Alma Mater Studiorum · Università di Bologna

Dottorato di Ricerca in Salute, Sicurezza e Sistemi del Verde
Ciclo 35°

Settore Concorsuale: 07/C1
Settore Scientifico Disciplinare: AGR/10

Big Data approaches as a support for Precision Livestock Farming techniques

Presentata da:
Miki Agrusti

Coordinatore Dottorato:
Prof.ssa Patrizia Tassinari

Supervisore:
Prof. Stefano Benni

Co-supervisori:
Dott. Ing. Marco Bovo
Prof. Daniel Remondini
Prof. David Nail Manners

Esame finale anno 2023

"*Ignoranti quem portum petat*
nullus suus ventus est"

Seneca

Abstract

With the advent of new technologies it is increasingly easier to find data of different nature from even more accurate sensors that measure the most disparate physical quantities and with different methodologies. The collection of data thus becomes progressively important and takes the form of archiving, cataloging and online and offline consultation of information. Over time, the amount of data collected can become so relevant that it contains information that cannot be easily explored manually or with basic statistical techniques. The use of Big Data therefore becomes the object of more advanced investigation techniques, such as Machine Learning and Deep Learning. In this work some applications in the world of precision zootechnics and heat stress accused by dairy cows are described. Experimental Italian and German stables were involved for the training and testing of the Random Forest algorithm, obtaining a prediction of milk production depending on the microclimatic conditions of the previous days with satisfactory accuracy. Furthermore, in order to identify an objective method for identifying production drops, compared to the Wood model, typically used as an analytical model of the lactation curve, a Robust Statistics technique was used. Its application on some sample lactations and the results obtained allow us to be confident about the use of this method in the future.

Contents

Abstract

Contents

Chapter 1

Introduction and goals

The attention paid to animal health and welfare is increasingly predominant in modern farming techniques. Better living conditions for animals, a lower amount of stress and a more targeted use of medicines to combat the onset of diseases has a direct consequence in the food chain and therefore indirectly also in humans.
Furthermore, the climatic changes involving the last decades have produced a sudden increase in average temperatures, with serious repercussions on the habits of the animals, on their well-being and productivity.
Precision Livestock Farming has among its objectives a more scientific and precision management of farms, in order to guarantee animals greater well-being and living conditions.
From the PLF perspective, new techniques can be used to continuously monitor the vital parameters of animals, starting with the use of motion sensors to characterize their habits, up to sensors that measure the percentage of substances present in milk, ending with devices capable of to record the rhythm of breathing.
Over the years, technological advancement and the introduction of new computational techniques have made it possible to make the best use of the data collected by the available instrumentation, making it easier to exploit the potential of Big Data.
At the same time, also another aspect can be highlighted: the computational power of computers is increasing from year to year and this allows operations that previously required days to be completed in fractions of a second.
This makes it faster today to train algorithms on a huge amount of data, without the need for an incredibly long time.
In some cases an in-line learning process is even possible, whereby the algorithm incorporates the new examples into the training process and becomes more and more accurate.

Modern analysis techniques have proved to be valid for the interpretation of complex problems, such as classifications, clustering and forecasts. An example of this are the Machine Learning and Deep learning algorithms which today form the basis of operation of household appliances, devices and vehicles, as well as being a valid scientific research tool.

1.1 Main goals

The introduction of technology in farms and the increasingly frequent use of precision measuring instruments for the control and monitoring of production and microenvironmental parameters makes it easier to use statistical techniques already widely exploited in other fields.
Furthermore, Precision Livestock Farming (PLF) introduces the need to explore animal data by observing the characteristics of each individual and customizing these techniques as much as possible.
This thesis was born in the context of the PRIN projects and ECPLF conference, with the intention of finding an application of the main data analysis techniques to the world of precision zootechnics.
In particular, it is known that production is influenced by the microclimatic conditions of the barn. Hence the need to continuously monitor the temperature and humidity of the stables and use this data for an increasingly accurate forecast of milk production.
A first objective of this work is to organize the reading of the reports produced by the milking robots. This is essential for a correct handling of the data, which must be as homogeneous and comparable as possible.
A second goal is to find a mathematical model to predict the response of the individual animal to microclimatic variations.
The study of the state of the art, the search for documents and articles where the models used before are described are themselves part of the thesis work. It was also necessary to identify an appropriate model for the amount of data available and to use a type of training that would make the most of the information available.
In addition to this, this thesis tries to introduce an objective method for the detection of production anomalies, caused, as known in literature, by different factors like heat stress, disease and infections.
All these applications agree in showing how it is possible to positively use technology in favor of animal and human well-being, by predicting or recognizing abnormal animal behavior in advance and guaranteeing

a healthy and safe product.

1.2 State of Art

Sustainability is an unavoidable goal for animal-derived products due to the mounting pressure on the livestock sector to meet the growing demand of an increasing population with rising incomes and the need to reduce the exploitation of resources and the environmental impact, while safeguarding animal welfare (Dawkins, 2017).
At the same time, global climate change and environmental crises are also challenging the dairy sector, and they will represent increasingly important issues to be addressed to ensure its economic, environmental, and social sustainability.
In the dairy sector, the cornerstones of sustainability can be recognized as milk production and quality, cow health and welfare, efficiency in the use of resources, and emissions reduction. Animal welfare is strictly related to sustainability, due to the consequences in terms of milk quantity and quality, which affect the efficiency of the use of natural resources. For this purpose, a crucial point is the prevention of heat stress, as it markedly jeopardizes animal welfare in several countries in the Mediterranean area.
Equipment based on Information and Communication technology (ICT) are increasingly installed in livestock barns to perform a wide range of operations from climatic control to milking, from precision feeding to cleaning (Tassinari et al., 2021). These devices are coupled with manifold sensors which collect data needed for a proper operation of the equipment and, therefore, large amounts of data are recorded nowadays in a livestock farm equipped with ICT systems. This is a significant aspect of the Precision Livestock Farming (PLF) approach, which is involving the livestock farming sector in a fast process and is providing farms with great opportunities of improvement of the production performance and the conditions of animal welfare, independently of the farm size (Berckmans, 2014) (Fournel et al., 2017).
In the dairy cattle sector, the availability of data recorded in real time concerning the environmental conditions of the barn and the production performances of the individual cows represent a quantitative knowledge basis with a huge potential of development of further informatic and electronic tools, able to achieve optimal conditions of animal welfare and more sustainable productions, in addition to improvements in milk

