Machine Learning for Dairy Cattle Genomic Selection

Emiliano López Taranto
School of Computer Science
National University of Ireland Galway

Supervisor(s)
Dr. Desmond Chambers

In partial fulfillment of the requirements for the degree of
MSc in Computer Science (Data Analytics)
September 6, 2022

DECLARATION I, Emiliano López, hereby declare that this thesis, titled "Machine Learning for Dairy Cattle Genomic Selection", and the work presented in it are entirely my own except where explicitly stated otherwise in the text, and that this work has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Emiliano López

Acknowledgements

First, I would like to take this opportunity to thank my thesis supervisor, Dr. Desmond Chambers, for guiding me through the completion of this project. His suggestions were essential for conducting this research with academic rigor. I would also like to thank the Irish Cattle Breeding Federation for providing me with the data needed to conduct this study. In particular, Paschal Coughlan, a member of the HerdPlus team, assisted me in the process of acquiring the appropriate datasets. Finally, I would like to thank Mario J. Ularte, a molecular biologist and graduate student of the Master's in Business Analytics at NUI Galway. He helped me incorporate knowledge related to the field of genetics and molecular biology needed to carry out this research. Mr. Ularte's technical support over the past few months allowed me to enter an unknown territory and pursuit a fascinating objective.

Abstract

Modern agriculture is a constantly developing field that merges a combination of sciences to find better and more efficient production techniques. The motivations for this chase are plenty and highly compelling. First and foremost, agricultural producers are naturally interested in using more efficient practices to increase their profits and get a competitive edge in the international market [1]. Also, global collaboration in research looks for ways of transforming current production techniques to reduce the resources needed to achieve the same output and reach a more sustainable system through increased efficiency [1]. Genomic selection is a tool used to systematically improve the productive potential of different species through a laborious genomic analysis [2]. The dairy industry uses this technique to find the best bulls and sires to breed the most efficient herds possible [3]. The impact genomic selection induced on dairy production since its implementation in the late 2000s is undeniable [4]. Optimizing phenotypic traits like milk yield, milk composition, fertility, and physical conformation has led to an accelerated increase in the per-cow productive potential, thus reducing the environmental impact of each unit of dairy produced [1]. Predicting phenotypic traits of a single animal involves analyzing extremely long DNA sequences called single-nucleotide polymorphisms (SNPs) to identify specific markers associated with known phenotypic attributes [2]. The highly quantitative nature of this procedure opens

the door for computer science to help process this data. Furthermore, it is evident by observing the precision of current predictions that there is plenty of room for improvement [5]. This study explores the potential use of Machine Learning to enhance phenotypic predictions in dairy cattle. Four different machine learning algorithms were trained to predict a selection of thirteen phenotypic targets. The results showed that these algorithms have the capacity not only to produce more accurate predictions than the original method but also to do it more consistently throughout the different targets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Introduction

Over the last 15 years, genomic selection has massively transformed livestock farming due to breakthrough technological advances in single nucleotide polymorphisms (SNP) genomic sequencing [4]. Since then, multiple national and international-wide initiatives have been launched to analyse large amounts of cattle genomes, allowing continuous improvement in farmers' herd profitability. Dairy farmers' focus has been to enhance the phenotypic traits relating to milk production and fertility, which are the two most significant factors for the animal's profitability [3]. Ireland started a national-wide campaign to collect cattle genomic data in 2009 that strongly complemented its already existing dairy-cattle phenotype measurement program [4]. The program was (and still is) in the hands of the Irish Cattle Breeding Federation (ICBF) which had developed an economic index called EBI that signifies the expected profitability of a dairy cow per lactation compared with a base cow [3]. The introduction of genomic selection allowed to predict dairy bull's phenotypic values before there were any records of their progeny's lactations. Therefore, the information that used to take six to seven

years to get now was available only after two years after the bull's birth [2]. Nevertheless, the phenotypic predictions' reliability at this early stage is in between 50% and 55% [2, 5]. This project will attempt improve the accuracy of phenotypic predictions using machine learning. This approach will use data collected from the ICBF database at two different points in time and train four separate models for its later comparison against the traditional EBI predicting method. The motivation for this study is to find room for improvement in a predicting procedure that could instantly benefit Irish dairy farmers. Also, in the long-term, more accurate predictions would accelerate the phenotypic transformation that the EBI has started, which has led to a more sustainable dairy production nation-wise.

1.2 Background

1.2.1 Genomic Selection in Livestock Farming

In animal biology, genomic selection refers to the analysis of DNA samples to predict the breeding values of a specimen's offspring. The field of genetics dedicated to analyzing genetic variations along genetic sequences to identify their relationship with observable phenotypical traits is called Genome-Wide Association Studies (GWAS). Implementing genomic selection in breeding schemes is highly beneficial, as it offers information on bulls without lactating progeny. Otherwise, this information would take about six years to be collected, or it would have to be estimated based on parental average breeding values with much less accuracy [2]. The first cattle genomes to be sequenced worked as a reference for further studies on other ruminants in 2002 [10]. It was not until 2008 that technological advances in genomic sequencing started showing their potential and genomic selection gained relevance in breeding schemes [10]. The breakthrough was made possible by implementing the newly developed single SNP genotyping, which,

Table 1.1: Key features of implemented genomic selection programs in selected countries at April 2011 [9]

Feature	Australia	Ireland	NZ	France	Germany	Netherlands	Denmark-Sweden-Finland	USA-Canada
Year in which genomic evaluation commenced nationally	2011	2009	2008	2009	2010	2010	2008	2008
Size of reference population (males; production traits)	2247	4500	3600	19377	19377	19377	19377	12152
Reliability (total merit index) (%)	43	54	55–60	65	65	60	55–60	62
Reliability (protein yield) (%)	50	61	55–60	65	72	66	63	71
Females included in reference population	Soon (10k)	Not yet	16 000	Not yet	0	0	0	11 473
Number of young bulls genotyped per year	300	1000	1500	12–15 000	6000	2100	1800	13 070
Number of bulls progeny-tested	100	70	160	0	<500	140	175	2000
Age at which young bulls are widely used (months)	16	24	14	16	15	20	20	12
Price relative to proven bulls	Same	Less	More	Less	Same	Same	Same	Same
Number of young genomically tested bulls in the top 20 bulls ranked on country's index	11	10	20	20	17	11	12	20
Market-share of genomically tested bulls (bulls without milking daughters) (%)	n.a.	50	30–35	30	<30	25	45	43

Several methods exist for calculating the reliabilities of genomic breeding values; so in some cases, the reliabilities between countries are not directly comparable. New Zealand (NZ) included Holstein, Jersey and crossbreds from Livestock Improvement Corporation. For all other countries only Holsteins are reported

in addition to being highly reliable, was significantly cheaper than its predecessors [10]. This technology kickstarted a wave of research and development that placed genomic selection in a central position in breeding schemes globally. The founding of national and international breeding programs (e.g., EuroGenomics, IGenoP, USDA genomic evaluations, Teagasc genomic selection project) in the late 2000s shows the increased global interest in genomic selection. Table 1.1, taken from J.E. Pryce (2011) [9], shows the year when selected countries started performing genomic selection nationally. Nowadays, genomic selection plays a prominent role in cattle breeding programs, especially in Artificial Insemination (AI) predominated production systems, such as dairy farming. These programs introduce young bulls to their records between 14 and 20 months old. At that moment, the reliability (measure of EBI accuracy) of the genomic values predicted revolves around the 50% mark [2, 5]. As soon as the bull's first offspring

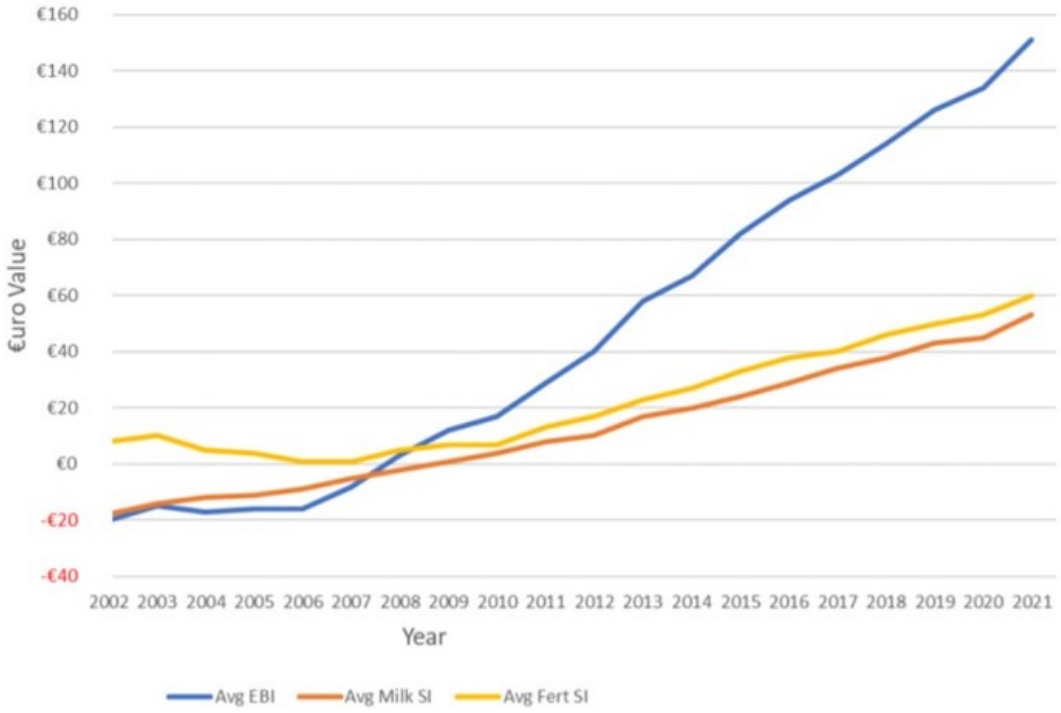
are born, calving records are added to the bull's genomic proof, incrementing the overall reliability of the genomic values predicted. Finally, as the bull's daughters become old enough to start lactating, records from these lactations are used to complement the genomic values previously predicted. These daughter-proven genomic values can reach reliabilities of over 95% [2]. The two most significant genomic traits that dairy cattle breeding programmes focus on globally are milk production and fertility. Generally, milk production contemplates the total milk yield, fat yield, and protein yield. These last two are critical because it is what producers get paid for rather than their net volume. Cattle fertility also affects dairy farmers' revenue as it comprises the cow's survival rate and calving interval (number of days between a calving and the next calving), which impacts its total expected number of lactations and lactation yield. Other traits that also are taken into consideration by breeding researchers are calving difficulty, beef production, maintenance, management, and health [3]. In contrast to its many benefits, genomic selection has occasionally rendered the excessive use of specific bloodlines, which reduces the herd's genetic diversity and ultimately causes inbreeding depression [11]. Higher rates of inbreeding index in herds are related to a decrease in milk yield (both in net and solids yield) and an increase in calving interval, directly affecting the herd's profitability. Multi-generation databases and more intensive genomic analysis (such as Genome-Wide Association) are viable tools to keep genomic diversity levels under control [11]. Despite this downside, it is safe to say that genomic selection has pushed dairy production to a sustained increase in yield per cow since 2008 [12]. The use of genomics in breeding schemes became the standard and is now inescapable for dairy farmers trying to maintain their profitability.In 2019, the Organization for Economic Co-operation and Development projected a 1% increase in milk yield and a 0.6% decrease in the dairy cow population per year between 2019 and 2029 [13]. This development has

