{intp}\left(1 - \frac{m}{100}\right). \quad (1)$$

It concerns the birth, taking into account the average values of the previous 12 months, without taking into account important factors related to the weaning of the calf. The effects of the two months following this event are significant, since in the immediate period following the birth, i.e. 60 days, the physiological development of the calf reaches completion. Multiple factors can affect the growth, such as its own ability to adapt, the genetic factors, congenital calf's defects, such as arthrogryposis or macroglossia (Lynch et al., 2019). Moreover, the environmental and food conditions can contribute positively or negatively to the quality of life of the newborn, becoming a remarkable source of stress, compromising the growth rate and the well-being of the animals. Among the farms considered in the study, (description in Section 2.2), we compared the reported number of calves that died at birth and the sixtieth day after. As straighforward from Figure 1, during birth almost all the farms did not report any deaths, while at the end of weaning the number of farms with zero deaths drops drastically.

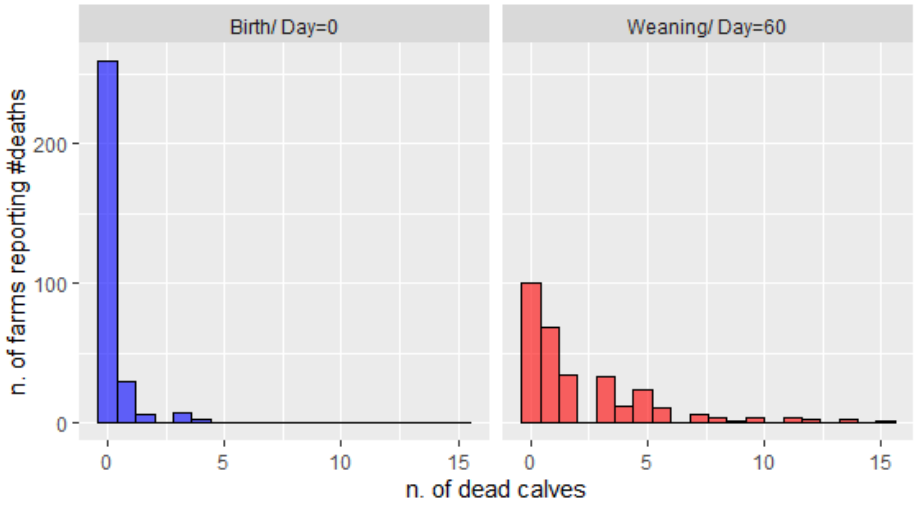

Figure 1. Distribution of reported deaths for 304 farms during 2017, respectively at birth and after 60 days. All the breedings show extremely different values between the dead calves at birth (in blue) and after 60 days (in red) (Kruskal-Wallis test: p-value << 0.001).

In fact, it is noted that the breeders reported a high number of dead calves at 60 days. For the farmer, the loss of the calf means the loss of economic value. The high mortality rate reduces then the number of young animals to be used to increase the farm size and the genetic potential of the herd. This makes it necessary to formulate a model that can predict the number of calves weaned per cow per year, shaped on data. Like equation (1), it should incorporate the influential variables affecting the output, and at the same time provide a not linear expression of simple interpretation, in order to be able to understand and explain zoologically the link with the output afterwards. On this porpouse, we approach the problem with GP, since it can produce accurate and explainable models, uncovering relevant predictors, with similar performance to other well-extablished common methods, such as linear regression, k-nearest neighbour, neural networks, and random forest.

2.2 Datasets and ML framework

The available database provided by ANABORAPI is the event history for all farms registered in the Herd Book of the Race. For every farm the average values of the measurements, recorded by technicians during routine controls, veterinarians, and directly by farmers, are stored. Data are registered with a devoted device available to technicians (a mobile computer Workabout) or directly with the smartphone and the personal computer. In fact, ANABORAPI designed a web service, accessible to registered users, which provides the situation of the farm. The data entered, both from the PC and from the other devices, are sent in real time to the servers, stored and processed, to return the updated situation at last. There are several records for each farm, since it keeps track of every visit. The content of the database is processed by the system on the date of elaboration going back 365 days, starting from the last check. Statistics are finally provided, resumed in a web page that the farmers can consult on their own. In addition to ID data of the breeding farms, all information on the consistencies, information on the parturitions and births, the type of inseminations carried out (natural or artificial), controls' dates, Estimated Breeding Values (EBV), consanguinity of all registered bovines, and perinatal mortality rates are kept. Since the database contains data of the last twenty years of all farms, including those that are no longer active, we have processed it in order to create a representative data set on which to apply the ML techniques.

In ML, it is necessary to define a set of data intended for "learning" and a set on which to "test" the obtained models. In other words, the learning set is a list of records (instances) on which the technique will build a knowledge

base. The algorithm analyses the data to find links between a series of input variables (i.e. consistencies, CI, mortality, EBVs, consanguinity, etc.) and a specific output variable (the target variable, i.e. the number of calves weaned per cow per year). This type of task is called supervised learning since the attribute to be predicted is known (target attribute). The test set must verify the effectiveness of the model just constructed, subjecting it to new situations, by checking the validity of its answers. Thus, the learning set is required to build the model, while the test set to measure its effectiveness. It is therefore strictly necessary for the two sets to be disjoint. For an optimal management, besides the contemporary situation of the farm, it is relevant for the breeder to know the prediction of the future trend. Therefore the core of the issue is concentrated in the prediction of the number of viable calves after the weaning phase, which each cow will produce over the next 365 days. The conditions that allow the survival of the calf are partly related to the calf itself and its temperament. However, in large part they are due to its genetic characters and those of its ancestors, and even to the choices that the breeder adopts for the management of the animals and the environment. The variable which serves to supervise the construction of the model is the target value for the following year. Considering for each farm the corresponding values in 2018 for the number of the calves born alive and those unable to survive during weaning period and the number of cows, (i.e. N_BALIVE, N_ELIM, $COWS$), the target attribute Y was obtained with the following:

$$Y = \frac{N\_BALIVE - N\_ELIM}{COWS}.$$

In a preliminary study, we investigated the dataset and performed a Genetic Programming approach, in order to explore the possibility to address this task with GP (Abbona et al., 2020). The method performed well, entailing that ML horizon should be investigated further and that comparisons with other techniques should be carried out, even on larger datasets containing more features. In the previous experiment we extracted and processed 19 variables and we kept stricter filters on data: to perform GP method, we selected the farms, based on the date of checks recorded between 2017 and 2018. In this study, only the farms that show constant visits between 2014 and 2019 were considered. In this way, the effects related to farm management are consolidated and just the breedings with substantial data were kept. Indeed, even if the investigation is based on farms with data from 2017 as input and from 2018 as target, as a change in the type of management stabilizes over time, we considered breedings with historical records updated between 2014-2019, in order to focus on farms with a solid management. A newly started company does not have completely representative data. Moreover, the summary produced by anaborapi elaborates the average values among recordings

related to the previous 65 days from the reference date. To avoid data from farms not yet fully operational, with gaps in registrations or close to resigning at the end of 2018, we set the restriction to companies active in the previous 5 years.

As in our pilot study, filters were imposed on breedings located in Piedmont with at least 30 cows and a percentage of artificial insemination between 90% and 100%. For each breeding, data recorded in 2017 and 2018 were considered and the record for the last check in the corresponding year was considered. Subsequently, in correspondence with each farm, input and target variables were extracted, respectively from 2017 to 2018. However , two further conditions were added for this research study. As already stated, in order to keep in the pool of currently active breedings those with stable and consolidated situations, inspections had to be constantly carried out for at least more than two years. Hence, as already explained, only farms exhibiting constant data recordings during the time interval 2014-2019 were considered. Moreover, the fact that breeders carry out between 90% and 100% of artificial insemination, means that a part of the considered farms own bulls and carry out also natural impregnations. Most of the time, instead of recording the date on which the insemination took place, a period of several days followed by the diagnosis of the pregnancy is set. These farms were therefore excluded from the analysis. A main group of 304 representative Piemontese cattle farms results from the selection. Since the performance of the farm mainly focuses on fertility, data concerning multiparae cows were considered to elaborate the number of deliveries and the calving intervals. In the same way, data on bulls used for artificial insemination were maintained (i.e. EBVs, that represent namely estimations of the additive genetic effect of a subject). Parameters on heifers were included in the dataset, since these are bovines that did not give birth but, in breeding farms, are mostly intended to the production of calves. Moreover, since many breeders carry out also natural impregnation besides artificial insemination, related to the bulls used for natural insemination were added to the analysis, as well as the levels of consanguinity of calves that will be born from ongoing pregnancies. The only strictly environmental measure available in the dataset, that was hence kept, is the Livestock Unit (LU or LSU): it has the purpose of synthetically expressing the zootechnical load, to easily compare the environmental impact of different farms. Based on the age of the animals, appropriate coefficients are applied to the number of animals for each age category in the breeding: cattle over 2 years old (1 * LSU), cattle aged between 6 months and 2 years (0.6 * LSU) and cattle less than 4 months old (0.4 * LSU) (Sistema Piemonte-UBA).