quality and production efficiency (Lovarelli et al., 2020b).
In particular, the ever more widespread Automatic Milking Systems (AMSs) provide farmers with detailed data concerning health conditions and parameters connected to the milk produced, which are of great interest to optimize the production (John et al., 2016) (Rotz et al., 2003). Moreover, in technological farms, data concerning different parameters of behavior and activity of cows, animal health and welfare are collected from different sensors (e.g., individual cow data recording system, activity tags such as pedometers or neck collars, ear tags for rumination monitoring, automatic concentrate feeders), and used for the daily management of the herd (Halachmi et al., 2019).
To gain a comprehensive understanding of these phenomena and monitor and control the production processes in relation to climate change in a sustainable intensification perspective, sophisticated and high-throughput data acquisition is needed, providing very heterogeneous and multichannel datasets.
Several studies have shown that a proper storage of collected data in structured databases represents a necessary preliminary step for the development of numerical models suitable to characterize the conditions and performance of individual cows (Bonora et al., 2018a) and to quantify the effects of particular thermo-hygrometric conditions on milk production (Bonora et al., 2018b)(Benni et al., 2020). In a climate change scenario, the welfare of dairy cows exposed to heat waves is becoming increasingly important (Cowley et al., 2015).
Moreover, cow activity response to heat load was recently investigated (Heinicke et al., 2019). Cows in the advanced lactation stage proved to be more sensitive to heat load than cows in early lactation. Moreover, multiparous cows showed less pronounced activity responses than primiparous ones. In fact, heat load accumulation and individual cow-related factors proved to be significant factors for prediction models based on the individual susceptibility of animals to heat stress (Lovarelli et al., 2020c) (Tullo et al., 2019).
Applied statistical methods used in the literature (Piwczyński et al., 2020) showed that milking frequency, lactation number (parity number), month of milking, and type of lying stall represent important factors responsible for the monthly milk yield of dairy cows in farms with AMSs.
In this context, Machine Learning (ML) algorithms have been already applied in some areas of dairy research, particularly to predict data, and they represent a promising tool, useful to develop and improve decision support for farmers (Cockburn, 2020) in order to increase both

milk yield and animal welfare and, on the other hand, to reduce the resources needed, hence increasing the sustainability of the sector (Strpić et al., 2020), (Lovarelli et al., 2020a).
Further studies are thus necessary to identify how factors related to animal welfare and cow performance can be combined with indoor conditions inside the barn.
To this end, continuous and real-time monitoring of the animals and the environmental parameters of the barn contributes to the knowledge of the welfare conditions of the individual cows: it can provide important information for the management of the barn environment (Bovo et al., 2020) and for the prevention of problems related to the longevity of the cows, their productivity, and the quality of the milk.

Chapter 2

Big Data approaches and methods

Nowadays, information flows are routed through huge amounts of data that are continuously collected, transmitted and stored thanks to increasingly advanced technologies. The unit of information contained in each piece of data actually contains hidden global information, waiting to be discovered.
Big Data analysis aims to explore global information, identify links between data groups, make predictions and classifications, zooming out with respect to the initial look.
In this chapter some of the most important analysis approaches will be mentioned, ranging from the simplest to the most advanced, still evolving.

2.1 Big Data Approaches

When it comes to data, and especially Big Data, it is important to introduce a series of operations that prove to be necessary for the collection, management, storage and analysis of information.
First, **Data Generation** refers to the process of generating data. The tools with which this phase is carried out can be innumerable depending on the type of data. Think for example of sensors, cameras, video cameras, sound recorders: each of these produces data of a different nature.
After creation, a selection and pre-processing operation is required, often called **Data Acquisition**.
The data collection phase ends with the storage of data, called **Data Storage**.
Finally, the term **Data Analytics** indicates the process of qualitative and qualitative data analysis. It includes different sub-processes:

- *Data transformation*: After Gathering, Selection and pre-processing

data, transforming preprocessed data into data-mining-capable format is required.

- *Data analysis*: After transforming data, analysis can be done using various statistical methods and data mining algorithms such as regression, classification, clustering ...
- *Data visualization*, divided in:
  - *Evaluation*: Measure the results of data analysis;
  - *Interpretation*: displaying the output of data analysis by an interactive way

2.2 Big Data: why so "big"?

The term Big Data is one of the most present in scientific production today, and is often used improperly to indicate large amounts of data. The adjective "Big" is to be related to a broader characterization of the data, in fact it refers not only to the quantity, but also to their other characteristics. Big data requires a revolutionary step forward from traditional data analysis, characterized by its three main components: *variety*, *velocity* and *volume* (Shobana and Kumar, 2015).
Variety is related to the inhomogeneity and diversity of information that can be gathered from the data. For example, it is possible to think of the multidimensionality of a dataset made up of different features.
Velocity, on the other hand, is a term that refers to the speed with which a data set can be expanded, for example with systems that acquire and transfer new information live.
Finally, **Volume** is the term that probably comes closest to the most common concept of Big Data, as it is superficially received. In fact, it refers to the quantity: exabytes were generated each day in 2012. This amount is doubling every 40 months approximately (Oussous et al., 2018).