positive implications on the overall sustainability of the dairy industry, given that the same output is possible with a reduced number of cows and consequently the environmental footprint (mostly greenhouse gases) per tonne of dairy products gets decreased [1].

1.2.2 Economic Breeding Index

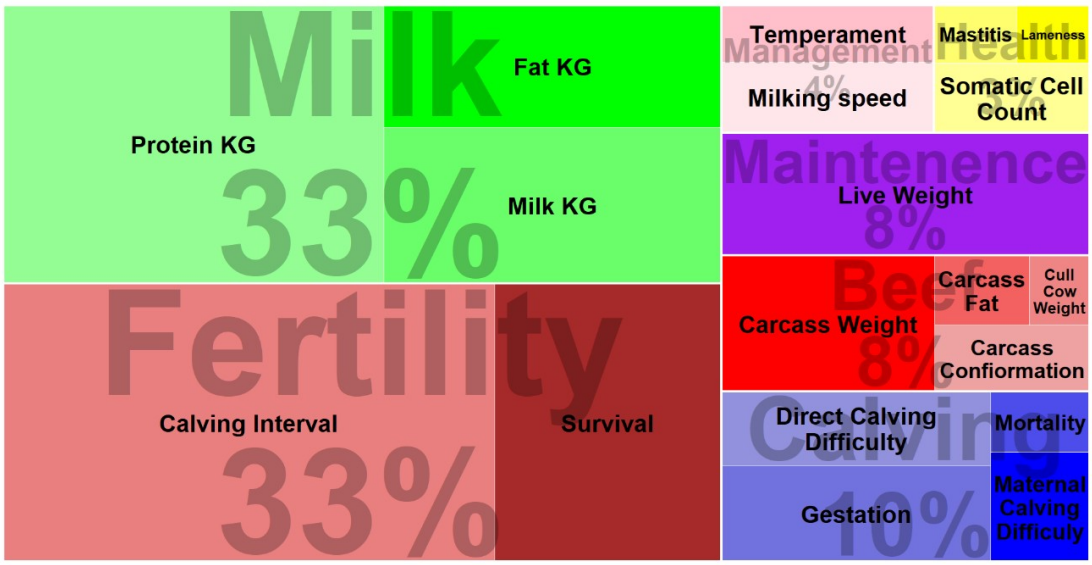
The Economic Breeding Index (EBI) is an economic measure of a bull's progeny expected profitability per lactation compared to a base dairy cow [3]. Although it is a single-digit index, it is composed of seven sub-indexes that can be analysed and targeted for improvement singularly [3]. The EBI is an initiative developed by the ICBF in 2001 [14]. This first edition contemplated only five traits in total (milk yield, fat yield, protein yield, calving interval, and survival rate) [14]. In its first five years, the EBI's complexity -and subsequent usefulness- of the EBI grew with the addition of the lifespan trait in 2004, the calving performance and beef performance traits in 2005, and the health trait in 2006 [14]. The most consequential development was the introduction of genomic selection in 2009, provoking a rapid acceleration in the increment of the average milk and fertility sub-indexes, which added to an even higher increment in the net average EBI, as shown in figure 1.1 [4].

Figure 1.1: Dairy EBI Trends 2002-2021 [4]



The EBI was first divided into sub-indexes in 2004 [14]. At that time, the sub-indexes utilized were milk production, fertility/survival, calving performance, and beef performance [14]. The present-day EBI also features a maintenance sub-index (relative to the cow's live weight), a management sub-index, and a health sub-index [3]. Each subindex has a particular weight over the total EBI and is composed of a combination of different traits, as shown in figure 1.2 [3]. The division of the EBI into sub-indexes allows dairy farmers to select specific bulls for artificial insemination to improve given features in their herd.

Figure 1.2: EBI Formula [3]



Twenty years after this initiative's launch, its impact is undeniable. The current average dairy cow is over €170 more profitable than the average dairy cow from 2001 [4], producing higher milk yield, higher percentages of milk solids, and shorter calving intervals [1]. In addition, because of the previously mentioned advances, the carbon footprint per kilogram of milk solid has decreased. Table 1.2 [1] shows the distribution of greenhouse gasses emitted by dairy herds per milk solids kg across different EBI categories. Currently, the EBI bull list contains values of genomically selected (GS) and daughter-proven (DP) sires (cow's male parent) [15]. DP sires' EBIs have an average 22% higher reliability than GS sires born in the same year. This higher reliability lies in the additional accuracy that the daughter-proven records add to the original genomic selection when predicting future performance. Regardless, the ICBF encourages producers to use GS sires, as they still show good average performance and cost around €50 less than DP sires [16]. Also, they suggest using teams of at least eight sires equally in a herd to "spread the risk" and secure an increase in the herd's EBI [16]. This practice

Table 1.2: Herd EBI and Carbon Footprint [1]

Herd EBI Category	Herd EBI	Kg CO2/Kg FPCM
Bottom 20%	€61	1.04
20-40%	€102	1.00
40-60'%	€121	0.98
60-80%	€139	0.95
Top 20%	€165	0.90
Average	€118	0.98

The summary of carbon emissions produced on dairy farms based on herd genetic merit. The carbon footprint is expressed as the number of kilograms of carbon dioxide produced per kilograms of fat and protein corrected milk (kg CO2/kg FPCM)

ultimately benefits both the producers and the breeders, as their young GS sires get performance data, which makes them DP sires, increasing their EBI reliability and sample price. As a remark, it should be mentioned that out of the ten most used dairy bulls in Ireland in 2021, six were GS sires with no progeny records [17]. This fact shows the widespread use of GS sires, despite their relative lack of individual reliability.

Chapter 2

Related Technology and Research

2.1 Overview

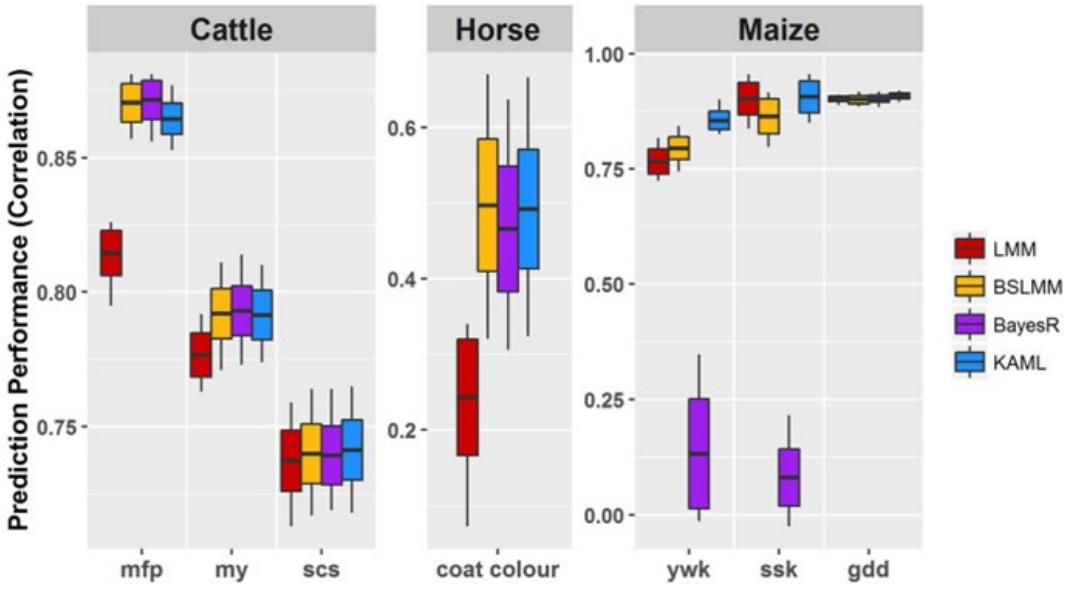
After more than a decade of extensive cattle genomic data collection, it seems natural to assume that the application of quantitative approaches to genomic prediction has been widely explored and made public. Even more so, considering that cattle genomic data comes in large dimensions, not only because of the size of populations genotyped but also because of the number of features included in each genomic sample. Nevertheless, most of the attempts to implement machine learning techniques publicly available aren't dated much earlier than 2020. One previous research worth mentioning explored the use of machine learning to detect subsets within SNPs to create genomic relationship matrices for predicting breeding values [18]. However closely related, this study is not directly comparable to the one presented in this document because the goal was not to predict breeding values but to find representative SNP subsets. More recently, one other research attempted to predict a specific phenotypic trait in dairy cattle [19] without satisfactory results. The target trait was the residual feed intake, which describes the cow's feed efficiency. The researchers attributed the low relia-

bilities of the predictions (not surpassing the 34% and being as low as 13% in one subset) to the low amount they could gather. In addition, the genetic uncorrelation of the target trait was a contributing factor to the low performance. Finally, this research concluded that its necessary to collect more data to achieve more accurate predictions. Although including a deeper review of this case would not be particularly constructive, it is worth mentioning to show that this approach to genomic predictions is still relatively undeveloped. The following review will dive deeply into the details of three directly comparable investigations. Although there are more studies published, these three are representative of the universe of approaches regarding data and algorithm selection. Conclusions taken from this analysis will later serve to design the methodology.