The final dataset counts 304 records, each one standing for a single farm, and a total of 48 input attributes (referring to year 2017) and one target variable, that is the actual number for weaned calves recorded in 2018. All variables represent positive quantities and are described in Table 1.

Table 1. Final attributes used in the studied dataset. The last line (variable Y) represents the dependent variable, target attribute for the predictive models generated by ML techniques.

	Attribute	Description
1	***BOVINES***	Consistency of all animals in the farms, i.e. size of the cattle.
2	***COWS***	Consistency for cows, i.e. number of cows.
3	***HEIFERS***	Consistency for heifers, i.e. number of heifers.
4	***F_CALVES***	Consistency for female calves, i.e. number of female calves.
5	***BULLS***	Consistency for bulls, i.e. number of bulls.
6	***M_CALVES***	Consistency for male calves, i.e. number of male calves.
7	***PERCENT_FA***	Percentage of Artificial Insemination.
8	***C_AGE***	Mean value of age of cows expressed in days.
9	***C_PAR***	Mean number of parturitions per cow.
10	***N_PAR***	Number of occurred deliveries.
11	***SALXGRAV***	Mean number of necessary inseminations, which resulted in positive pregnancy diagnosis.
12	***N_CORRECT***	Percentage of calves born without birth defects, such asMacroglossia or Arthrogryphosys.
13	***H_EASE***	Number of easy parturitions for primiparae, that did not require human intervention and that did not cause stress
14	***H_DIFFICULT***	Number of difficult parturitions for primiparae, that partly required human intervention.
15	***H_CESAREAN***	Number of parturitions that required caesarean section for primiparae.
16	***C_EASE***	Number of easy parturitions for multiparae, that did not require human intervention and that did not cause stress
17	***C_DIFFICULT***	Number of difficult parturitions for multiparae, that partly required human intervention.
18	***C_CESAREAN***	Number of parturitions that required caesarean section for multiparae.
19	***C_N_IND***	Number of cows that gave birth in the last year.
20	***C_PARTIND***	EBV referred to ease of parturition of the cows.
21	***H_PARTIND***	EBV referred to ease of parturition of the heifer.
22	***N_TFA***	Number of AI bulls whose semen has been used during the last year.
23	***TFA_BIRTH***	Mean value of EBV referred to ease of birth of the bulls, which semen has been used on artificial inseminations
24	***TFA_PAR***	Mean value of EBV referred to ease of parturition with which the bulls, which semen has been used on artificial
25	***N_TFN***	Number of NI bulls used during the last year.
26	***TFN_ BIRTH***	Mean value of EBV referred to ease of birth of the bulls, which is used for natural impregnation
27	***TFN_PAR***	Mean value of EBV referred to ease of parturition with which the bulls, which is used for natural impregnation, were
28	***C_GRAVID***	Number of pregnant cows at the end of 2017.
29	***C_INS***	Number of inseminated cows at the end of 2017.
30	***C_POSTPARTUM***	Number of post-partum cows at the end of 2017.
31	***C_EMPTY***	Number of cows in dry period.
32	***LSU***	Total Livestock Unit.
33	***LSU1***	Livestock Unity for bovines older than one year.
34	***LSU06***	Livestock Unity for bovines between 6 months and 1 year old.
35	***LSU04***	Livestock Unity for bovines between 4 and 6 months old.
36	***INTP***	Mean calving interval, i.e. the average number of days that elapse between a parturition and the following one.
37	***CONSANG_NEW***	Level of consanguinity of calves that will be born from ongoing pregnancies.
38	***N_CONSANG_NEW***	Number of ongoing pregnancies.
39	***BIRTHW_M***	Mean birth weight of male calves.
40	***BIRTHW_M***	Mean birth weight of female calves.
41	***MORT***	Mean value of neonatal mortality.
42	***ABORT***	Percentage of abortions.
43	***N_ABORT***	Number of abortions.
44	***N_ELIM***	Number of calves dead within 60 days after their birth.
45	***N_TOT***	Total number of newborns.

46	***N_BALIVE***	Total number of calves born alive.
47	***BORN_FA***	Total number of newborns with artificial insemination.
48	***BORN_FN***	Total number of newborns with natural impregnation.
49	***Y***	Number of expected weaned calves per cow per year.

2.3. Application of ML techniques: basic steps

As previously described, one of the basic steps for applying ML techniques is the subdivision of the dataset into two disjoint parts: the learning set and the test set. Therefore, the main dataset is split, the chosen algorithm builds a model, learning the hidden relations between the data on the "Learning set", and its generalization performances are finally evaluated on the “Test set”. In order to compare the performance of different approaches, it is necessary to analyse the median behaviour of the models obtained. Therefore, we split 30 times the dataset in order to obtain 30 different sets, each one with constant learning-test partitioning (75%-25%) and randomly selected instances. In this way, each experimental phase will determine one solution on the corresponding dataset, for a total of 30 prediction models for each technique. Regarding GP, the "Learning set" was further split, exactly in a similar way to the division into learning and test set. Each of the 30 learning sets was randomly divided, with a constant partitioning (75% -25%), into a Training Set and a Validation Set. The choice of this methodology, i.e. a 75%-25% split repeated for both partitions, is due to the size of the dataset. The division between learning and test entailed a learning set of size equal to 228 instances. We initially considered partitioning the learning set in training-validation through a *k*-fold cross validation, but the reduced size did not allow us to find a suitable value of *k*: for example, k smaller than 10 leads to a restrained number of training-validation subsets. On the contrary, a value of *k* greater than 10, led to a restrained number of records within the validation sets (i.e. less than 21 farms). Using a 30-fold would imply a validation of size 7, not representative at all. For this reason, we repeated further the subdivision 75%-25% to obtain disjoint training and validation sets, paying attention to avoid too much overlap between the 30 subsequent partitions. Finally, the sizes for learning and test were respectively 228 and 76, whereas 171 and 57 for training and validation.

Very often, during the construction and development of a model, two opposite problems may appear: underfitting or overfitting (Domingos, 2012; Bhattacharya, 2013). The first arises when the created model is too simple and fails to generalize neither on the learning set nor on the test set. In the opposite case, the model adapts extremely well to the learning set, but fails to generalize, which leads to small errors on the learning set and very large on the test. The noise, naturally present in the dataset, can result in a problem: the model learns the noise instead of the true hidden

relationship among the variables. Setting correctly the regularization hyperparameters with which the technique learns and adding complexity to the model are crucial steps. Once the parameters are set and the model is obtained, the error is evaluated. When dealing with ML, the error is measured with a fitness function, that is an objective function that is used to evaluate how close a given solution is to achieving the experimental aims. In this case, since we are dealing with a regression problem, i.e. we want to calculate the continuous value of a dependent variable starting from the independent variables, we chose the Root Mean Square Error (RMSE) as fitness function:

$$RMSE = \sqrt{\sum_i \frac{\left(y_i - \phi(x_i)\right)^2}{n}},$$

where $i=1,\ldots,n$ and n is the number of instances (depending whether it is calculated on the learning or the test set). The predictor ϕ is evaluated at x_i (values of input variables on 2017) and y_i are the target values (on 2018). A good fitness value means a small RMSE and viceversa. Moreover, RMSE is expressed in the response variable's unit and it is an absolute measure of accuracy. The choice of this fitness function was also determined by the comparison of different ML techniques, that build non-linear models. This issue excludes a discussion based on the coefficient of determination R^2, as its definition assumes linearly distributed data. When the assumption is violated, R^2 can assume misleading values (Spiess and Neumeyer, 2010).