2.3 Exploratory Data Analysis

In statistics, exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what

the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.
Tecniques usually used in EDA aims to show the highlights of data. The most common graphical tecniques are:

- Box plot
- Histogram
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot (2D/3D)
- Stem-and-leaf plot
- Parallel coordinates
- Odds ratio
- Targeted projection pursuit
- Heat map
- Bar chart

and for the dimensionality reduction:

- Multidimensional scaling
- Principal component analysis (PCA)
- Multilinear PCA
- Nonlinear dimensionality reduction (NLDR)

2.4 Machine and Deep Learning algorithms

Artificial intelligence (A.I.) is one of the most commonly used expressions in the panorama of new technologies. Informally, the term "artificial intelligence" is applied when a machine is able to perform functions that humans associate with other human minds, such as "learning" and "problem solving" (Shinde and Shah, 2018).
Over time, computer scientists have focused their attention on solving increasingly specific problems and to do this they have thought about developing computational machines that are capable of learning and then independently reproducing choices, judgments and forecasts.
Machine learning is a sector of A.I. which utilizes algorithms whose working mechanism is completely clear and understandable from the human point of view.
The accuracy of these machines has grown over time, to the point that their use has become increasingly frequent even in areas that require particular attention, precision and a guarantee of success.
Among the most modern examples of applications we find Computer Vision, which is widely used for the recognition of objects, faces and fingerprints, classification tasks, useful in making predictions and analyzing images.
One of the most engaging features of computational machines is their ability to learn from examples, and in other cases to use statistical information to analyze the case and make decisions on their own. These characteristics are fundamental for the use of technologies in modern devices that we use every day such as PCs, tablets, smartphones. A particular type of algorithms, which have come back into vogue in recent years, is also increasingly present: these are neural networks. These algorithms were born with the idea of imitating the mechanism of transmission of impulses between the neuronal cells of our brain and of perfecting synaptic connections thanks to training.
Artificial neural networks (ANNs) are essentially made up of 3 blocks: input features, some hidden layers and an output which can be a number or a class. When the number of hidden layers is very high, these neural networks are called "deep neural networks". They have the advantage of solving modeling the weights in their hidden layers in such a way as to be able to apply nonlinear transformations in multidimensional spaces. Neural Networks are used extensively in forecasting of weather and climatic change which is helpful in human safety and security of properties

such as buildings, environment, installation, houses, and transportation (Abiodun et al., 2018).

2.5 Robust statistics approaches

Robust statistics seek to provide methods that emulate popular statistical methods, but which are not unduly affected by outliers or other small departures from model assumptions. In statistics, classical estimation methods rely heavily on assumptions which are often not met in practice. In particular, it is often assumed that the data errors are normally distributed, at least approximately, or that the central limit theorem can be relied on to produce normally distributed estimates. Unfortunately, when there are outliers in the data, classical estimators often have very poor performance, when judged using the breakdown point and the influence function, described below.
The practical effect of problems seen in the influence function can be studied empirically by examining the sampling distribution of proposed estimators under a mixture model, where one mixes in a small amount (1–5% is often sufficient) of contamination. For instance, one may use a mixture of 95% a normal distribution, and 5% a normal distribution with the same mean but significantly higher standard deviation (representing outliers).

2.6 Python Libraries for data analysis

The raw data obtained from the milking robot used in this analysis required extensive manipulation and reorganization.
Milking robots collect different data:

- 'animal_id' (identification number of the animal)
- 'datetime' (date of record)
- 'robot_id'(identification number of the robot)
- 'milking_yield' (milking yield of the cow)
- 'my_expected' (expected Milking yield)
- 'n_milking' (number of milking event)
- 'milking_interval' (milking interval)

- 'milking_speed' (milking speed)
- 'max_milking_speed' (maximum milking speed)
- 'dim' (day in milking),
- 'box_time' (time spent in the box)
- 'treatment_time' (treatment time),
- 'milk_temp' (milk temperature),
- 'weight' (weight)
- 'milking_time_as' (milking time front left breast),
- 'milking_time_ad' (milking time front right breast)
- 'milking_time_ps' (milking time rear left breast)
- 'milking_time_pd' (milking time rear right breast)
- 'dead_time_as' (dead time),
- 'dead_time_ad' (dead time front right)
- 'dead_time_ps' (dead time rear left)
- 'dead_time_pd' (dead time rear right)
- 'as_cond' (front left breast milk conductivity)
- 'as_color' (front left breast milk color)
- 'ad_cond' (front right breast milk conductivity)
- 'ad_color' (front right breast milk color)
- 'ps_cond' (rear left breast milk conductivity)
- 'ps_color' (rear left breast milk color)
- 'pd_cond' (rear right breast milk conductivity)
- 'tot_intake' (total food intake)
- 'label' (milk infected or not)

Since the main goal of the study is the prediction of milk yield anomalies, for the analysis carried out in this thesis, we will use the milk yield as main features among the ones collected by the robot.
The data transmitted by the temperature and humidity sensors and those directly collected by the milking robots are written in files of the ".diff" type, which have been read as ".csv" in which the separator was ";".
A library called *pandas* allows you to read, view and manipulate large dataframes, leaving ample space for managing missing values and filling them.
Scikit-learn (Pedregosa et al., 2011) features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries *NumPy* and *SciPy*.

Chapter 3

Results and applications

In this section of the thesis some applications of some of the analysis methodologies identified in chapter 1 will be described.
In particular, in the first part, the results obtained in the exploratory analysis of the temperature data will be explained. The analysis aims to represent, describe and characterize the temperature measured inside and outside an example barn, in order to highlight the relation between the two measures.
In the second part, an application of the Random Forest algorithm to milk production data in a barn equipped with milking robots is first described, and subsequently the results of this study are compared to the results obtained with the application of the algorithm to German data.
Finally, in the third and final part, a robust statistical method is used to identify anomalies in the lactation curve.

3.1 Exploratory Data Analysis results

3.1.1 Description of Data

Temperature data analyzed in this subchapter were collected from sensors of an experimental barn used as example.
The internal temperature was measured by the thermo-hygrometer data logger PCE-HT71, the external one is obtained by an external meteo station, placed very near to the barn.
The time range used for this analysis is from july 2021 to february 2022.

3.1.2 Study of temperature time series

One of the first operation was the visual representation of the series. In Figure 3.1 the blue line is the external temperature of the barn; the orange line is the internal temperature.