2.2 Improving genomic predicting accuracy using machine learning

A group of researchers from the Huazhong Agricultural University (Wuhan, China) conducted this study and later published it in early 2020 [6]. Its objective was to explore the performance of machine learning models for the genomic prediction of complex traits in different species. The inputs taken for this study were multiple datasets of SNPs collected from humans, cattle, horses, and maize. The nature of SNP data means that every event presents many features, and accurate modeling requires many events. The cattle dataset corresponded to 5024 German Holstein bulls previously genotyped and used for previous research, for which their phenotypic data were available. The conventional method for genomic prediction starting from SNPs data is the linear mixed model (LMM), and it supposes that all SNPs contribute to the heritability of traits. This study compared LMM against several Bayesian framework-based methods and a new algorithm

Figure 2.1: ML vs LMM Performance Summary [6]

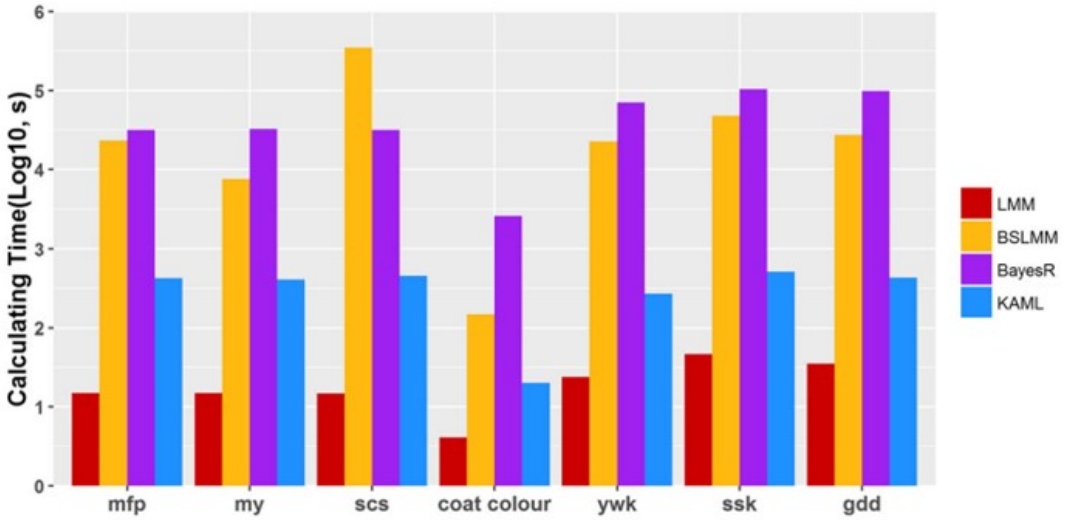


Comparison of prediction accuracy performances of LMM (red), BSLMM (yellow), BayesR (violet), and KAML (blue) in cattle, horse, and maize datasets. The prediction accuracy performance of each method was measured by the correlation method, which is the average Pearson correlation between predicted values and phenotypic values of 20 replicates in the validation subset. In each replicate, the dataset was randomly split into a reference subset containing 80% of individuals and a validation subset containing the remaining 20%. For each boxplot, the middle line represents the average value, the bottom and top are the standard deviation, and the upper and lower ends of each box represent the maximum and minimum, respectively

developed by the researchers called Kinship-adjusted-multiple-loci (KAML). Over the studies on cattle SNP data, the target features selected for prediction were milk-fat percentage, milk yield, and somatic cell count. The results of the three best-performing algorithms and the classic LMM method were quantified using the Pearson's correlation coefficient, and they are visible in figure 2.1 [6].

The three machine learning methods outperformed the classic LMM calculation in all three of the phenotypic traits targeted. It is relevant to mention that considering the studies on human SNP data in addition to the ones mentioned before, the KAML algorithm proved to be the most consistent. During these

Figure 2.2: Calculating Time Per Method [6]



The comparison of computing performances (in seconds) of LMM (red), BSLMM (yellow), BayesR (violet), and KAML (blue) for cattle, horse, and maize datasets. The y-axis represents the computing time in log10 scale

studies, the researchers had to make constant compromises to keep their calculation times under control. As mentioned before, SNP data brings a substantial dimensionality challenge, which often results in methodology decisions that affect the final accuracy achieved by the algorithms. Figure 2.2 shows the running time of every method included in the research. Although the newly developed KAML algorithm shows improvement, the LMM method still shows a clear advantage over the machine learning approaches.

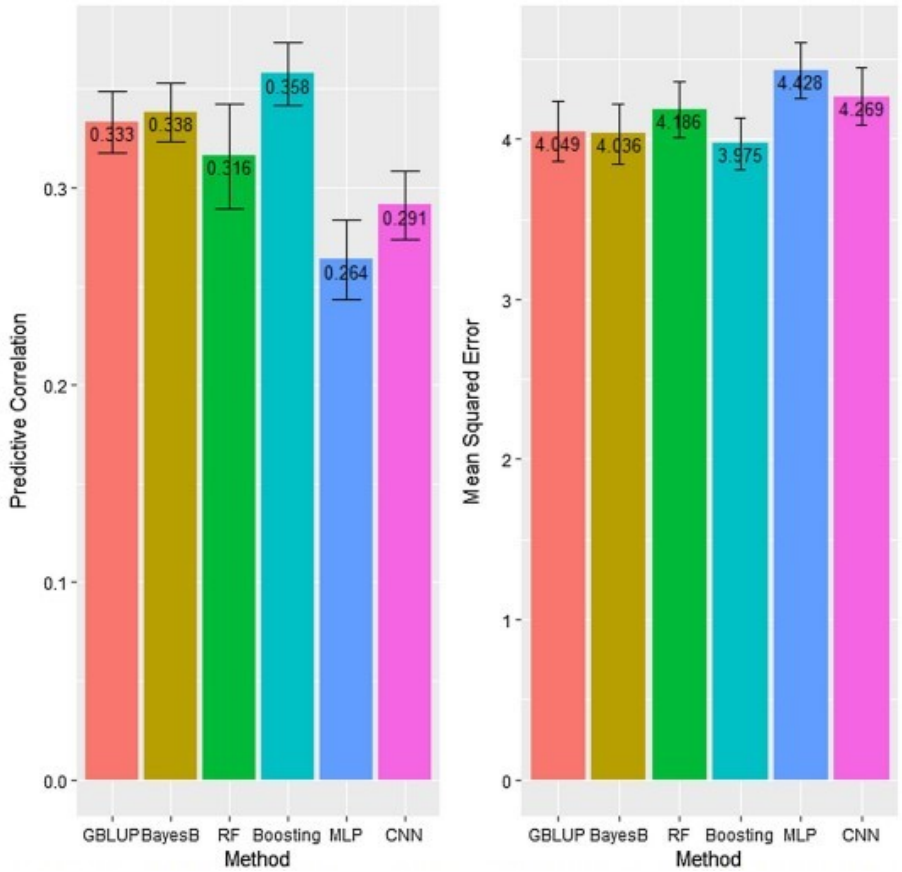
Finally, it is fair to say that this study contributes to advances in the use of machine learning for genomic prediction. In particular, the promising results across all species of the KAML algorithm suggest that the potential of machine learning is high in the field of genomics in general.

2.3 Deep learning vs parametric and ensemble methods for genomic predicitions

Similar to the previous research, this is also a genome-wide association study that attempts to predict phenotypic values from SNP data and dates back to early 2020 [7]. Nevertheless, this approach evaluates the performance of deep learning algorithms (convolutional neural networks and multi-layer perceptron) in addition to ensemble methods (random forest and gradient boosting). The researchers compared these deep learning algorithms to two classically used parametric methods: genomic best linear unbiased prediction (GBLUP) and Bayes B. The study contemplated both an additive and a non-additive genetic scenario. The first scenario studies the genetic influence of genes from a linear standpoint, while the second scenario studies the genetic impact of the relationship between genes. The mean square error (MSE) was the measure of choice to evaluate the method's performance within a five-fold cross-validation scheme.