Finally, GPLab package built in Matlab ("The package 'caret'"; Silva, 2007) was performed, and comparisons with other techniques were applied to the benchmark problem, with R software library caret (Table 2).

Table 2. ML techniques adopted, the corresponding belonging area and the respective used package.

Method	Description	Package
'GP'	Genetic Programming based algoritm (GP)	GPLab library built in Matlab
'knn'	k-Nearest Neighbour algorithm (kNN)	R software library caret
'nnet'	Neural Network algorithm (NN)	R software library caret
'lm'	Linear regression algorithm (LM)	R software library caret
'ranger'	Random Forest Tree-based algorithm (RF)	R software library caret

2.3 . Application of ML techniques: Genetic Programming

The algorithm based on GP creates a population of models, whose number is set by the user in the parameters' settings (Silva, 2007; Poli et al., 2008). It is a tree-based algorithm, which, exactly as in an evolutionary process, with the

passing of generations lets the initial population evolve, through mechanisms of selection, mutation, and recombination of individuals (i.e. mutation and crossover). GP evolves individuals, represented as tree structures, that can be recursively evaluated. The tree nodes are operator functions and every terminal node is an operand. By selecting, recombining, and mutating the best individuals at each generation, at each evolutionary step (i.e. new generation), the members of the new population are on average fitter than previously generated individuals, i.e. show a smaller error.

The user sets the size of the population, initially constructed randomly by the algorithm, to find at the last generation (parameter determined also by the user) a population of the same size but with evolved individuals. In our implementation, initial population was generated with the Ramped half and half method: half the initial population is constructed using the *full* method (generates trees where all the leaves, i.e. the variables, are at the same depth) and half is constructed using *grow* method (creation of trees of different sizes and shapes). Among other parameters, it is possible to set the conservation of the best individual at each run (Elitism) and the selection method: we decided to set the lexicographic parsimony pressure, since this parameter optimize both fitness and tree size, as fitness is treated as the primary objective and tree size as a secondary objective in a lexicographic ordering. This peculiarity leads to the conservation of the most influential variables over generations. The algorithm performs, hence, an implicit feature selection and among all the input variables, only the most relevant are encapsulated in the solutions.

At the end of the evolution process the population size consists of 500 members (population size), whereas a single model should be extracted at the end of the run on the learning set. It is necessary to evaluate the 500 individuals, obtained on the learning set, on the validation dataset: GP models evolve on the training set and finally the best ones are selected among all the evaluations on the validation set. These models are hence evaluated on the test set, in order to measure their generalization ability on unseen data. Parameters are summarized in Table 3.

Table 3. Parameters used to perform GP.

Parameter	**Description**
Maximum number of generations	40
Population size	500
Selection Method	Lexicographic Parsimony Pressure
Elitism	Keepbest
Initialization Method	Ramped half and half
Tournament Size	2
Subtree Crossover Rate	0.8
Subtree Mutation Rate	0.1
Subtree Shrinkmutation Rate	0.05
Subtree Swapmutation Rate	0.05

2.4 Application of ML techniques: Linear Model, k-Nearest Neighbour, Neural Network and Random Forest

We compared GP performance with other classical black-box ML approaches used for regression tasks (Hastie et al., 2009). Differently from GP, these methods do not carry out an automatic feature selection. By the end of the learning process for each run, the final population already consists of one model. This model was built on the examples given in the learning set and it only needs to be evaluated on the test set, that is a validation set is not necessary. The hyper-parameters were obtained with a tuning process. A tune grid was manually specified for each algorithm, that is the parameters to test were explained in a vector and the best results were then selected. Corresponding main parameters for all ML approaches are listed in Table 4. The unmentioned parameters were kept as default values, since during tuning no tangible improvements in terms of fitness were achieved.

The k-Nearest Neighbors algorithm (kNN) is an instance-based method, in which the input consists of the k closest instances (neighbors) in the feature space, and the output is the mean of the output values of k nearest neighbors. To predict the value of any new data point, the distance between the point and the k nearest ones are selected and the mean value of their output is assigned as prediction. A small value of k leads to results highly influenced by noise and a large value would be computationally expensive. We configured k equal to 15, that is the square root of the size of the learning set (Lantz, 2015). Values greater larger than 15, generated overfitting models.

Neural Networks (NN) emulate the complex functions of the brain. A NN is a simplified model of the structure of the biological neural network and consists of interconnected processing units organized in a specific topology. A set of nodes is arranged in at least three layers, including an input layer, where the data enter the system, one or more hidden layers, where the learning takes place, and an output layer, where the prediction is given. Learning occurs by changing connections weights, based on the error affecting the output. At each update, the weights of the connection between nodes are multiplied by a factor in order to prevent the weights from growing too large and the model from getting too complex. In this experimental study, we set a size of 15 hidden units that fit a single hidden layer.

Linear regression (LM) is a linear model, that basically assumes a linear relationship between the input and the single output variables. More specifically, the representation is a linear equation that combines a specific set of input values, whose solution is the expected output. As such, both the input values and the output are numeric. Learning a

linear regression model means estimating, with the available data, the values of the coefficients used in the representation.

Random forest (RF) is an ensemble learning method, which operates by constructing a multitude of decision trees (i.e. a forest) during learning phase and returning the mean prediction of the individual trees. Basically, a bootstrapping is first performed by the algorithm and a tree learns from a random sample of the training observations. The samples are drawn with replacement, which means that some samples will be used multiple times in a single tree. When dealing with a large number of features, it is hence common to reach greater bias. The choice of the optimal cut-point is, indeed, responsible for a large proportion of the variance of the induced tree. Instead of learning on bootstrap copies, it is possible to grow the trees by splitting nodes at fully randomly chosen cut-points. Extra-Tree parameter (extremely randomized trees) drops the attempt to find an optimal cut-point for each chosen variable at each node. The algorithm uses then the whole learning sample, instead of a bootstrapped dataset, and all the variables are selected at each split. Consequently, we set extraTrees as split rule and the number of variables available at each split was configured equal to the total number of features in the dataset.

Table 4. Parameters used to perform ML techniques with caret package in R.

ML technique	Parameters
kNN	k = 15
Nnet	size = 15; decay = 0.2
Lm	Intercept = TRUE
ranger	mtry = 48; splitrule = extratrees; min.node.size = 5

3. Results and Discussion

3.1. Interpretability of GP models.

The section is dedicated to the discussion among the models obtained with GP. The expressions obtained with this approach are readable and interpretable: this is a crucial characteristic, since we aim to put the breeder in the conditions of understanding the meaning of the model, to understand which aspects (i.e. the variables) of management are more relevant in the definition and the measure of the performance of the farm. It is possible to make some interesting considerations, in order to interpret the achieved models. First of all, it is possible to analyze the frequency with which

the variables are used by the 30 best models, that is those that show the best fitness (in other words, the lowest error) on the validation set and that have been evaluated on the test set.

In order to highlight key role variables, it is useful to investigate the median frequencies among the best solutions on all the runs. Parameters showing non-null median values were used in over 50% of the models, whereas negligible features respectively correspond to null median values. Results are reported in Table 5.

Namely, the most frequent variable is the number of calves born from natural inseminations (BORN_FN), followed by the number of cows (COWS), the total number of born calves (N_TOT) and the number of calves dead in the first 60 days after birth (N_ELIM). In exactly half of the models the EBV referred to ease of parturition of the heifers was used (H_PARTIND). It is straightforward possible to infer that GP models detected the majority of information in the aforementioned features. On the contrary, it should be noted that none of the final prediction models encapsulated th number of parturitions that required caesarean section for multiparae (C_CESAREAN) and the mean value of EBV referred to ease of birth of the bulls, which semen has been used on artificial inseminations (TFA_BIRTH).