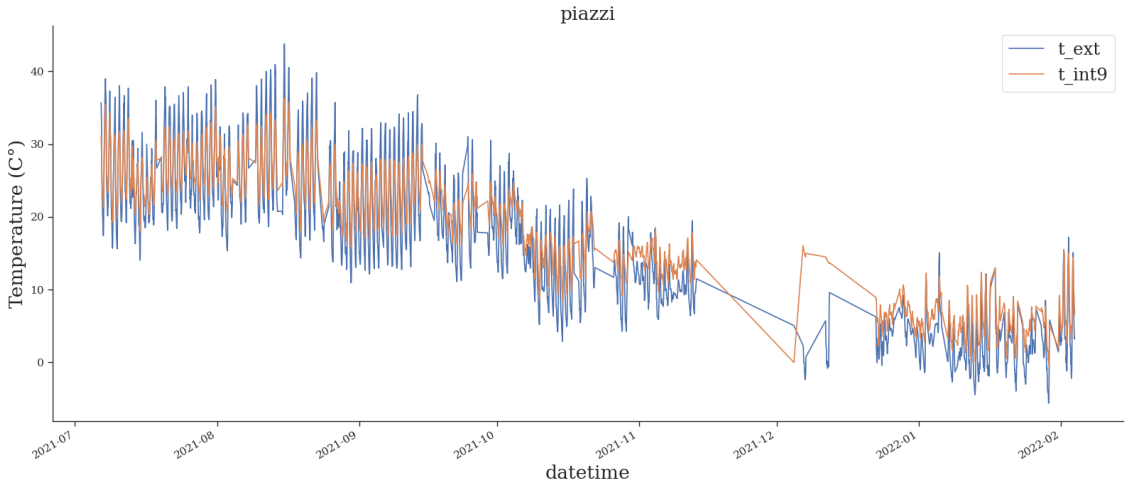

Figure 3.1: Plot of temperature recorded inside and outside the barn.

The simplest analysis that can be performed on the two time series is a Pearson correlation. The coefficient, for the experimental barn analyzed and for the period available is 0.967.

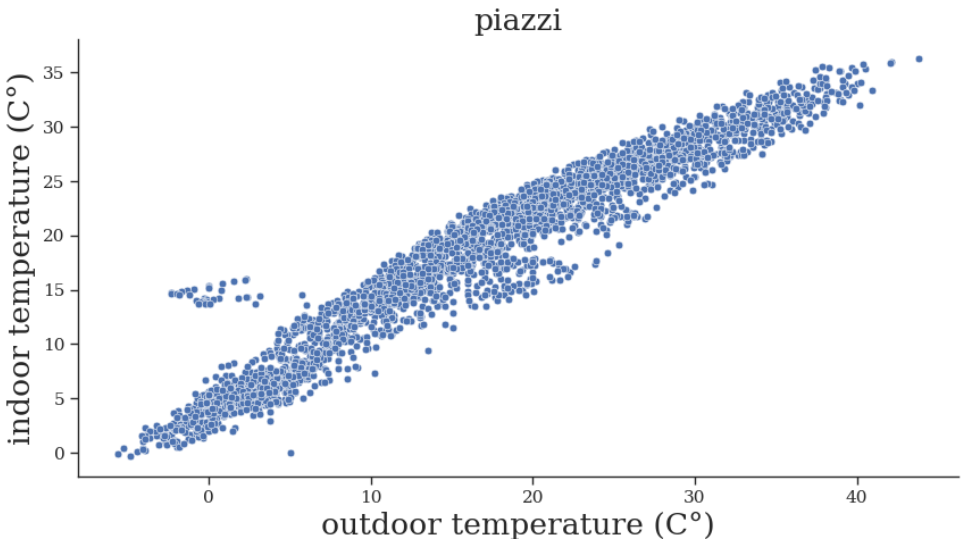

Figure 3.2: Scatterplot of

Figure 3.2 is the scatterplot of the indoor vs the outdoor temperature. In Figure 3.3 a zoom of the plot in figure 3.1 in the range august - september 2021 is shown.
Looking at this portion of the time series we can notice that the descending portions of the internal temperature time series decrease less quickly than the corresponding descending portions of the external temperature. The same is not observable in the ascending portions, at least in such an evident way.

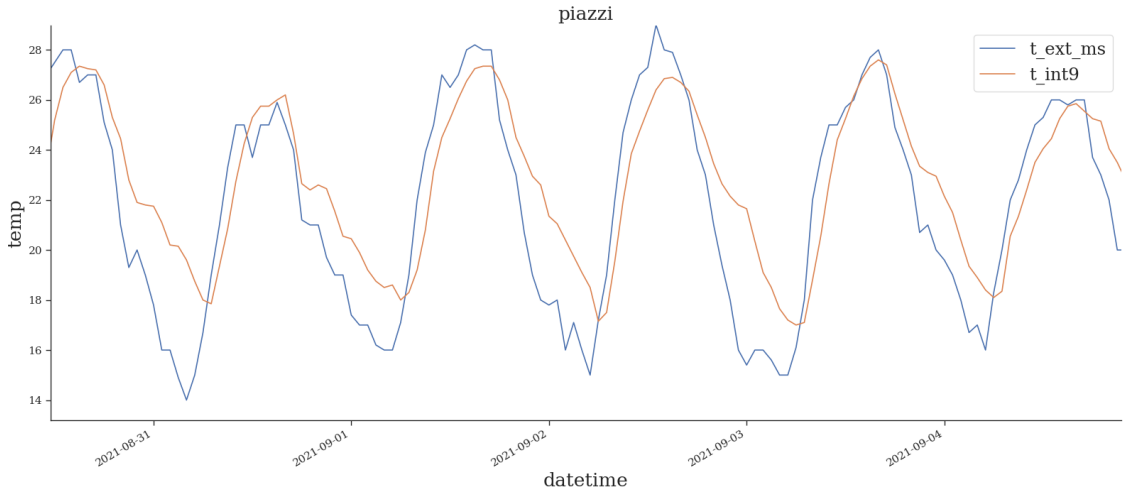

Figure 3.3: Plot of temperature recorded inside and outside the barn (zoom).

3.1.3 Cooling coefficient

The remark made at the end of the paragraph 3.1.2 led to think of modeling the descending traits of the time series.
The equation assumed as model was the exponential:

$$y = m \cdot e^{-\frac{x}{\tau}} + b \tag{3.1}$$

where $m, \tau, b \in \mathbb{R}$.

The rate of descent of the exponential depends on the alpha coefficient, which, taking into account the type of quantity taken in analysis, will in this case be called **cooling coefficient**.