Figure 2.3 shows a summary of the performance of each method. It is observable that the only machine learning algorithm that managed to outperform the classical methods was gradient boosting (GB), while the deep learning methods showed the worst performance overall. Nevertheless, it is noteworthy that this study also arrived at some interesting conclusions. First, the machine learning methods showed better performance in non-additive than in additive scenarios, indicating a potential use of these methods for some particular traits. Second, using larger batches of SNPs improved the performance of machine learning methods but showed no improvement in the classic parametrical methods. This is no surprise given the nature of machine learning but is worth mentioning because it indicates a potential advantage for these methods. Finally, it is noteworthy that, although the researchers used the UF Research Computing HiPerGator su-

Figure 2.3: Calculating Time Per Method [7]



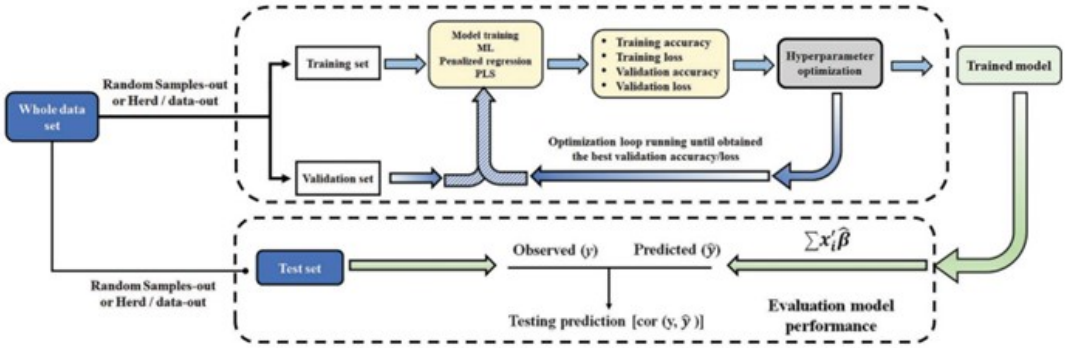
Predictive correlation (left panel) and mean squared error of prediction (right panel) of two conventional statistical methods (GBLUP and Bayes B) and four machine-learning methods including random forests (RF), gradient boosting (Boosting), multilayer perceptron (MLP) and convolutional neural network (CNN) using a real dataset of sire conception rate records from US Holstein bulls. The whiskers represent 95% confdence intervals

percomputer (https://www.rc.uf.edu) for this experiment, the data dimension partially limited the dimensions of the data.

2.4 Machine learning for predicting phenotypic traits in dairy cattle

This research [8] was carried out by a group of academics from Italian and American universities, with the help of the Breeders Federation of Trento Providence (Trento, Italy), and published in the Journal of Dairy Science of The American Dairy Science Association (ASDA) in 2021. The approach chosen involved taking Fourier-Transform Infrared Spectroscopy (FTIR) data from milk samples and comparing the predictive ability of different machine learning algorithms against classical phenotype prediction methods. FTIR is a widespread tool for analysing milk composition and phenotyping dairy cows. In a few words, this technique shines infrared light through a milk sample and analyses the wavelength range absorbed by the milk solids. Then, partial least squares (PLS) regression is used, among other less usual methods, to predict phenotypic trades. The algorithms chosen for comparison against PLS regression were random forests (RF), gradient boosting machine (GBM), and elastic nets (EN). The phenotypic traits targeted were selected for their difficulty and expensive predictability when using classical methods. These traits were the body condition score (BCS), β-hydroxybutyrate (BHB as mmol/L), and the κ-casein (κ-CN as a percentage of nitrogen). The data was collected daily from 471 Holstein-Friesian cows and split in two ways. The first one was a "herd/date-out" split, assigning cows for training (70% of herds), validation (10% of herds), and testing (20% of herds) based on the date of the sample and herd of origin. The second one was a "samples-out random" split consisting of 10-folds. Training data took over 8-folds, and validation and

Figure 2.4: Random Search [8]



Workflow of grid search for the hyperparameter optimization of supervised machine learning and penalized regression methods. The general process for the cross-validation includes splitting the whole population into training, validation, and testing subsets. The training and the validation subsets are used to select the main hyperparameters for each approach used for phenotypic prediction. The trained model with the best adjustment is then evaluated on the disjoint test subset in the cross-validation scenario. FTIR = Fourier-transform infrared spectra.

training took 1-fold each. These two scenarios served for cross-validation to subject the research to a high performance-evaluation standard. The fundamental difference between each scenario was that, in the first one, the data used for testing is independent of the one used for testing and validation because they belong to different herds. On the contrary, in the second scenario, the data is randomly split, allowing for a higher dependence.

To optimize the selection of hyperparameters, the researchers performed a random grid search on the machine learning algorithms (RF and GBM). As shown in figure 2.4 this consists of an optimization loop that iterates over the testing and validation data, calculating the loss and accuracy and adjusting the hyperparameters. Table 2.1 summarizes the results of this experiment. The machine learning methods presented a higher performance than PLS regression for the three traits in both cross-validation scenarios. The cross-validation comparison shows slightly higher performance in the samples-out random than in the herd/date-out scenario. Nevertheless, this difference is expected, given that the testing and

Table 2.1: ML vs PLS Reslults Summary [8]

		Samples-out random cross-validation				Herd/date-out random cross-validation			
Trait	Model	r, training	r, validation	RMSE validation	Slope prediction	r, training	r, validation	RMSE validation	Slope prediction
BCS	EN	0.92 (0.003)	0.59 (0.030)	0.27 (0.025)	1.22 (0.035)	0.85 (0.001)	0.55 (0.051)	0.24 (0.042)	1.28 (0.041)
	GBM	0.91 (0.002)	0.63 (0.023)	0.25 (0.017)	1.07 (0.030)	0.88 (0.002)	0.58 (0.048)	0.23 (0.042)	1.13 (0.039)
	RF	0.95 (0.001)	0.61 (0.028)	0.26 (0.038)	1.29 (0.038)	0.82 (0.007)	0.58 (0.051)	0.24 (0.048)	1.30 (0.045)
	PLS	0.95 (0.001)	0.57 (0.034)	0.35 (0.036)	0.89 (0.047)	0.89 (0.004)	0.53 (0.057)	0.25 (0.053)	0.79 (0.095)
BHB (mmol/L)	EN	0.89 (0.004)	0.78 (0.023)	0.10 (0.012)	1.22 (0.034)	0.86 (0.003)	0.70 (0.039)	0.12 (0.019)	1.29 (0.042)
	GBM	0.90 (0.001)	0.80 (0.023)	0.09 (0.009)	1.03 (0.026)	0.90 (0.001)	0.73 (0.030)	0.11 (0.016)	1.14 (0.033)
	RF	0.90 (0.001)	0.79 (0.027)	0.10 (0.011)	1.30 (0.034)	0.85 (0.003)	0.73 (0.040)	0.12 (0.017)	1.44 (0.037)
	PLS	0.88 (0.002)	0.76 (0.030)	0.10 (0.012)	0.89 (0.037)	0.92 (0.002)	0.68 (0.048)	0.12 (0.015)	0.92 (0.061)
κ-CN (% N)	EN	0.96 (0.001)	0.79 (0.027)	1.25 (0.049)	0.92 (0.035)	0.86 (0.004)	0.74 (0.030)	1.29 (0.067)	1.20 (0.061)
	GBM	0.97 (0.001)	0.81 (0.025)	1.08 (0.046)	1.06 (0.034)	0.88 (0.002)	0.77 (0.029)	1.14 (0.048)	1.13 (0.047)
	RF	0.96 (0.001)	0.80 (0.030)	1.18 (0.052)	1.21 (0.037)	0.90 (0.002)	0.77 (0.031)	1.17 (0.054)	1.25 (0.062)
	PLS	0.90 (0.008)	0.77 (0.034)	1.41 (0.062)	1.37 (0.041)	0.90 (0.003)	0.73 (0.035)	1.26 (0.073)	0.90 (0.067

Predictive ability (r), root mean square error (RMSE), and slope prediction (SD in parentheses, refers to variation between replications used in the cross-validation scenarios) of the standard model (PLS) and machine learning methods for BCS, BHB, and κ-CN using milk spectra data

training data are independent in the second scenario, which ultimately improves the overall reliability of this study. In conclusion, this research provides a successful example of the application of machine learning for phenome prediction in dairy cattle. This instance shows the potential that data analytics can reach applied to genomic selection. Additionally, the superior performance of GBM and RF suggests that ensemble methods might be particularly convenient in this field and should be paid attention to in further studies.

2.5 Conclusion

The room for improvement in phenotypic prediction reliability is a global motivating factor for research, and the data-abundant nature of genomics gives machine learning the opportunity to contribute to these studies. On the positive side, the results reviewed show that there is at least a small margin to gain from implementing machine learning. In particular, ensemble methods appear to outstand from their competition when used in cattle data [7, 8]. Nevertheless, there are

still some challenges. First, using SNP data as inputs can potentially allow for higher accuracy, but it comes at a high computational cost [6]. Also, the model's predictive ability will always depend on the genetic correlation of the target trait, which means that hard-to-predict targets will continue to show low reliabilities [19]. In conclusion, the study conducted in this thesis must consider a wide selection of methods, target traits, and features. Also, the novelty of this application requires a highly critical approach to its validation.

Chapter 3

Proposed Research and Methodology

3.1 Overview

As mentioned before, the objective of these experiments is to evaluate the performance of a collection of data analytics algorithms when attempting to predict breeding values of genomically selected dairy bulls. The data used for this study (collected and published by the ICBF) describes the bull's progeny phenotypical values within the scheme of the EBI index. The targets selected for prediction will include an array of EBI sub-indexes, specific phenotypic traits (e.g., milk-fat percentage), and the overall EBI value.

3.2 Data description

The ICBF has made public, over the last number of years, datasets containing all the dairy and beef bulls available for artificial insemination and their projected

breeding values [20]. In addition, every bull can be searched for individually in the ICBF database using its identification code. Breeders and farmers genotyped the bulls included in this list using the IDB SNP chip [21]. This chip has been the genotyping instrument used for nearly all cattle genotyping in Ireland. This tool projects phenotypic traits of each sire, which then go into the database. Additionally, the values presented in regard to the sires' calving characteristics derive from data collected from each animal's progeny calving, and records from DP sires' progeny's lactations (among other traits) complement their values [3]. In conclusion, these schema results in large datasets showing a mixture of genomically predicted values and daughter-proven data, with different levels of reliability. Compared to the previous genome-wide association studies [6, 7], the data used for this study is radically different because this data is not genotypical data. Instead, these are phenotypical predictions based on genome sequencing. On the one hand, this detail affects the input data's organic nature, but on the other hand, its dimensionality as well. The bull's EBI dataset [22] the dataset bears 100 features for each sire, organized in the following structure:

1. **"Bull Details":** Includes general identification and heritage data such as the bull's ID, birth year, and main breed.