The emphasis placed by GP among the listed features entails that the prediction of yearly weaned calves per cow for 2018 depends above all on the quantity of natural inseminations in the farm is accomplished. It is also proportional to the total number of newborns and calves not weaned during 2017. The result suggests that these variables could be the main features involved in this kind of task, that is the prediction of weaned calves per cows per year. It does not imply, however, that the other parameters are not important in the management of the farm.

We thereafter investigated the interpretability of the expressions, considering the fitness obtained in each of the best final models, and taking into account also the number of variables involved in the formula. Considering the results reported in Table 6, we can deduce that the model entailing the best predictions on the test set encapsulates only three variables (*Model 13* in Table 6). The expression in infix notation is:

$$Y = \frac{X_{10} + \frac{X_2}{X_{45}}}{X_2 + \frac{X_{45}}{X_2} + \frac{X_{10}}{X_2} + \frac{X_{45}}{X_2 + \frac{X_{10}}{\frac{X_2}{X_{45}} + X_{10}}}}, \quad (2)$$

where X_2 is the number of cows (COWS), X_{10} is the total number of deliveries occurred during the year in the farm (N_PAR) and X_{45} is the total number of born calves (N_TOT). Since these quantities are always positive summed and

divided in Equation (2), the denominators never reach null values. The *mydivide* operator is actually a division and the model can be reformulated as

$$Y = \left(\frac{X_2 + \frac{X_{45}}{X_2}}{X_{10} + \frac{X_2}{X_{45}}} + \frac{\frac{X_{10}}{X_2}}{X_{10} + \frac{X_2}{X_{45}}} + \frac{\frac{X_{45}(X_2 + X_{45} \cdot X_{10})}{X_2(X_2 + X_{45} \cdot X_{10}) + X_{10} \cdot X_{45}}}{X_{10} + \frac{X_2}{X_{45}}} \right)^{-1}. \quad (3)$$

Table 5. Median frequencies and percentage of use of each variable among the best 30 individuals found by GP.

	Variable	*Median*	*% of use on 30 runs*		*Variable*	*Median*	*% of use on 30 runs*
X_1	*BOVINES*	0	27%	X_{25}	*N_TFN*	0	17%
X_2	*COWS*	1	57%	X_{26}	*TFN_ BIRTH*	0	13%
X_3	*HEIFERS*	0	7%	X_{27}	*TFN_PAR*	0	20%
X_4	*F_CALVES*	0	3%	X_{28}	*C_GRAVID*	0	3%
X_5	*BULLS*	0	17%	X_{29}	*C_INS*	0	10%
X_6	*M_CALVES*	0	13%	X_{30}	*C_POSTPARTUM*	0	20%
X_7	*PERCENT_FA*	0	23%	X_{31}	*C_EMPTY*	0	17%
X_8	*C_AGE*	0	10%	X_{32}	*LSU*	0	7%
X_9	*C_PAR*	0	7%	X_{33}	*LSU1*	0	20%
X_{10}	*N_PAR*	0	43%	X_{34}	*LSU06*	0	7%
X_{11}	*SALXGRAV*	0	13%	X_{35}	*LSU04*	0	23%
X_{12}	*N_CORRECT*	0	33%	X_{36}	*INTP*	0	13%
X_{13}	*H_EASE*	0	10%	X_{37}	*CONSANG_NEW*	0	27%
X_{14}	*H_DIFFICULT*	0	7%	X_{38}	*N_CONSANG_NEW*	0	17%
X_{15}	*H_CESAREAN*	0	7%	X_{39}	*BIRTHW_M*	0	7%
X_{16}	*C_EASE*	0	33%	X_{40}	*BIRTHW_M*	0	27%
X_{17}	*C_DIFFICULT*	0	7%	X_{41}	*MORT*	0	17%
X_{18}	*C_CESAREAN*	0	0%	X_{42}	*ABORT*	0	7%
X_{19}	*C_N_IND*	0	40%	X_{43}	*N_ABORT*	0	10%
X_{20}	*C_PARTIND*	0	40%	X_{44}	*N_ELIM*	1	57%
X_{21}	*H_PARTIND*	0,5	50%	X_{45}	*N_TOT*	1	57%
X_{22}	*N_TFA*	0	30%	X_{46}	*N_BALIVE*	0	20%
X_{23}	*TFA_BIRTH*	0	0%	X_{47}	*BORN_FA*	0	17%
X_{24}	*TFA_PAR*	0	17%	X_{48}	*BORN_FN*	1	60%

Table 6. Fitness on the test set, number of involved variables and corresponding percentage are reported for each model evolved by GP in each one of the 30 performed runs.

Model	***Fitness on Test***	***N. of variables***	***% of variables***	***Model***	***Fitness on Test***	***N. of variables***	***% of variables***
model 1	0,1274	9	19%	*model* 16	0,1946	18	38%
model 2	0,1361	7	15%	*model* 17	0,1097	10	21%
model 3	0,1480	9	19%	*model* 18	0,1238	8	17%
model 4	0,0999	13	27%	*model* 19	0,1373	6	13%
model 5	0,1262	9	19%	*model* 20	0,1263	3	6%
model 6	0,1263	7	15%	*model* 21	0,1404	9	19%
model 7	0,1088	6	13%	*model* 22	0,1242	4	8%
model 8	0,1309	11	23%	*model* 23	0,1130	8	17%
model 9	0,1330	8	17%	*model* 24	0,1390	7	15%
model 10	0,1617	12	25%	*model* 25	0,1385	10	21%
model 11	0,1325	10	21%	*model* 26	0,1391	6	13%
model 12	0,1370	12	25%	*model* 27	0,1177	5	10%
model 13	0,0974	3	6%	*model* 28	0,1222	13	27%
model 14	0,1025	7	15%	*model* 29	0,1075	10	21%
model 15	0,1328	20	42%	*model* 30	0,1502	10	21%

In Equation (3) it is possible to notice that the simplification led to an expression containing a sum of three terms. Whenever such result is reached, the following considerations can be developed:

a) the obtained expression is in the form of

$$y = (x_1 + x_2 + \cdots + x_n)^{-1}$$

where *y is* the result (i.e. the prediction) obtained for the values x_i, *i=1,…,n* of input variables, that is equivalent to

$$\frac{1}{y} = x_1 + x_2 + \cdots + x_n.$$

b)

c) By multiplying each x_i on the right side in the previous expression, we complete the standardization process and reach the final expression

$$1 = (y \cdot x_1 + y \cdot x_2 + \cdots + y \cdot x_n)$$

or equivalently in a more compact expression

$$1 = (\tilde{x}_1 + \tilde{x}_2 + \cdots + \tilde{x}_n).$$

d) The previous standardization process allows an analysis of the contribution of each component of the linear combination. The boxplots of each component for *i=1,…,n* give a visual idea of the distribution of data in the interval [0;1] and statistical tests highlight any difference between them and with respect to the range boundaries.

We standardized hence Equation (3), in order to evaluate the contribution of each of the three components isolated in the expression. Following the previous step and renaming the predictions *Y* obtained for all the instances *j=1,…,304* with the new label n_j, and with $v_{i,j}$ the three components in parenthesis (*i=1,2,3*), Equation (3) can be expressed as

$$1 = n_j v_{1,j} + n_j v_{2,j} + n_j v_{3,j}, \quad (4)$$

or equivalently

$$1 = \tilde{v}_{1,j} + \tilde{v}_{2,j} + \tilde{v}_{3,j}. \quad (5)$$

whether referring to the rescaled values $n_j \cdot v_{i,j}$ as $\tilde{v}_{i,j}$.

Since the distributions of $\tilde{v}_{i,j}$ are not normal (Lilliefors test: $p<0.05$), the statistical significance was checked with the non-parametric Wilcoxon test with Bonferroni correction ($\alpha=0.017$) for paired data: all components are significantly different ($p < 0.001$) , that is the difference of the mean values is not zero, in particular comparing $\tilde{v}_{2,j}$ and $\tilde{v}_{3,j}$. The boxplots for each one of them (Figure 2) show that the predictions obtained with Equation (5) are mainly due to the first addend, that is most of the information is contained in $\tilde{v}_{1,j}$. Stated otherwise, in Equation (3) the corresponding value

$$\frac{X_2+\frac{X_{45}}{X_2}}{X_{10}+\frac{X_2}{X_{45}}} \quad (6)$$

is the part of the individual almost completely concurring in the prediction. The remaining components play a minor role, with a minimal effect on the performance of the individual obtained, corresponding to a refinement of the value gained with the main Component (6).