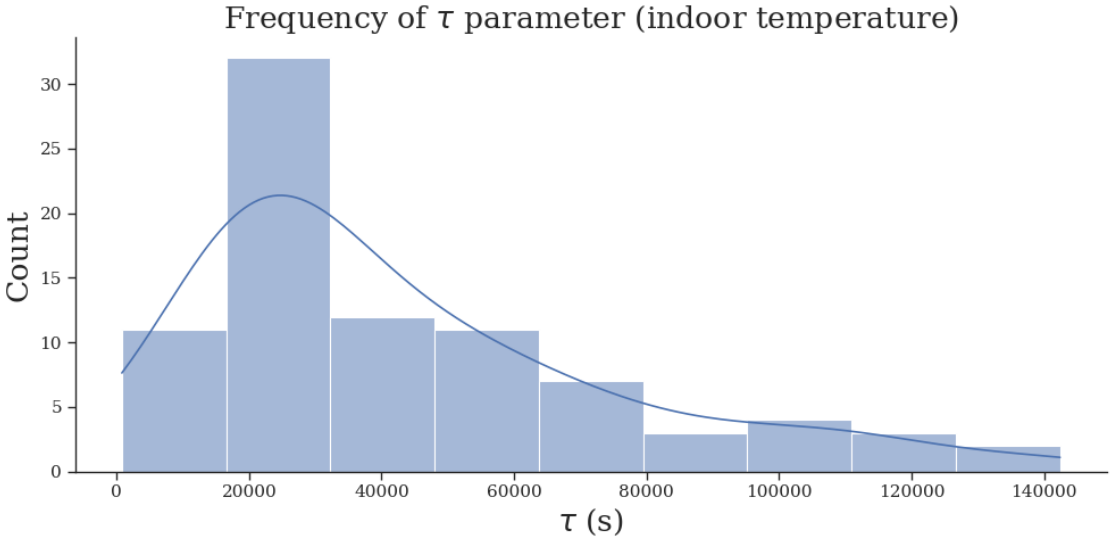

Figure 3.4: Histogram and density plot of the tau values (indoor temperature).

Since each descending portion can be fitted with the same equation, obtaining several different values of τ, the histogram in Figure 3.4 can be useful to understand what is the most frequent value.
In our case the most frequent value is around 25000s, so that it means that, according to the classical definition of τ, in about 7 hours the barn loses the 63% of the medium temperature delta observable in a day.
The τ value can be considered as representative of the barn used as example because it depends on its shape.
The *cooling coefficient* assumes an important role in the microclimatic data analysis because it can be used to compare different barn and to modelize the indoor thermal behaviour looking at the external one.
As confirmation of what has just been deduced, a similar analysis was carried out on the time series of the outdoor temperature. Looking at figure 3.5, representing the distribution of τ values for the outdoor time series, i is visually clear that the most frequent value are around 14000s, equivalent to about 4 hours.
This results shows how, for the external sensor, the delay in the cooling process is shorter.

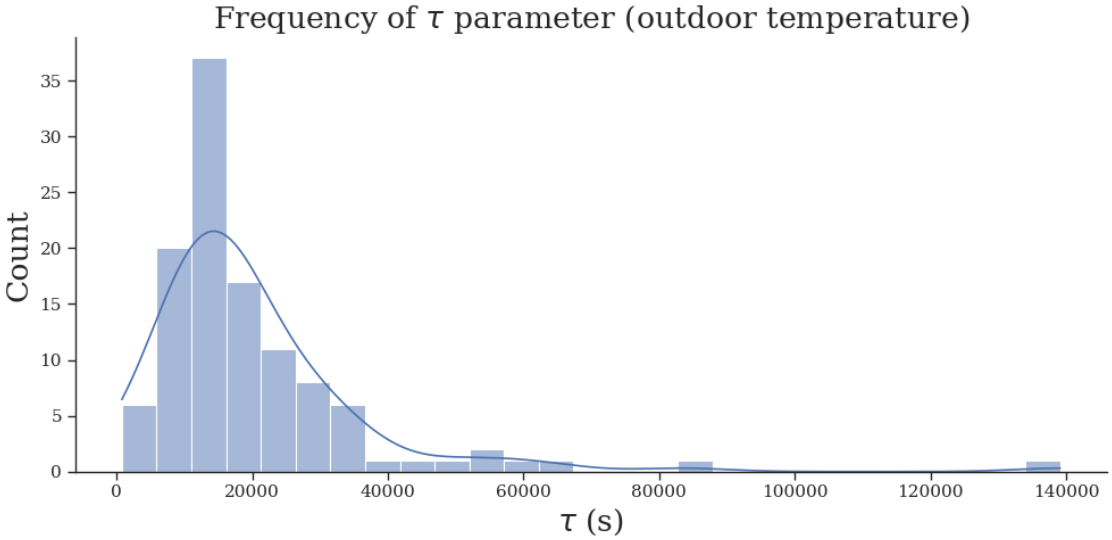

Figure 3.5: Histogram and density plot of the tau values (outdoor temperature).

3.2 Random Forest regressor for milk yield predictions

3.2.1 Description of Data

3.2.1.1 Housing and animals

Data used for this study were collected by Italian and German barns. For the Italian case, as reported in (Bovo et al., 2021) the data collection and the validation of the model have been carried out with reference to a case study dairy farm located in the municipality of Budrio, about 15 km NE of Bologna (Italy). The region is characterized by hot summer seasons with high percentage of humidity; in fact, considering the warmer months of the year (i.e., June, July and August), the average of the daily maximum temperature typically ranges from 27 to 29 °C, with daily average relative humidity, for the same period, from 75% to 85%. The rectangular layout of the barn is 51 m long and 23 m wide, with the longitudinal axis SW–NE-oriented, a ridge height of 8.52 m and gutter heights of 4.95 m on the NW side and 6.65 on the SE side. It consists of a hay storage area on the SE side, a resting area in the central zone of the building and a feeding area with feed delivery lane on the NW side (see Figure 3.6). The resting area has a partially slatted floor and hosts 78 cubicles with straw bedding. Two blocks of head-to-head rows are in the central part of the resting area, while another row runs along