2. **"EBI & Sub-Indexes":** Includes the economic value and reliability for every sub-index. The EBI, also presented with its reliability, is the sum of all the sub-indexes.

3. **Sub-Index breakdown:** The dataset shows every specific trait affecting every sub-index, some of them with their correspondent reliability. The number of features corresponding to each sub-index varies between thirteen and two, depending on the sub-index.

4. **Additional traits:** This section shows several phenotypic characteristics

regarding health and physical conformation.

5. **Kinship information:** The final section identifies the bull's sire, dam, and maternal grand-sire.

The proposed study requires data collected at two different points in time. The earlier data will contain the predicted values based on genetic selection, while the latest data will provide the targets based on daughter-proven records. Given that bulls take between 14 to 20 months to get genotyped (thus entering the dataset) and between 6 and 7 years to get daughter-proven, the ideal chronological separation between the two datasets used in this study would be 5 years. This gap would give just enough time for most of the youngest GS sires of around two years of age in moment t, to be seven years old and have daughter-proven records at the moment t+5. A wider gap would assure that all the GS sires in moment t would be DP sires in moment t+k, but at the cost of a higher dissociation between the animals' physical state in each moment. According to this logic, the optimal records to select for this research would be the 2022 and 2017 [23] bull EBI lists. After obtaining these datasets, a complex process of filtering and rearranging is necessary to reduce the universe of data to a two-dimensional dataset, containing nothing but the records relevant to the study. This task is made possible with the help of a relational database engine. The two datasets must be loaded to the relational database engine, selecting the bull identification code as their primary key. The final dataset will be the outcome of a query that will select all the features from the 2017 dataset, all the target values from the 2022 dataset, filtering by the 2017 proof type (to select only sires that were GS at that moment), and organizing them by the bull ID. Finally, an intense data cleaning process will be necessary to deal with missing values, outliers, data misspelling, and several inconvenient data type cases.

3.3 Target Selection

The array of targets selected for this research will be broad and diverse, attempting to explore every possibility this approach might have to show improvements over the classic prediction methods. Three different types of targets compose this list:

1. **Index:** The first type refers solely to the EBI. This value is a relevant feature because of its power to summarize the animal's profitability and because it is the most used measure to rank the bulls.

2. **Sub-Indexes:** All seven sub-indexes are included in the target selection. The sub-indexes provide more utility to the farmers, allowing them to choose the sire that will enhance their herds' performance in the way they particularly need. This option makes the sub-indexes highly valuable and consequently a subject of prediction in this research.

3. **Traits:** The number of traits contemplated in the dataset is overly extensive to be entirely included in the target selection. Nevertheless, accurately predicting a highly relevant trait subset would increase the reach of this study without creating a labour overflow issue. The selected targets are milk yield (kg), fat yield (kg), protein yield (kg), calving interval (days), and survival percentage. The criteria for this selection were the trait's impact on the cow's profitability. This criterion assures that the selected parameters are worthy of the farmers' attention.

In summary, this research will have thirteen target features, each specifically selected to give a potential interested party important information about the sires' phenotypical characteristics.

3.4 Algorithm Selection

3.4.1 Introduction

This study intends to give a broad and complete view of the application of data analytics for phenotypic prediction. That is why the algorithm selection contains an extensive and diverse set of methods. Most of the selected algorithms showed success in previous studies except for the last one, which was chosen in hopes of it being benefitted from the new style of data used for this research.

3.4.2 Random Forests

The first algorithm selected is Random Forests (RF) [24]. This method showed one of the best overall performances when used in the research [8]. RF is an ensemble method that combines the results of multiple decision trees to reach a unique consensus. Its design derives from the conclusion that as long as they are independent and better than random, the predictions' accuracy improves when the results of different algorithms are combined. The RF algorithm maintains independence between the decision trees in question by feeding each one different random subsets of features. This design also makes RF particularly useful for highly dimensional tasks like the one taking place in this study. RF uses a weighted average method to reach a final consensus, which diminishes the interference of insignificant attributes that otherwise would have deteriorated the prediction's quality.

3.4.3 Gradient Boosting

Continuing with ensemble methods, the second algorithm in this list is Gradient Boosting (GB) [25], the best-performing machine learning algorithm in the deep

learning [7] research. Equal to RF, GB uses a weighted average system to discern between multiple decision tree's results to arrive at a final prediction. However, it constructs every new decision tree based on the performance of the previous one, focussing on those cases that caused problems in it. The idea behind this design is that each new tree complements the forest, specializing in deeper and deeper levels of complexity. This method is particularly convenient for this research because of the same dimensionality characteristics explained for RF.

3.4.4 Elastic Net

The third method is called Elastic Net [25], and its precedent in the [8] research showed better performance than the classic phenotypic prediction tools. This algorithm is a least-sum-of-squares-based linear regressor that uses lasso and ridge regression for variable selection and regularization. Lasso (or L1) regularization adds a penalizing system that annuls useless features, tackling the dimensionality problem. The ridge (or L2) regression introduces a bias that minimizes the overall variance (in training and testing data) and avoids overfitting. However simple, this algorithm seems to be a perfect suite for this study, as it has an appropriate design to handle the data's dimensions.

3.4.5 Multy-Layer Perceptron

Finally, the last algorithm to be included in this study is the multi-layer perceptron (MLP) [26]. Although this method performed sub-optimally in past genome-wide association studies [7], it would be premature to discard it. In a few words, this algorithm is an array of parameters organized in layers of nodes through which data flows. The output layer (the network's last layer) makes the final prediction, and optimizer algorithms are used to minimize the network's loss by correcting each parameter. Two significant deep learning advantages are

its capacity to precisely model highly complex systems and its ability to handle many features without affecting the prediction's accuracy. These qualities are the encouraging factor that makes MLP worth including in this research, despite MLP's past performance.

Chapter 4

Experimental Settings

4.1 Cleaning and Pre-processing

4.1.1 Introduction

Before carrying out the experiments, the data must be thoroughly cleaned and organized. This step is particularly complex in this study since the data originates from two large datasets with numerous features. Although there is a five-year gap between the making of these two datasets, this difference only manifests itself in a few columns with different names. This concordance is logical since both datasets were designed by the same organization and with the same purpose. Rather than inconsistency, the main problems these datasets presented involved inadequate formats, useless and redundant attributes, and missing values in several features. Next, a detailed explanation of the techniques used to overcome these obstacles will be presented, along with an overview of the final dataset.

4.1.2 Dropping Useless Features

A total of eleven features were discarded during the data cleaning process. This decision seems less radical when considering that the number of attributes of the two original datasets adds up to one hundred and ninety-six. The first and most frequent reason that led to the removal of attributes was that these attributes offered no additional discriminatory power to any possible model. They either presented only one unique value or a different categorical value for every event, becoming useless from an analytical standpoint. The **"HBNO"** feature was the first one dropped for this reason. It refers to the "Herdbook Number" [27], an identification code assigned to herds used for disease control purposes [28]. It would be logical that, as it is assigned to herds and not individual animals, these would allow to locate every bull within a herd and consequently be analytically valuable. Nevertheless, since every value in this column is unique, this possibility is eliminated, and the attribute is therefore useless. The case of the **"Name of Bull"** feature is very similar. As it is predictable from its nature, there are no two bulls with the same name, and this column brings no potential value for any possible model. Finally, four evaluation traits in the 2017 dataset (**"Overall Type TPS"**, **"Overall Type Frame"**, **"Carcass Trait Evaluation Wean Wt"**, and **"Carcass Trait Evaluation Rel Wean Wt"**) that aim to provide specific information about the animal's physical composition. However, they all fail to do so in this sample. The reason is the lack of any diversity of values in said columns. All four of these attributes presented one unique value repeated for every event, offering no potential discriminatory power to a model and thus provoking their removal. The second reason that caused the discarding of features was as simple as their duplication. That was the case for the **"Production Trait Evaluation Rel %"** and **"Fertility and Survival Rel %"** features. These values show the reliability levels for the milk and fertility sub-indexes, re-

spectively. However, the **"Milk Prod SI Rel"** and **"Fertility SI Rel"** features had already provided this information, making them redundant. The third and last motive that caused the removal of three attributes was simplification. These three attributes contain very sparse categorical data, which is problematic to deal with when working with many machine learning algorithms. Additionally, these features showed many missing values, which would be very difficult to fill given their sparse nature and significantly reduce the number of events in my final dataset. For this combination of reasons, the **"Name of Sire"**, **"Name of Dam"** and **"Test Stat"** features were eliminated. As a final remark, it is appropriate to mention that only the Bull ID, the type of proof (DP or GS), and the target features were taken from the 2022 dataset since those were the only ones relevant to the current study.

4.1.3 Correcting Format Issues

This problem arose in the "Birth Date" feature of the 2017 dataset, where the values kept their original five-digit Microsoft Excel date format. The values had to be changed to the same format as the one used in the 2022 dataset. This format allowed easier comparison and handling of the data. It consisted solely of the year (four-digit integer) in which the animal was born. The built-in "TimedeltaIndex" Pandas function was used to correct this issue and transform the values into DateTime format, keeping only the year and ignoring the day and month.