In order to further investigate the mentioned concept and the interpretability of GP models, we focused on a second individual, namely Model 20 in Table 6. The model encapsulated 3 variables, showing a larger error. Despite this, the model gains a great interpretability, since the expression released at the end of the run is given by

$$Y = \frac{X_{45}}{X_2+X_{44}}, \quad (7)$$

where X_{44} is the number of calves that did not survive during the weaning period.

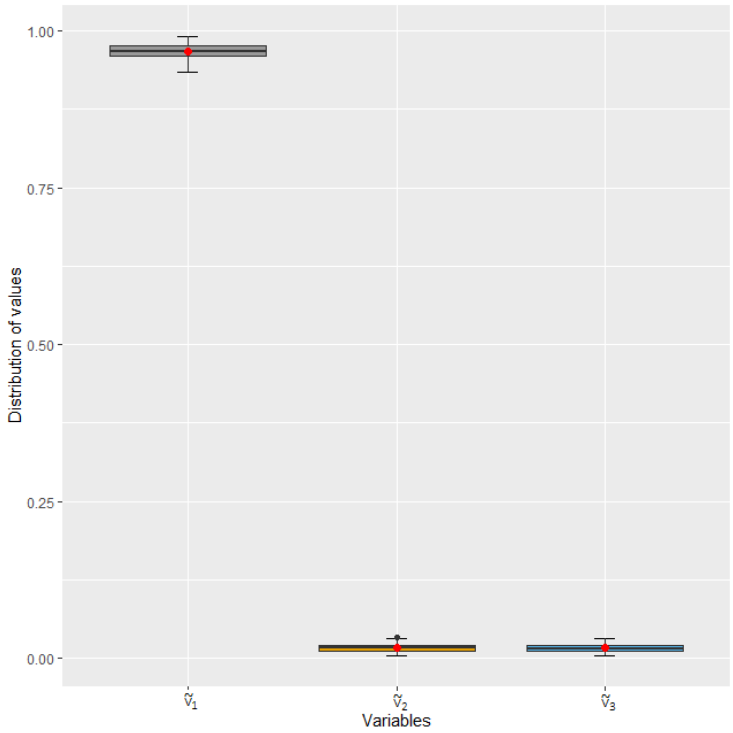

Figure 2. Boxplots of the distributions of the variables in Equation (5). Wilcoxon test with Bonferroni correction at α=0.017 reported significantly difference between the median of the three distributions (p<0.001). The single sample Wilcoxon test, with α=0.05, finally showed for each distribution mean values $\mu \neq 1$ and $\mu \neq 0$ (p<0.001). Mean values are respectively μ_1=0.9671, μ_2=0.0166 and μ_3=0.0163 (red dots).

Because of the same reasons entailing the simplification of Equation (2) into (3), the previous expression leads to the following:

$$Y = \left(\frac{1}{\frac{X_{45}}{X_2}} + \frac{X_{44}}{X_{45}} \right)^{-1}, \quad (8)$$

otherwise stated as

$$Y = (Calves^{-1} + DeadCalves)^{-1}, \quad (9)$$

where *Calves* is the yearly number of calves per cow and the number of calves per cow that do not survive during weaning period is labelled as *DeadCalves.*

As in the previous case, we investigated how the prediction is distributed between the two variables *Calves* and *DeadCalves*. We performed again the standardization procedure, supporting the analysis with an equivalent expression of (8):

$$1 = \tilde{u}_{1,j} + \tilde{u}_{2,j}, \quad (10)$$

where, for $k=1,2$, $\tilde{u}_{k,j}$ are the rescaled quantities $\tilde{u}_{k,j} = m_j \cdot u_{k,j}$, the prediction Y obtained with Model (7) are renamed as m_j, the variables $u_{k,j}$ are respectively $Calves^{-1}$ and *DeadCalves*.

Performing once again the non-parametric single sample Wilcoxon test, we obtained extremely significant p-values, supporting that the two components *Calves* and *DeadCalves* mean values are different respectively from the range boundaries 0 and 1. Both variables are crucial in predicting the output, with more relevance given by *Calves* (Figure 3). As we could entail for the first inspected model, the first component of the Expression (8) is crucial one in predicting the output, since it assumes values close to the result. However, this second model is also interesting, as the two plotted distributions assume the same complementary behavior. Boxplots in Figure 3 visually express the concept and in particular we inspected the corresponding instance. The lower outliers of the first distribution correspond to farms where cows produce a smaller number of calves. It seems reasonable that a higher portion of calves will not even survive during the weaning period, values corresponding hence to the upper outlier of the second distribution.

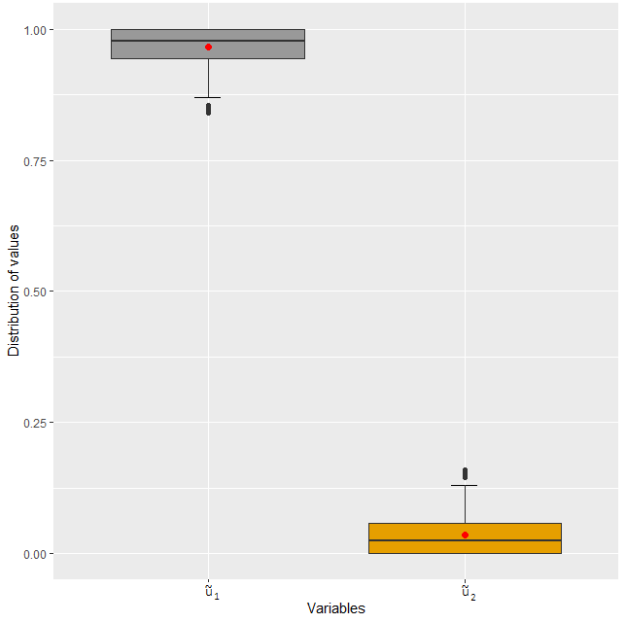

Figure 3. Boxplots of the distributions of the variables in Equation (10). The single sample Wilcoxon test, with α=0.05, showed for each variable a mean values μ≠1 and μ≠0 (p<0.001).

3.2. Comparison with other ML techniques.

In this section, the performance achieved with the five methods are compared**.** The 30 models obtained by each ML technique were first evaluated on the Test Set to measure capacity of generalization of each, analyzing the median Fitness. Finally, the best model (i.e. the one that presents the best fitness) was extracted for each technique.

We analyzed the fitness distribution among the thirty models, to assess the ability of the models to learn and to generalize. In particular, we first commented the results obtained on the learning set and thereafter on the test set. Figure 4 displays the boxplots of the fitness distribution for each technique.

For all statistical tests the significance level was set at α = 0.05. The normality of the distributions among all sets was analyzed and Lilliefors test showed a deviation from the normal distribution for the results of the LM method ($p = 0.006$). Therefore, in order to compare the performance of the achieved models on the learning sets, a non-parametric test was performed, to asses wheter there is a significant difference between the samples' performace medians. The median values were compared with Kruskal-wallis test and the null hypothesis that all median values are equal was rejected ($p < 0.001$). Indeed, all the distributions resulted significantly different to Wilcoxon signed-rank test with Bonferroni correction ($\alpha = 0.005$, since there are 10 comparisons), meaning that all performances differ one from the other on the learning set (p-values for all considered couples showed $p < 0.001$). Among all the involved method, as

straightforward from Figure 4, the models obtained with RF are indeed the best performing ones in the learning phase, whereas GP produced less accurate models.