the entire length of the barn close to the storage area. Mechanical ventilation is controlled by three high volume low speed (HVLS) fans with five horizontal blades which were activated by a temperature-humidity (TH) sensor situated in the middle of the barn at about 3 m of level. Lactating cows are fed with a total mixed ratio kept available along the feeding lane. About 65 Friesian cows are milked everyday using an AMS "Astronaut A3 Next" (Lely, Maassluis, The Netherlands) placed at the SW extremity of the barn.
During the period of the study, the robot was programmed to ensure a particular number of daily visits for each cow depending on her productivity and her expected optimal milk yield per visit, with a minimum of two and a maximum of four daily visits as constraints.
Animals with fewer than two visits in one day were signaled by a warning, while the cows which have been milked four times in one day can only pass through the AMS box without being milked and fed further.
The milk room is located on the SW side of the building, next to the offices and the technical rooms. The robot also manages the supplement feeding, which is calculated based on daily milk yield and days in milk (DIM) value: it linearly increases with time from 3.0 kg to 3.5 kg for cow during the first 15 DIM, then it is proportional to milk yield with a coefficient 0.157 kg/L up to the limit of 7.50 kg. Finally, during the last 14 days of the milking period, the supplementation decreases linearly with time to the lower limit of 1.5 kg.

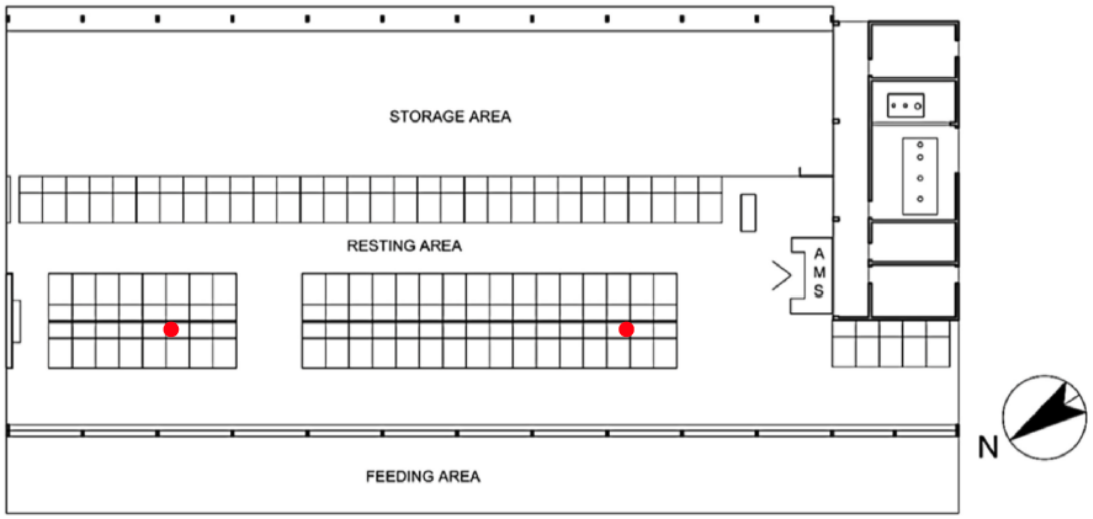

Figure 3.6: Scheme of the experimental barn (AMS: Automatic Milking System; the red dots represents the position of the two thermo-hygrometer data loggers).

3.2.1.2 Milk Yield and Enviromental Data

The period of study spans the years of 2016 and 2017. In this period, 132 cows were milked by the AMS system, although 91 animals were considered in the study—i.e., the cows with more than 100 daily milk production values. All of this ensured a robust dataset for each cow necessary to perform a reliable training of the numerical model described in the following. Among the 91 cows, at the beginning of the study, 41 were single-parity and 50 were multiparity. The data of the various milking events recorded by the AMS were downloaded, together with the cow tags and the DIM in a large dataset. Then, the daily milk yields were calculated for each cow. The dataset was then filtered by eliminating the exceptional events (e.g., daily milk yields of cows with mastitis or other factors that can influence animal production). This allowed us to create a cleaned dataset for each cow, collecting the time series of the milk yields during the monitored period. The cow datasets considered in the study range from 100 to about 550 milk daily yields. As far as the recording of environmental data is concerned, two thermo-hygrometer data loggers, PCE-HT71, with an accuracy of ±3% on the relative humidity (RH) and ±1 °C on the temperature T, were positioned inside the barn (see Figure 1) and recorded the indoor temperature T_{in} and relative humidity RH_{in} from 1 January 2016 to 31 December 2017. Outdoor thermo-hygrometric parameters were measured, for the same period, by a weather station located in the proximity of the building. The thermo-hygrometric data loggers recorded temperature and humidity at 30 min intervals and for each couple of values the THI was calculated following Equation 3.2, described by the National Research Council (Rowsell, 1972):

$$THI = [(1.8 \cdot T_{db} + 32) - (0.55 - (0.0055 \cdot RH) \cdot (1.8 \cdot T_{db} - 26)] \quad (3.2)$$

where T_{db} is the dry bulb temperature (T_{db} in °C) and RH is the relative humidity (RH in %). Then, the daily average THI was calculated for the two thermo-hygrometer sensors. The values of the two showed a maximum difference of only 0.8, confirming the environmental homogeneity in the barn. In the study, the mean values of THI obtained by the two thermo-hygrometers were considered. They are showed in Figure 3.7.

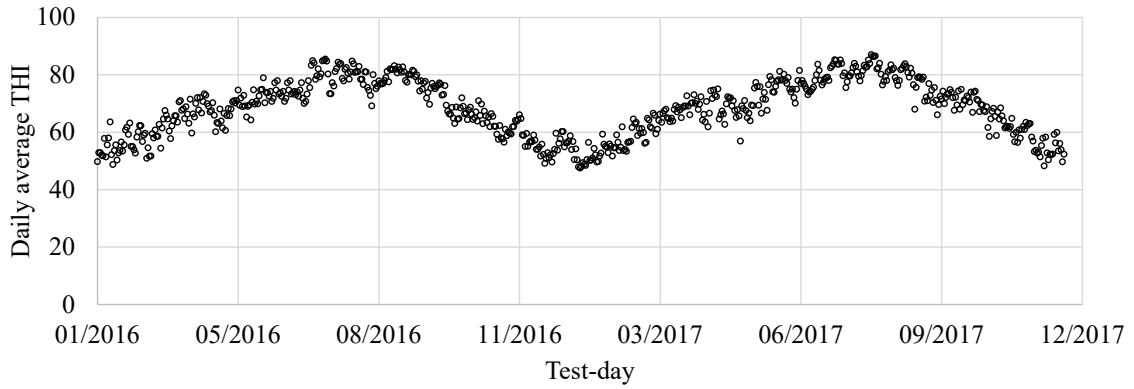

Figure 3.7: Barn indoor THI values calculated for the years 2016–2017 considered in the study.