4.1.4 Missing Values

Before going through the problems related to missing values and the techniques used to sort them out, it is pertinent to mention the significance of every single event in every dataset. Although the original 2017 and 2022 datasets contained over 4200 and 5700 events respectively, which seems to be more than enough

data to conduct the experiments, the reality is more complex. Out of all these bulls, only those that were genomically selected in 2017 and daughter-proven in 2022 are relevant for the ongoing study, drastically reducing the number of records. In the 2017 dataset, out of 4248 bulls, only 291 were originally labeled as genomically selected. Nevertheless, there were a total of 203 missing values in this feature. This high volume of missing values in such a critical feature like this one motivated the decision to fill the missing values using a simple binary classification algorithm to reduce the loss of records in the dataset to a minimum. The model used for this task was the Decision Tree Classifier, taken from the Scikit-Learn library. The main concern for this procedure was the threat of including many mislabelled records into the final dataset, which would probably cause more harm than benefit in the long run. However, the decision tree showed a highly reliable performance, reaching a recall score of over 95% in the testing data for the 2017 "Type of proof" missing value imputation. The same algorithm was later used to fill the missing values in the same feature containing the same data but for the 2022 dataset. Similarly to the previous case, this column had 222 missing values that threatened to significantly reduce the final volumes of data available for the experiments. In this instance, the model showed an even higher recall score of 99.7%, which caused further reassurance on the method selected for the imputation. The result of this imputation allowed the final dataset to hold a total of 379 records. Had all the missing values been discarded immediately, the final dataset would have only been composed of 272 records making this data imputation very valuable for the development of the study.

4.1.5 Merging and Pre-processing

The merging process was simple and straightforward. The shared features on which the 2017 and 2022 datasets were joined were the "AI Code" and "Bull"

columns, respectively. These two contain a unique identification code that allows tracking each animal's breeding records on the ICBF dataset, making it perfect for aligning every event on one dataset with its corresponding in the other. An "inner" merge was used to keep only those records present in both datasets. As previously mentioned, not every event is relevant for this study. For that reason, the merged dataset kept only the records of bulls that were GS in the 2017 dataset and DP in the 2022 dataset, dropping the rest. And lastly, this merge caused four features to become redundant, for which they were discarded. These features provided the bull's identification code and type of proof in each dataset. Since now there is no need to individually identify the bulls because they all carry the same proof type (for each year of measurement respectively), these features had no contribution to the remainder of this study. Regarding the pre-processing needed to carry out the experiments appropriately, there was one common issue to tackle. Four different categorical features were expressed in non-numeric values. These values must be numeric to be fitted to the machine learning algorithms used in the experimental process. The methodology to solve this issue was selected on a case-by-case basis. The first two features were transformed using binary encoding. This method converts the values into either a 0 or a 1. Even though these two features were not originally binary, their values' frequency demonstrated that this simplification would not cause a lot of information to be lost in the process. The 'Main Breed" feature offered a short abbreviation of the name of the bull's main breed. However, as the Holstein breed is the most used breed for dairy production worldwide (and by a long stretch), this breed represented almost 75% of the values in the column. The replacing column was named "Main Breed = HO" and contained a 1 for every Holstein bull and a 0 for other breeds. The other column transformed using this method was the "Cat" feature, which showed the bull's ancestral category. Since more than 70% of the

bulls belonged to the same category (pedigree), the replacing column was named "Cat = Pedigree", representing the pedigree bulls with a 1 and others with a 0. Finally, the "Owner Name" feature presented a more complex challenge. In this case, the values were more sparsely distributed, with its mode repeated 128 times (around one-third of the total number of records), which indicated that a binary encoding approach might be overly simplistic. Another commonly used method for encoding categorical variables is called "One-Hot Encoding", and it works by creating a new column for each possible category and assigning a 1 to the cases in which that category is present and a 0 when it is not. However, performing one-hot encoding on this feature would significantly increase the dataset's dimensions since it would need to create 15 additional columns. The final approach came as a compromise between the two previously mentioned options, using one-hot encoding for the four most frequent bull owners (representing 78% of the total number of records) and blending the reminding twelve as "Other". As a result, only four new columns were needed, and only 22% of the values were unified as "Other". After this process, the data was ready to be separated between the input data and the targets. The input data had a final size of 87 features with 379 events.

4.2 Hyperparameter Tunning

4.2.1 Optimization approach

The methodology proposed in this research is broad and explorative. It aims to test four machine learning algorithms for each of the thirteen selected targets. This approach adds up to fifty-two different models that, after being fine-tuned and optimized, will be compared against each other to determine which are better than the others. Not only the training of such a large number of models

brings a practical burden to the research, but the multiplicity of approaches that can be taken while training an algorithm puts the objectivity of the evaluation in question. Using one single methodology replicated for each algorithm would make the comparison more impartial. However, this problem shines a light on one of the benefits of working with a set of algorithms that can be found in the Scikit-Learn library. In addition to a long list of machine learning models, this library offers multiple tools that can be used for predictive data analysis. In the case of this study, the "Randomized Search" and "Grid Search" tools for hyperparameter tunning were used as a standardized way of optimizing each algorithm. The Randomized Search algorithm is a method used to randomly test multiple hyperparameter combinations from a long array of options [29]. This function automatically keeps records of the scores showed by each combination of hyperparameters tested. Since the procedure does not run every possible combination of hyperparameters, it is frequently used to explore a wide range and quantity of settings. These settings are defined as a grid from which the optimizer selects one value per hyperparameter in a random fashion, repeating the process until it reaches a certain number of iterations. The Grid Search approach has a similar procedure [30]. It also runs tests using hyperparameter values taken from a grid and keeps records of every trial's setting and performance. However, this method is more exhaustive since it tests every possible combination of hyperparameters in the grid. The fact that there is no limit to the number of iterations and the objective is to run every possible combination means that every hyperparameter value added to the grid will demand an increasingly high amount of computational effort. Consequently, this technique is appropriate for narrower and more restricted hyperparameter grids and is widely used for fine-tuning the models. For this research, the Randomized Search and Grid Search tools were implemented sequentially for each algorithm. The objective was to test a large and diverse

number of options using the Randomized Search tool and subsequently make a new hyperparameter grid based on the best-performing hyperparameter combination resulting from this search. The new grid would contain a spectrum of values very close to the original and then be used for a final Grid Search. This second search would potentially allow for additional improvement with a more fine-tuned set of hyperparameters. Next, this report will provide deeper insights related to the design of these grids for each of the four selected algorithms.

4.2.2 Random Forest

Every algorithm has a different set of adjustable hyperparameters, some more influential than others. Therefore, an efficiently designed grid for Randomized Search would include an array of values for only the most consequential hyperparameters. In the case of the Random Forest Regressor, the selected group of hyperparameters was the following:

1. "n_estimators" [31]: The number of estimators defines the number of trees that will make the forest and is set to one hundred by default. The randomized search in this study used a set of ten evenly distributed numbers between ten and four hundred. That list was automatically generated using the "linspace" function from the NumPy library.

2. "max_features" [31]: This hyperparameter defines the number of features considered when calculating the best split within each tree. The values included in the grid were 1.0 (meaning that all features would be taken into consideration) and "sqrt" (meaning that the square root of the total number of features would be taken into consideration).

3. "max_depth [31]": The max depth hyperparameter allows to set of a limit for the maximum depth of the trees in the forest. Similar to the "n_estimators"

hyperparameter, the "linespace" function was used to generate an evenly distributed list of numbers from ten to one hundred and ten. Additionally, the "none" option was added to the list, which allows the trees to continue expanding without a limit.

4. "min_samples_split" [31]: This hyperparameter sets the minimum number of samples required to conduct a split. The default value is two, which is the lowest possible value. A short list of small values ([2, 5, 10]) was selected for the randomized search.

5. "min_samples_leaf" [31]: The minimum samples leaf offers the option to limit the number of samples required for a partition at any given tree depth. This could be used to avoid an excessive number of small-sized leaves. The array of values selected for the randomized search consisted of one, two, and four leaves.

6. "bootstrap" [31]: It is possible to determine whether the algorithm should use the entire dataset for each tree or if a "bootstrap" set of samples should be used for each tree, by setting this hyperparameter to either "True" or "False". Both options were considered for the randomized search.

4.2.3 Gradient Boosting

Although it included a smaller number of hyperparameters, the grid designed for the Gradient Boosting Regressor contained a wide array of values for each of them. This grid was composed of:

1. "n_estimators" [32]: This hyperparameter is similar to its namesake in the previous algorithm. It defines the number of boosting stages to be carried out by the model. The randomized grid experimented with an exponentially

growing set of nine values starting from one and finishing at five hundred for this hyperparameter.

2. "max_leaf_nodes" [32]: The maximum leaf nodes hyperparameter allows to set a limit to the final number of leaves. It has a similar effect to the one that the "min_samples_leaf" has on the Random Forest Regressor, which is to smooth down the model by lowering the number of leaves. The array selected for this stage contains nine exponentially growing values from two to one hundred.

3. "learning_rate" [32]: This rate determines the ratio in which every tree will attempt to reduce the previous tree's loss, therefore presenting a trade-off between accuracy and computation time. For this rate, the randomized search used a list of ten evenly distributed numbers between 0.1 and 1.

4.2.4 Elastic Net

The hyperparameters selected for optimization for this algorithm were the following:

1. "alpha" [33]: This value determines the magnitude of the penalization designed to make useless features less impactful. A list of nine exponentially growing values starting from cero and finishing in one hundred were selected for this hyperparameter.

2. "l1_ratio" [33]: This measure regulates the proportions of L1 and L2 regularization. The "linespace" formula was again used to automatically generate one hundred evenly distributed values between zero and one.

4.2.5 Multi-Layer Perceptron

The final algorithm presented a large variety of impactful hyperparameters to adjust, which is logical given the complexity of its nature. The final selection of hyperparameters included:

1. "hidden_layer_sizes" [34]: This hyperparameter sets the network's shape by setting the number of hidden layers and nodes in each layer. The array of shapes used for this experiment included combinations of two and three hidden layers containing twenty-five, fifty, and one hundred nodes.

2. "activation" [34]: The activation hyperparameter allows for the selection of the network's activation function. The available options include the "identity" (also known as linear), "logistic", "tanh", and "relu" functions, all of which were used for this study.

3. "solver" [34]: This hyperparameter sets the solver to be used for adjusting the weights. Like in the previous parameter, all possible options were taken into consideration. These options are the "adam", "sgd", and "lbfgs" solver algorithms.