The results on the test set were therefore investigated. Predicted values were plotted against the observed data to check their dispersion among the 30 test sets (Figure 5 (a)). In a supervised learning issue, a predictive model is more accurate as the predicted values are close to the observed ones. In order for the model to be very accurate, the regression line of the scatterplot should tend to overlap the bisector of the plane. For each technique we hence plotted the regression line of all the predicted values versus the observed values on the test sets and compared the coefficients of the line: intercepts and slopes are reported in Figure 5(a). All the techniques overestimated target values smaller than ~0.85 (i.e. the coordinates value of the intersection between the bisector and the regression lines). For values larger than ~0.85, the models underestimated the target. Indicating with $\bar{x}$ the abscissa of the intersection, the observed values $x < \bar{x}$ were estimated with greater prediction values. On the contrary, for $x > \bar{x}$ the predicted values are lower than the observed data. The slope of the fitting line obtained with LM is the closest to 1 (β_1=0.613): the predictions follow a linear distribution on each test set by construction and therefore the assumed value is expected. Among the other techniques, NNET, GP and kNN, reported slopes β_1=0.417, β_1=0.391 and β_1=0.248 respectively. Finally RF showed a slope equal to 0.002 and a corresponding larger value for the intercept (β_0=0.856).

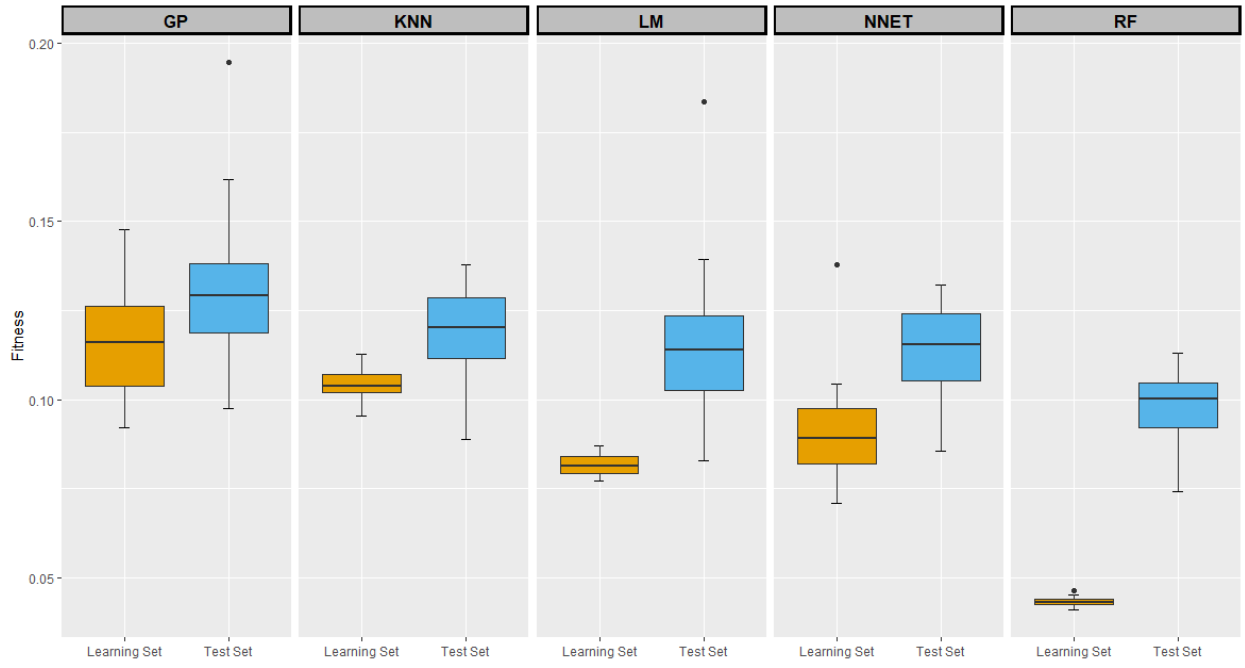

Figure 4. Fitness distribution for all the applied method. Respectively for each technique, the fitness among Learning (in yellow) and Test (in blue) sets are shown in boxplots.

Although the latter showed a lower median RMSE compared to the other techniques, two almost symmetrical regions with respect to the bisector were identifyed, entailig that predictions vary into a fixed interval (0.63;1), also for values outside the previous range. It is clear that the models are not able to generalize. Regarding the corresponding achieved errors, all fitness samples showed normal distributions of the variables (conclusion supported also by the representation of q-q plot in Figure 5 (b)), and parametric tests were performed. Since the Levène test did not show any difference between the variance of the distributions ($p = 0.139$), we carried out the one-way ANOVA test: the result was extremely significant, entailing that at least one sample had a mean performace different from the others. Finally, the Tukey test with Bonferroni correction was performed, in order to highlight which samples' average performances are actually significantly different. As it is tangible from the previous boxplots, similar results were achieved on the Test set ($p > 0.005$), that is the techniques that showed a lower median fitness on the learning set, revealed a lower median error also on the test set, compared to the other techniques. Moreover, the following pairs of techniques showed not significantly different fitness distributions among the 30 runs: GP-kNN, kNN-NNet and kNN-LM, NNet-LM, stating that the pairs of considered methods performed likewise among the test. What is clear is that, once again, models obtained with RF are the best performing models also on the test, with respect to all other techniques. It is a crucial step to assess the robustness of the model over unseen data with respect to its ability to generalize.

On this porpouse, we finally compared the fitness within each technique among learning and test sets respectively. Apart from LM results, analyzed with non-parametric Kruskal-Wallis and Wilcoxon signed-rank tests, all couples of results for each technique were tested with the Student's t-test. All techniques showed significant difference between learning and test results, extreamely remarkable among kNN, NNet, LM and RF ($p < 0.001$). Regarding the results achieved with GP, high significance was detected comparing the learning and test results ($p = 0.006$).

The statistical tests entail that all the models can achive good results on unseen instances, in particular Random Forest algorithm, since it outperformed all other techniques on both learning and test sets. It is followed by LM, NNet, kNN, with similar results as stated in the previous paragraph, and finally GP. However, by analyzing the results on the test sets, their ability to generalize tends not to be as accurate as that obtained during the learning phase. In fact, only the application of RF led to significantly better results. It must be considered that, among all methods, GP is the technique that actually produced models that show a median error on the test set not too different from the one obtained on the learning set. The final models built with kNN, NNet, LM and RF involve all variables available in the datasets (in the study under consideration, the dataset contains 48 variables). The techniques can easily perform better

and show a better fitness (Figure 4), since the predictions receive the contributions of all parameters. Feature selection is usually carried out previously in ML approaches. However, this is a not intrinsic operation in the indicated algorithms, unrelated to their structure. Considering the best model obtained with GP, i.e. showing the lowest RMSE, analyzed in the previous section (Model 13 in Table 6, i.e. Expression (3)), its performance is comparable to the median behavior obtained with RF models, even incorporating only three variables among the 48 in input, without imposing a priori hypotheses. We also managed to provide a zootechnical interpretation, which would not be possible with black-box techniques. This fact outline that the different architecture of the evolutionary algorithm can be a good alternative, balancing overfitting issues, whereas other techniques could slightly be affected. The characteristics of GP outline models that combine few variables, leading to a great interpretability of the formula and allowing further speculations on influencial parameters.