3.2.1.3 Statistical Model

The general statistical model used to determine the effect of environmental conditions on milk yield at the single animal level has the following general form:

$$y_{i,j} = f(DIM_{i,j}, THI_{i_j}, THI_{i,j-1}, THI_{i,j-2}, THI_{i,j-3}, THI_{i,j-4}, THI_{i,j-5}) + e_{i,j} \quad (3.3)$$

where $y_{i,j}$ is a test-day milk yield for cow i at day j; $DIM_{i,j}$ denotes the effect on milk yield of the DIM of cow i at day j; THI i,j is the effect on milk yield for cow i of the daily average THI at day j; $THI_{i,j-1}$ — $THI_{i,j-5}$, respectively, represent the effect on milk yield for cow i of the daily average THI at day from j-1 to j-5; $e_{i,j}$ represents the random residual effect, a priori assumed to be independently and identically distributed as N(0,s_e^2), where s_e^2 is the residual variance. In particular, several statistical models have been tested also considering a longer period, starting from 10 days prior to testing. Then, it was gradually reduced to 5, removing one day at a time with the value of the average relative error that remained almost unchanged (modifications lower than about 0.1%). Only with the removal of the THI value of the fifth day prior to testing did the average error increase significantly, thus leading to the decision to consider a preceding period of 5 days. In order to predict the heat stress effects at the level of a single cow, seven different features (i.e., predictors) have been used as input data to the Random Forest algorithm, better detailed in the following section, and the dataset of each animal has been divided into data for the training phase and data for the testing phase.

3.2.2 Random Forest algorithm

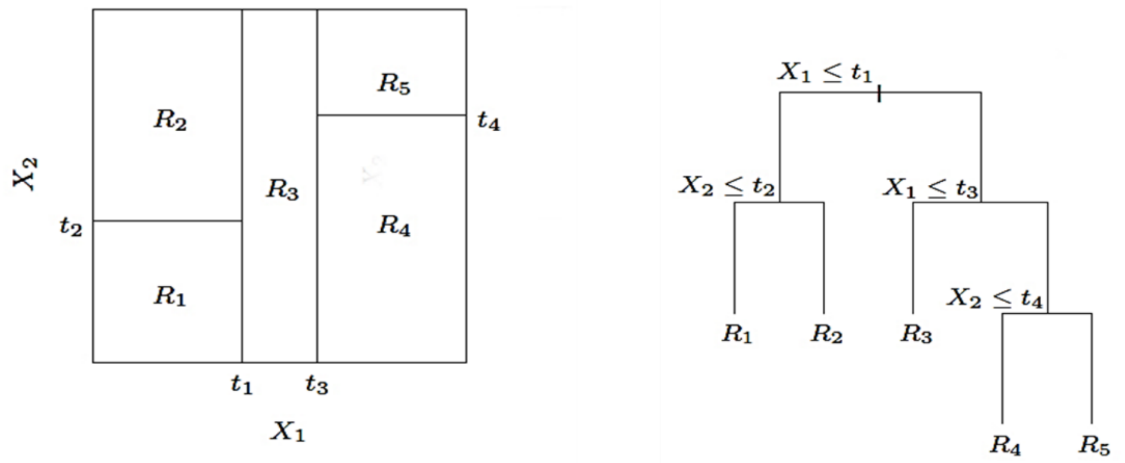

Figure 3.8: Partitions and CART. Left scheme shows a partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data. X_1 and X_2 are sample features, R_i is the i-th region of the features' space. The panel on the right shows the tree corresponding to the partition. The variable t is a generic parameter.

The regression analysis of the collected data was performed by using the Random Forest algorithm (Breiman, 2001), an ensemble learning method that makes predictions by averaging over the predictions provided by several independent random models (Denil et al., 2014). The algorithm (see Figure 3.8) was originally conceived as a method of combining several classification and regression trees (CARTs) (Breiman et al., 1984) using bagging (Breiman, 1996), and as the name suggests, it is a tree-based ensemble with each tree depending on a collection of random variables. Random decision trees have found widespread applications thanks to several features, such as the ability to capture interactions between predictors, to deal well with irrelevant predictors, being robust in terms of outliers in the predictors and well scalable for large sample sizes (Hastie et al., 2009). In the present work, the algorithm was adopted for regression purposes by using the Scikit-Learn Python library (Python Software Foundation, 2020) in order to establish the random forest model (RFM) best fitting the data values of each cow. A key advantage of the recursive binary tree is its interpretability. The feature space partition is fully described by a single tree. With more

than two inputs, partitions such as that in the top scheme of Figure 3.8 are difficult to draw, but the binary tree representation works in the same way (Hastie et al., 2009). Furthermore, the Random Forest algorithm can provide useful indications on the most important predictors between those included in the training dataset. On the other hand, Random Forests are frequently used as "black box" models, as they generate reasonable predictions across a wide range of data even if they sacrifice the intrinsic interpretability present in the decision trees.