4. "learning_rate" [34]: The learning rate hyperparameter is similar to its namesake in the Gradient Boosting Regressor algorithm. However, instead of offering the option of setting a fixed rate, it provides three different schedules ("constant", "invscaling", and "adaptive"), all of which were used in this research.

5. "alpha" [34]: This final hyperparameter sets the influence of the L2 regularization term. The values selected for the Randomized Search were 0.001, 0.0001, 0.00001, and 0.000001.

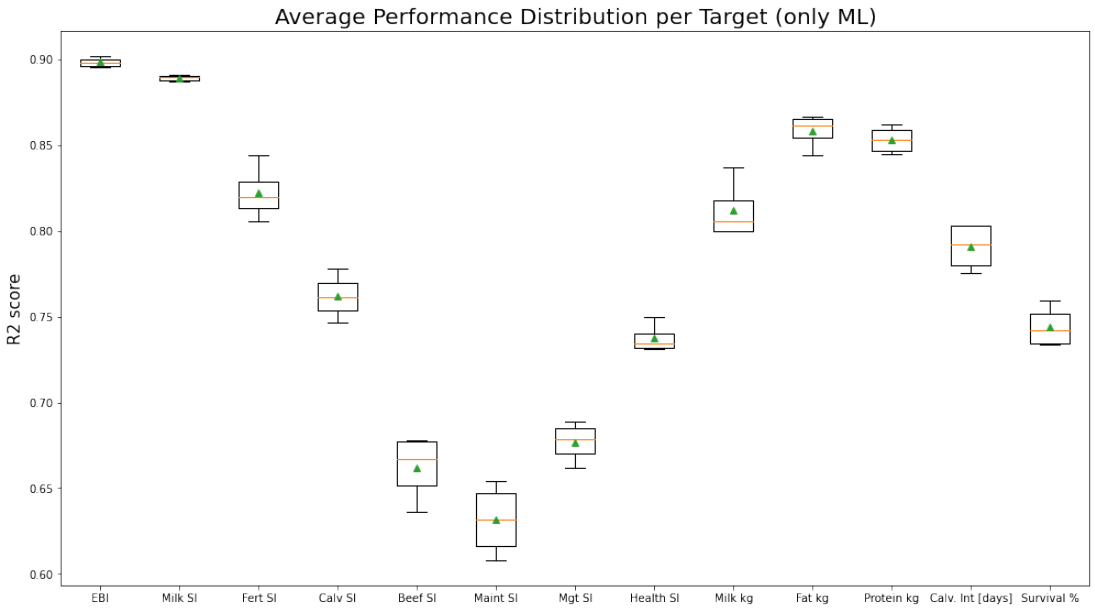
Chapter 5

Results

5.1 Introduction

The R2 score was the selected evaluation method for all the targets in this study. This metric shows the variation or error proportionally to the variable [35]. The fact that the error is expressed proportionally and not in absolute numbers allows for easier comparison between variables with different ranges, which was the reason it was selected for evaluating all thirteen targets. Figure 5.1 illustrates the performance distribution shown by the ML models for each target. Across the different targets, the performance showed by the algorithms tested was notably different, averaging scores between 60% and 90%. However, the original GWAS predictions also fluctuated among the different targets in a similar fashion. In fact, those predictions maintained a correlation of 0.79 with the average R2 score achieved by the ML models for each target, indicating that the prediction difficulty for each target was similar between the two approaches.

Figure 5.1: Performance Distribution per Target



For this reason, comparing each method's performance as an average of its R2 score for all targets would be overly simplistic and it would dismiss valuable information about the differences between the machine learning and the GWAS approach. Therefore, the only appropriate way to evaluate the machine learning algorithms' performance is by comparing their R2 score against the benchmark established by the GWAS on a target-by-target basis. This comparison can be appreciated in figure 5.2, where the R2 score of each algorithm and each target is shown in relation to its relative benchmark score.

Figure 5.2: Experimental Results

5.2 Target Evaluation

Figure 5.2 reveals that all four models show relatively similar performance for each target, which allows them to be analysed as a whole in comparison to the benchmark. In this context, the first remark to be made is that, for eight out of the thirteen targets, the machine learning models showed a significant improvement. These targets are the EBI, Milk Sub-Index, Fertility Sub-Index, Calving Sub-Index, Maintenance Sub-Index, Management Sub-Index, Calving Interval, and Survival percentage, which showed an increase in R2 score of between 3% and 30%, with a combined increase of 13%. In the second group, there are three targets for which the machine learning algorithms showed little to no difference in performance. These targets are the Health Sub-Index, Fat Yield, and Protein Yield, which reached R2 scores of between -1% and 1% relative to the GWAS

benchmark for those targets. Finally, the Beef Sub-Index and the Milk Yield targets showed the worst performance relative to their respective benchmarks, with an R2 difference of -8% and -4%, respectively.

5.3 Method Comparison

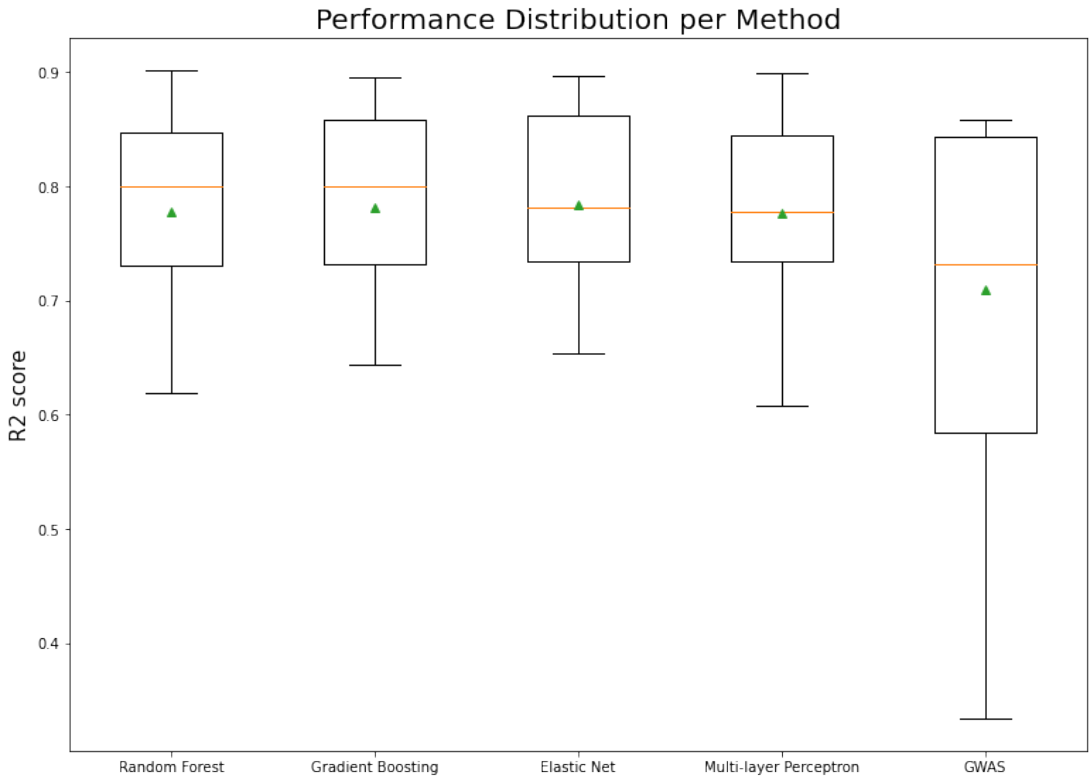
Since each model was used to predict thirteen different targets, assessing their performance using a single accuracy metric would not be the most appropriate way of comparing them. The average R2 score across all targets is a valuable metric since it allows to put all four models in a unidimensional space where their preciseness can be easily compared. Nevertheless, this doesn't account for the differences in performance across the targets, opening the gates for highly inconsistent algorithms to rank higher than the others, which is not ideal. The best algorithm should be the one that makes the best and most consistent predictions. To consider this variable, the R2 score variance across the targets will complement the average R2 score by showing how reliable this score is across the different algorithms.

Table 5.1: Performance Per Method

	Random Forest	Gradient Boosting	Elastic Net	Multi-layer Perceptron	GWAS
R2 Variance	0.008022	0.006824	**0.006751**	0.006765	0.022960
Average R2	0.777576	0.781550	**0.784021**	0.776359	0.709621

Figure 5.3 illustrates the performance distribution of each predictive method, including the GWAS approach. Additionally, the average R2 score and R2 variance for each methodology are displayed in table 5.1 to provide a more precise outlook of the algorithms' performance.

Figure 5.3: Performance Distribution per Method



The Elastic Net algorithm showed the highest R2 score and the lowest variance, attaining the best record in the two selected evaluation metrics. The Gradient Boosting Regressor's metrics were very close to those shown by the Elastic Net algorithm, ranking second highest in average R2 and third lowest in R2 variance. The Multi-Layer Perceptron ranked second lowest in R2 variance, but its accuracy was the worst among the ML methods. Finally, the Random Forest Regressor showed the worst overall performance, with the highest R2 variance and the second lowest average R2 score. As a final remark, it is pertinent to mention that, when looking at all thirteen aggregated targets, the original GWAS predictions' performance showed substantially inferior results than what the machine learning methods did. Although the original method achieved a higher R2

score in the prediction of the Beef Sub-Index and Milk Yield targets, its average R2 score, and R2 score variance are significantly outside the range in which the machine learning models performed.