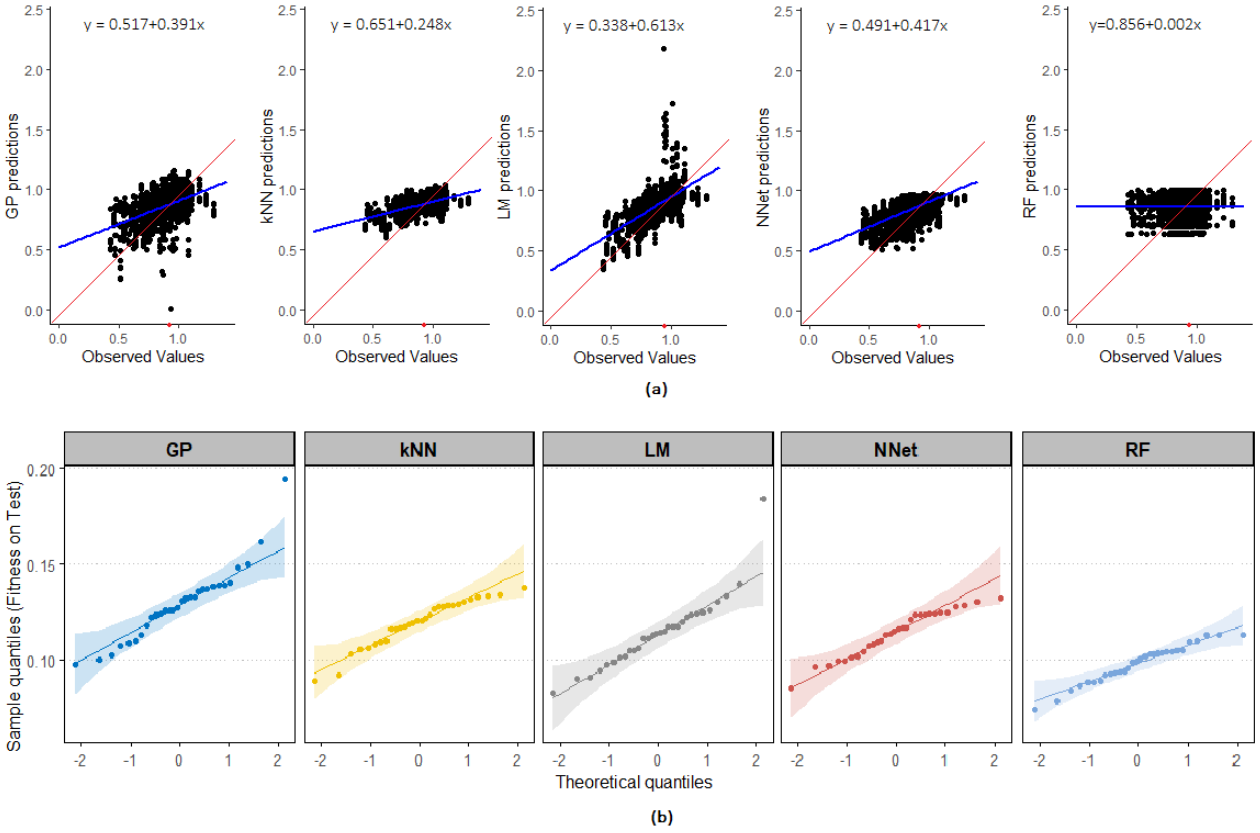

Figure 5. (a)Scatterplots for predictions among the test sets. Predicted values among test set are plotted against the corresponding observed data, for each method on all the 30 test sets. The blue line represents the linear regression fitting line, whereas the red line is the bisector. Corresponding slopes and intercepts are reported for each plot, as well as the corresponding x coordinate (red dot on the abscissa axis).

(b)Q-Q plots for the fitness among the test set. Normality of the of the RMSEs obtained is visually inspected. The quantiles obtained with all the performed techniques of the fitness on the test versus the theoretical ones are plotted. The joint distribution in each case follows the diagonal and

4. Conclusions

In this study, we investigated the performance of Piemontese cattle breedings, namely the number of weaned calves per cow produced per year. The sought prediction model should include relevant factors that describe the weaning period, that is the 60 days after the birth. Many calves do not survive during this time lapse, entailing great losses to the economic revenues of the breedings and affecting the performance. The expression was expected to predict the value without imposing any kind of a priori assumption on its formulation, but it is shaped on the available data.

Medium to large farms located in Piedmont were considered. The dataset provided by the National Association of Piemontese Cattle Breeders was accurately filtered, imposing some conditions: since the number of involved variables was much greater (we processed 19 variable in the previous study), we extracted records from biennium 2017-2018, among the most representative farms, i.e. with solid data during all the time lapse between 2014 and 2019. The final dataset consisted in 304 farms and 48 variables, referring to information on cows and artificial inseminations, as well as heifers, natural inseminations and levels of consanguinity of calves resulting from ongoing pregnancies.

ML techniques can provide prediction models without making any kind of a priori assumptions. On this purpose, the dataset was divided into learning and test sets, and a GP approach was proposed. The technique is a white-box method, suitable to provide performing models, while automatically selecting significant features. This characteristic was quite useful and let us develop considerations among the achieved expressions. Whenever a GP model can be expressed as a sum of terms, it is possible to perform an analysis among the standardized equation. We could deduce that the first term of the considered sum is the most important one, assuming values close to the output, whereas the other components concurred minimally in the prediction. We also recall that one of the aims of the study was to produce models that can be easily read by farmers, highlighting possible important factors, that can explain directly the measure or the performance of the farm. GP models detected the majority of information in five features, outlining their possibly crucial role in the prediction of the performance of the breeding farm. The number of calves born from natural inseminations is the most significant variables, followed by the number of cows, the total number of born calves, and the number of calves dead in the first 60 days after birth. In exactly half of the models the EBV referred to facility of parturition of the heifers was used.

Comparisons with other classic methods, such as k-Nearest Neighbor, Neural Network, Linear Regression, and Random Forest were developed. Compared to other techniques, GP is not the best performing method, considering the median RMSE among 30 runs. On the contrary RF produces models with the best fitness on the test. This could be mainly due to the different architecture of the algorithms and the fact that, differently from GP, the other performed methods encapsulate all the features into the prediction models. On one side, we handle with classic techniques, producing on average outperforming models, showing lower fitness but complex expressions. On the other side, evolutionary algorithms such as GP led to less accurate models, since their error is slightly greater, but easy to read and interpret. GP can model straightforward expressions, which combine a few variables, selected during the evolution process. At the end of the procedure, the best models performed as well as those obtained with other commonly used techniques, that are however characterized by non-dynamic algorithms as evolutionary ones.

In conclusion, considering all the results in relation to the kind of dealt task, we could assert that GP could represent the most suitable technique. Evolutionary algorithms can be applied on zootechnical data, achieving performing models, able to learn on the available data. Further investigations are encouraged, in order to explore the role of other variables in predicting the considered output. In this sector it is common to associate cow-calf problems to genetic and pathological factors, related to pregnancy and childbirth. However, many are the factors usually considered as marginal: difficult to detect and assert as critical points, quality of water and air, illumination, available space and surface, composition of the food ration could influence the weaning period, being key information that lay into the environment of the farms. Furthermore, comparisons on other time frames are requested. The management of the farm and the choices made by the farmer drag on over time and have delayed effects. It is necessary to analyze the problem, taking into account the data on several years as learning set, to investigate whether ML techniques could detect crucial factors, that did not emerge in this study.

Author Contributions: All authors have read and agree to the published version of the manuscript.

Funding: This work was partially supported by FCT through funding of LASIGE Research Unit (UIDB/00408/2020) and projects BINDER (PTDC/CCI-INF/29168/2017), GADgET (DSAIPA/DS/0022/2018), AICE (DSAIPA/DS/0113/2019) and PREDICT (PTDC/CCI-CIF/29877/2017).

Conflicts of Interest: The authors declare no conflict of interest and no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbona F., Vanneschi L., Bona M., Giacobini M., "A GP approach for precision farming", *2020 IEEE Congress on Evolutionary Computation (CEC) Proceedings*, Glasgow, Scotland, 2020.

Abeni, F.; Petrera, F.; Galli, A. A Survey of Italian Dairy Farmers' Propensity for Precision Livestock Farming Tools. Animals 2019, 9, 202. https://doi.org/10.3390/ani9050202

Abraham A., Nedjah N. and Mourelle L.D.M., 2006. Evolutionary Computation: from Genetic Algorithms to Genetic Programming. In: Nedjah N., Mourelle, L. D. M., Abraham A. (eds) Genetic Systems Programming. Studies in Computational Intelligence, vol 13. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-32498-4_1

Amrine, D. E., White, B. J., & Larson, R. L.: Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Computers and Electronics in Agriculture, 105, 9-19. (2014). https://doi.org/10.1016/j.compag.2014.04.009

Associazione Nazionale Allevatori Bovini Razza Piemontese, http://www.anaborapi.it

Berckmans, D., General introduction to precision livestock farming, Animal Frontiers, Volume 7, Issue 1, January 2017, Pages 6–11. https://doi.org/10.2527/af.2017.0102

Berckmans, D., Guarino, M., From the Editors: Precision livestock farming for the global livestock sector, Animal Frontiers, Volume 7, Issue 1, January 2017, Pages 4-5. https://doi.org/10.2527/af.2017.0101

Bhattacharya, M. (2013). Machine Learning for Bioclimatic Modelling. International Journal of Advanced Computer Science and Applications, 4(2), 1-8. http://dx.doi.org/10.14569/IJACSA.2013.040201

Bona, M., Albera, A., Bittante, G., Moretta, A., Franco, G.: L'allevamento della manza e della vacca piemontese, Supplemento al n. 44 dei Quaderni della Regione Piemonte-Agricoltura, pp. 65-129. (2005).