Chapter 6

Conclusions

6.1 Genomic Selection and Machine Learning

The use of genomic selection in pursuit of more phenotypically efficient dairy cows throughout the generations has provoked staggering advances in the last fifteen years[4]. The effect that the implementation of GWAS has had in this industry is undeniably beneficial for producers since they are achieving a more profitable output [1]. It's also benefiting the consumers, which are buying cheaper dairy products. And finally, it is beneficial for the environment since higher efficiency translates to fewer greenhouse gasses emitted per unit of milk produced by each animal [1]. Furthermore, the relative novelty of these methods suggests that there is still a lot to gain from their continuous development. Fifteen years of applied research is a relatively short time span, considering the complexity of the task of analysing SNPs and identifying the complex relation between genomic markers and phenotypic traits. The highly complex, lengthy, and laborious nature of this analysis opens the door for modern computational sciences to contribute to the matter, and Machine Learning is one of the most evident. This idea stems not only from the proven Machine Learning's capacity for modelling highly complex

systems and relations between multiple variables but also from the observable room for improvement that current techniques show, despite having achieved major advances [5]. This study explored the performance of several Machine Learning algorithms when predicting phenotypic traits in dairy cattle by taking GWAS-based predictions and additional factual data as the inputs for training the models. The results showed that these techniques have the potential to achieve more accurate and consistent results than what GWAS is currently reaching. In particular, the Elastic Net algorithm outperformed the benchmark in eleven of the thirteen selected targets and positioned itself as the best-performing model in the study with the best average R2 score and lowest variance throughout all the targets. In comparison with the original GWAS's predictions, the Elastic Net algorithm showed a 7.4% higher average R2 score and a 1.6% lower R2 score variance throughout the different targets. The results attained in this experiment reinforce the initial idea that Machine Learning has the capacity to enhance the performance that current phenotype predicting tools are achieving. Furthermore, these results suggest the possibility of reaching more reliable predictions through the use of different types of data. Starting from the fact that these algorithms were fed a combination of GWAS-based predictions and factual data, it presents the idea of adding further data. Feasible examples would be using milk infrared spectral data, like the one used in [8], or raw SNPs data, like the one used in [6] and [7]. However, it is pertinent to mention that the algorithms used would need to attenuate the influence of less relevant features and that the use of additional data would always come at the cost of higher running times.

6.2 GWAS Performance Fluctuation

As mentioned earlier, the average R2 score of all ML models for a specific trait is highly correlated with that trait's GWAS R2 score, indicating that these two methods have similar difficulties in predicting some targets. A possible explanation is that, since the Machine Learning models take GWAS predictions as part of their input data, these are inheriting the performance of the predictions they are being fed. Beyond the question of how correct this supposition is, a more fundamental question arises. What makes the GWAS' predictions vary so much among different targets? Answering this question would take a deeper genetic engineering analysis than what this study can carry out. Nevertheless, a strong relationship made visible during this study might help to guide future researchers toward the answer. The GWAS-based predictions' R2 score for each of the seven sub-indexes shows a correlation of 0.53 with the sub-indexes' ponderation in the overall EBI value. In simpler words, the most important and valuable sub-indexes show the most accurate predictions. This correlation suggests that, probably, the reason why the GWAS predictions are more accurate for some targets than others is that the predictive analysis is optimized to prioritize more economically influential targets, compromising the rest's performance.

6.3 Future Work

As mentioned in the previous paragraph, this research could be complemented by further studies, assessing the predictive ability of several Machine Learning models using an even more diverse selection of input data. Active collaboration with breeding organizations like the ICBF would be useful to effectively explore all the available data on active bulls, and build more comprehensive models and evaluate their performance. Additionally, this research did not include a specific

analysis that would be useful to project future studies. This analysis would explore which features were given more importance by the models and which were given less. Needless to say, this is not a simple task since not all algorithms allow for a comprehensible internal examination. Nevertheless, it is possible to carry out this analysis using the Elastic Net algorithm [33], which was the best performing model in this study, and use the insights discovered to guide future research in this field. Regarding the potential value of the models built in the curse of this study, their performance suggests that they could be used to enhance current phenotypic prediction's accuracy for some specific traits. In particular, the Elastic Net algorithm shows a substantial accuracy gain in eight of the targets selected for this study (EBI, Milk Sub-Index, Fertility Sub-Index, Calving Sub-Index, Maintenance Sub-Index, Management Sub-Index, Calving Interval, and Survival percentage), with an average 12.5% increase in R2 score over the original GWAS's predictions. The use of this model could therefore give dairy farmers more reliable information when selecting bulls to inseminate their herds, which would contribute to the long-term improvement of dairy production efficiency.

References

[1] ICBF, "Ebi continues to maximise profitability and sustainability on irish dairy farms," https://www.icbf.com/?p=17909, 2021.

[2] ——, "What is genomics?" https://www.icbf.com/?p=5831, 2019.

[3] ——, "Understanding the economic breeding index (ebi)," https://www.icbf.com/wp-content/uploads/2020/02/Understanding-EBI-PTA-BV-Spring-2020.pdf, 2018.

[4] ——, "20 years ebi and sub indices national averages," https://www.icbf.com/?p=17520, 2021.

[5] ——, "Understanding reliability and genetic index," https://www.icbf.com/?p=5806, 2018.

[6] L. Yin, H. Zhang, X. Zhou, X. Yuan, S. Zhao, X. Li, and X. Liu, "Kaml: improving genomic prediction accuracy of complex traits using machine learning determined parameters," *Genome biology*, vol. 21, no. 1, pp. 1–22, 2020.

[7] R. Abdollahi-Arpanahi, D. Gianola, and F. Peñagaricano, "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes," *Genetics Selection Evolution*, vol. 52, no. 1, pp. 1–15, 2020.

[8] L. F. Mota, S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato, "Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in holstein dairy cattle using milk infrared spectral data," *Journal of Dairy Science*, vol. 104, no. 7, pp. 8107–8121, 2021.

[9] J. Pryce and H. Daetwyler, "Designing dairy cattle breeding schemes under genomic selection: a review of international research," *Animal Production Science*, vol. 52, no. 3, pp. 107–114, 2011.

[10] D. Bickhart, J. McClure, R. Schnabel, B. Rosen, J. Medrano, and T. Smith, "Symposium review: advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection," *Journal of dairy science*, vol. 103, no. 6, pp. 5278–5290, 2020.

[11] Doekes, H. P, Veerkamp, R. F, Bijma, Piter, de Jong, Gerben, Hiemstra, S. J, Windig, and J. J, "Inbreeding depression due to recent and ancient inbreeding in dutch holsteinfriesian dairy cattle," *Genetics Selection Evolution*, vol. 51, no. 1, pp. 1--16, 2019.

[12] J. B. Cole, S. A. Eaglen, C. Maltecca, H. A. Mulder, and J. E. Pryce, "The future of phenomics in dairy cattle breeding," *Animal Frontiers*, vol. 10, no. 2, pp. 37–44, 2020.

[13] OECD, "Dairy and dairy products," https://www.oecd-ilibrary.org/sites/aa3fa6a0-en/index.html?itemId=/content/component/aa3fa6a0-en#section-d1e19383, 2019.

[14] D. Berrya, L. Shallooa, A. Cromieb, V. Olorib, R. Veerkampc, P. Dillona, P. Amerd, R. Evansb, F. Kearneyb, and B. Wickhamb, "The economic breed-

ing index: a generation on," https://www.icbf.com/wp-content/uploads/2013/06/economic_breeding_index.pdf, 2007.

[15] ICBF, "Proofs for bulls do change as they accumulate more data," https://www.icbf.com/?p=15831, 2020.

[16] ——, "Breeding 2019 – should i use daughter proven or genomic sires?" https://www.icbf.com/?p=10405, 2019.

[17] ——, "Top 10 most used sires 2021," https://www.icbf.com/?p=17657, 2021.

[18] B. Li, N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li, "Genomic prediction of breeding values using a subset of snps identified by three machine learning methods," *Frontiers in Genetics*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2018.00237

[19] B. Li, P. VanRaden, E. Guduk, J. O'Connell, D. Null, E. Connor, M. VandeHaar, R. Tempelman, K. Weigel, and J. Cole, "Genomic prediction of residual feed intake in us holstein dairy cattle," *Journal of Dairy Science*, vol. 103, no. 3, pp. 2477–2486, 2020.

[20] ICBF, "Dairy ebi," https://www.icbf.com/?page_id=202, 2022.

[21] ——, "Health and genotypes: Another big step for icbf," https://www.icbf.com/?p=6092, 2016.

[22] ——, "Icbf dairy ebi bull proof file," https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.icbf.com%2Fwp-content%2Fuploads%2F2022%2F05%2FWebVersion_All_Dairy_web_240522_20220524_SR.xlsx&wdOrigin=BROWSELINK, 2022.

[23] ——, "Proofs for ebi, sub indexes, and individual traits for dairy ai bulls with a progeny based calving proof decemeber 2017," https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.icbf.com%2Fwp-content%2Fuploads%2F2017%2F12%2Fall_ebi_dec2017.xlsx&wdOrigin=BROWSELINK, 2017.

[24] T. Yiu, "Understanding random forest," https://towardsdatascience.com/understanding-random-forest-58381e0602d2, 2019.

[25] T. Masui, "All you need to know about gradient boosting algorithm part 1. regression," https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502, 2022.

[26] C. Bento, "Multilayer perceptron explained with a real-life example and python code: Sentiment analysis," https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analys 2021.

[27] ICBF, "Notes and glossary," https://www.icbf.com/?page_id=210.

[28] F. Department of Agriculture and the Marine, "Apply for a flock or herd number," https://www.gov.ie/en/service/f90f21-application-for-a-flockherd-number/#:~:text=A%20herd%20or%20flock%20number%20for%20cattle%2C%20sheep%20or%20goats,kept%20under%20that%20herd%20number., 2020.

[29] Scikit-Learn, "Randomizedsearchcv," https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html#.

[30] ——, "Gridsearchcv," https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[31] ——, "Randomforestregressor," https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[32] ——, "Gradientboostingregressor," https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

[33] ——, "Elasticnet," https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.

[34] ——, "Mlpregressor," https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html.

[35] ——, "r2_score," https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

Appendix A

Code Base

https://github.com/e-lopez-taranto/MSc-DA-Thesis-21250773-Lopez