Bovine Diseases and Resources, available at: http://www.cfsph.iastate.edu/Species/bovine.php

Cole, J. B., Newman, S., Foertter, F., Aguilar, I., Coffey, M.,: BREEDING AND GENETICS SYMPOSIUM: Really big data: Processing and analysis of very large data sets, Journal of Animal Science, Volume 90, Issue 3, March 2012, Pages 723733. https://doi.org/10.2527/jas.2011-4584

Cozzi G., Brscic M., Gottardo (2009) Main critical factors affecting the welfare of beef cattle and veal calves raised under intensive rearing systems in Italy: a review, Italian Journal of Animal Science, 8:sup1, 67-80, https://doi.org/10.4081/ijas.2009.s1.67

Derner J.D., Hunt L., Filho K.E., Ritten J., Capper J., Han G. (2017) Livestock Production Systems. In: Briske D. (eds) Rangeland Systems. Springer Series on Environmental Management. Springer, Cham. https://doi.org/10.1007/978-3-319-46709-2_10

Domingos. S.P., A few useful things to know about machine learning. Commun. ACM 55, 10 (October 2012), 78-87. https://doi.org/10.1145/2347736.2347755

González-Recio, O., Rosa, G.J.M., Gianola D., Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits, Livestock Science (2014). https://doi.org/10.1016/j.livsci.2014.05.036

Guzhva, O., Ard, H., Herlin, A., Nilsson, M., Sturm, K., Bergsten, C.: Feasibility study for the implementation of an automatic system for the detection of social interactions in the waiting area of automatic milking stations by using a video surveillance system. Computers and Electronics in Agriculture, Volume 127, Pages 506-509, ISSN 0168-1699. (2016). https://doi.org/10.1016/j.compag.2016.07.010.

Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009, Springer, New York City, USA. https://doi.org/10.1007/978-0-387-84858-7

Hessle, A., Therkildsen, M., & Arvidsson-Segerkvist, K. (2019). Beef Production Systems with Steers of Dairy and Dairy × Beef Breeds Based on Forage and Semi-Natural Pastures. *Animals: an open access journal from MDPI*, *9*(12), 1064. https://doi.org/10.3390/ani9121064

Koza, J.R. Genetic programming as a means for programming computers by natural selection. Stat Comput 4, 87-112 (1994). https://doi.org/10.1007/BF00175355

Lantz, B., Machine Learning with R, (Second Edition), Packt Publishing. Cambridge University Press, Cambridge (2015).

Lo svezzamento del vitello Piemontese [The Weaning of the Piemontese Calf],

pp. 3-5, http://www.anaborapi.it/images/media/pdf/rivista/2012/2012-05.pdf

pp. 9 -11, http://www.anaborapi.it/images/media/pdf/rivista/2012/2012-06.pdf.

Lokhorst, C., de Mol, R.M., Kamphuis, C.: Invited review: Big Data in precision dairy farming. Animals. 13(7):15191528. (2019). https://doi.org/10.1017/S1751731118003439

Loyola-González O., "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View," in *IEEE Access*, vol. 7, pp. 154096-154113, 2019, https://doi.org/10.1109/ACCESS.2019.2949286.

Lynch E., McGee M., Earley B., Weaning management of beef calves with implications for animal health and welfare, Journal of Applied Animal Research, 47:1, 167-175, (2019) https://doi.org/10.1080/09712119.2019.1594825

Machado, G., Mendoza, M. R. & Corbellini, L. G.: What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet. Res. 46 (2015), https://doi.org/10.1186/s13567-015-0219-7

Morota, G., Ventura, R. V., Silva, F. F., Koyama, M., Fernando, S. C.: BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICU LTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. Journal of animal science, 96(4), 15401550. (2018). https://doi.org/10.1093/jas/sky014

Nasirahmadi, A. , Edwards, S.A., Sturm, B., Implementation of machine vision for detecting behaviour of cattle and pigs, Livestock Sci., 202 (2017), pp. 25-38, https://doi.org/10.1016/j.livsci.2017.05.014

Ortiz-Pelaez, A., Pfeiffer, D.U.: Use of data mining techniques to investigate disease risk classification as a proxy for compromised biosecurity of cattle herds in Wales. BMC Vet Res.;4:24. (2008). https://doi.org/10.1186/1746-6148-4-24

Poli, R., Langdon, W., McPhee, N.: A Field Guide to Genetic Programming. Lulu Enterprises, UK Ltd. (2008).
https://doi.org/10.1007/s10710-008-9073-y

Price E. O., Harris J. E., Borgwardt R. E., Sween M. L., Connor J. M., Fenceline contact of beef calves with their dams at weaning reduces the negative effects of separation on behavior and growth rate, *Journal of Animal Science*, Volume 81, Issue 1, January 2003, Pages 116–121.
https://doi.org/10.2527/2003.811116x

Relazione Tecnica e Statistiche al 31.12.2018 [Technical Reports and Statistics], Anaborapi, Carrù (IT) (2018)
Available at: http://www.anaborapi.it/images/media/pdf/stat/relazionetecnica2018.pdf

Rodero, E., González, A., Luque, M., Herrera, M., Gutiérrez-Estrada, J.C., Classification of Spanish autochthonous bovine breeds. Morphometric study using classical and heuristic techniques, Livest. Sci., 143 (2012), pp. 226-232 https://doi.org/10.1016/j.livsci.2011.09.022

Rutten CJ, Velthuis AGJ, Steeneveld W, Hogeveen H. Invited review: sensors to support health management on dairy farms. *J Dairy Sci*. 2013;96(4):1928-1952. https://doi.org//10.3168/jds.2012-6107

Savoia S., Brugiapaglia A., Pauciullo A., Di Stasio L., Schiavon S., Bittante G., Albera A., Characterization of beef production systems and their effects on carcass and meat quality traits of Piemontese young bulls, Meat Science, 153 (2019), pp. 75-85. https://doi.org/10.1016/j.meatsci.2019.03.010.

Silva, S.: GPLAB a genetic programming toolbox for Matlab, (2007). http://gplab.sourceforge.net/index.html

Sistema Piemonte – UBA, available at: http://www.sistemapiemonte.it/agricoltura/dw_rpu/glossario3.shtml

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC pharmacology*, *10*, 6. https://doi.org/10.1186/1471-2210-10-6

Tao, H., Guo, F., Tu, Y., Si, B. W., Xing, Y. C., Huang, D. J., & Diao, Q. Y. (2018). Effect of weaning age on growth performance, feed efficiency, nutrient digestibility and blood-biochemical parameters in Droughtmaster crossbred beef calves. *Asian-Australasian journal of animal sciences*, *31*(6), 864–872. https://doi.org/10.5713/ajas.17.0539

The package 'caret', available at: https://cran.r-project.org/web/packages/caret/caret.pdf

Williams, M.L., Parthalin, N.M., Brewer, P., James, W.P.J., Rose, M.T.: A novel behavioral model of the pasture based dairy cow from GPS data using data mining and machine learning techniques. J Dairy Sci., 99(3):20632075. (2016). https://doi.org/10.3168/jds.2015-10254

Yao, C., Zhu, X., & Weigel, K. A.: Semi-supervised learning for genomic prediction of novel traits with small reference populations: an application to residual feed intake in dairy cattle. Genetics, selection, evolution: GSE, 48(1), 84. (2016). https://doi.org/10.1186/s12711-016-0262-5