The Random Forest method is based, as for most of the data-driven methods, on the minimization of a function. Then, for a random vector X containing the values of the independent variable (i.e., the regressor or predictor) and a random vector Y collecting the values of the dependent variable (i.e., response), it is possible to assume an unknown joint distribution $P_{XY}(X, Y)$. The goal is to find a predictor function $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ to be minimized. Intuitively, $L(Y, f(X))$ measures the distance between vectors $f(X)$ and Y, and it should penalize values of $f(X)$ distant from Y. For regression purposes, a typical choice of L is the squared error loss function:

$$L(Y, f(X)) = [Y - f(X)]^2 \tag{3.4}$$

While L is usually a binary function (is a zero-one function in this work) for classification applications:

$$L(Y, f(X)) = \begin{cases} 0 & \text{if } Y = f(X) \\ 1 & \text{otherwise} \end{cases} \tag{3.5}$$

From the minimization of the loss function, the collection of the n base learners $b = [h_1(X), \ldots, h_n(X)]$ are identified. Then, they can be combined to provide the so-called ensemble predictor $f(X)$:

$$f(X) = \frac{1}{n} \sum_{j=1}^{N} h_j(X) \tag{3.6}$$

Providing the best approximation of Y (Cutler et al., 2012)

A significant advantage of the RFMs is the possibility of assigning a score to each individual feature composing the input of the statistical model. The scores are representative of the importance of the different

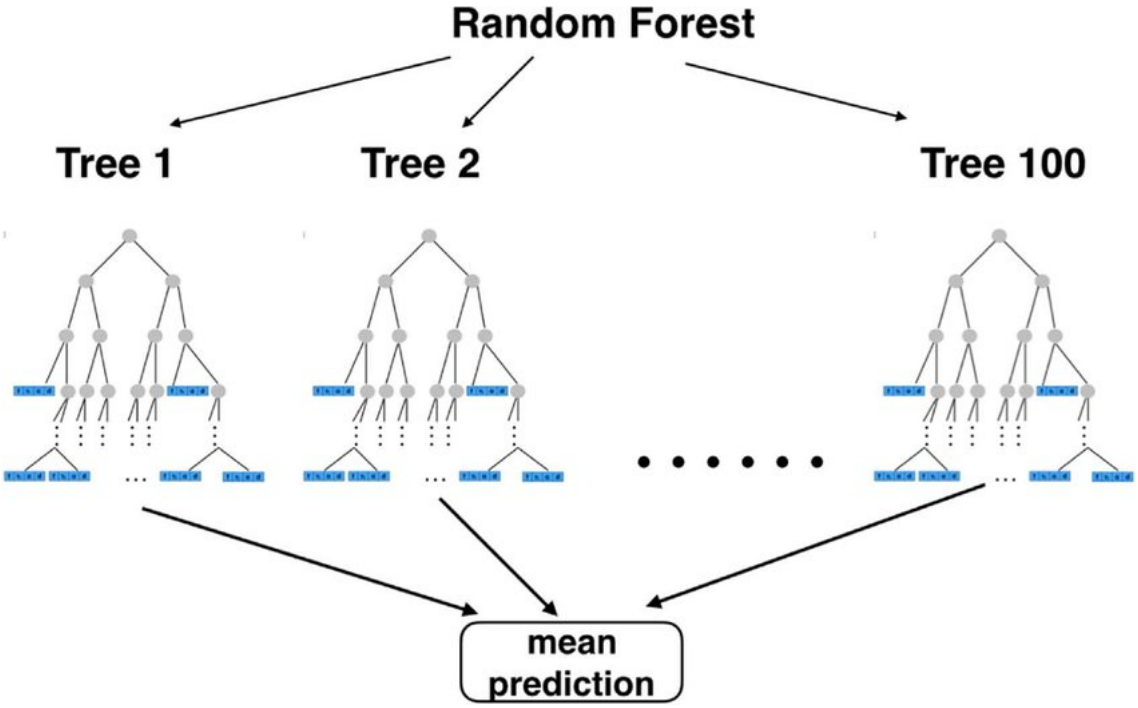

Figure 3.9: Scheme of Random Forest working principle.

features in the model output (i.e., the prediction).
A function of the scikit-learn library allows users to produce the ranking of the features and the evaluation of various scores. Two of the most important parameters for the application of RFMs are the size of each tree (i.e., number of nodes) and the number of trees adopted. If the parameters are too large, overfitting problems could appear, while if the values are too small for the complexity of the data, the model is not able to converge to a suitable solution.
In this work, for the first parameter, a self-expanded criterion, it was assumed that the nodes number expand by itself when the number of samples is bigger than 2. Instead, the number of trees has been set equal to 1000. The dataset of each selected cow was divided in two portions: one used for the training phase and the other for the validation, and a specific RFM was obtained for each animal.
More details about the training/test division are provided in the following subsections. The RFM has been developed for the assessment of the daily yield (the dependent variable) starting from the values of the independent variables.

3.2.3 THI-based Random Forest regressor

The RF model can be applied not only as classifier, but also as a regression tool or as predictive tool. The methodology proposed here can be applied for both purposes by considering three different numerical scenarios. In the statistical model, having seven predictor features, the daily milk yield is evaluated as a function of the position of the day in the lactation curve and the indoor barn conditions expressed in terms of daily average of the temperature-humidity index (THI) in the same day and its value in each of the five previous days, recognized as a statistically significant period for the production on the day under consideration. In this way, extreme hot conditions inducing heat stress effects can be considered in the yield predictions by the model.
Figure 3.10 shows the flow chart of the THI-based Random Forest regressor, from data collection to the prediction of daily milk yield.

Figure 3.10: Flow chart of the THI-based Random Forest regressor.

3.2.4 Training and Test configuration

The model has been calibrated and tested on the data collected on 91 lactating cows of a dairy farm, located in northern Italy and equipped with an AMS and two thermo-hygrometric sensors acquiring information on the environmental conditions, during two entire years—i.e., 2016 and 2017. To validate and test the forecasting potentials of the method, as well as to quantify its reliability, three different numerical scenarios, i.e.,

A, B, and C, have been considered. Scenario A has the objective to test the model for regression purposes, while B and C aim to evaluate the reliability of the model in providing the time series trend of future milk yields.

3.2.4.1 Scenario A

Scenario A has been used to train and test the RFMs for regression purposes. In this scenario, the dataset of each cow was sampled with a cross-validation procedure, a resampling procedure evaluating machine learning models on a limited data sample. In particular, the k-fold cross-validation procedure (Ng, 1998) was considered by adopting a k value equal to 20 (so adopting a 20-fold cross-validation procedure). In this procedure, the dataset was divided into 20 equal parts (i.e., groups) and the training/testing process ran 20 times each time with a group used as test, the holdout group, and the others 19 groups used to train the model. In this scenario, the train and test values are randomly selected by the extraction algorithm. The accuracy of each prediction was used to evaluate the performance of the model. The scheme of the 20-fold cross-validation procedure is shown in Figure 3.11.

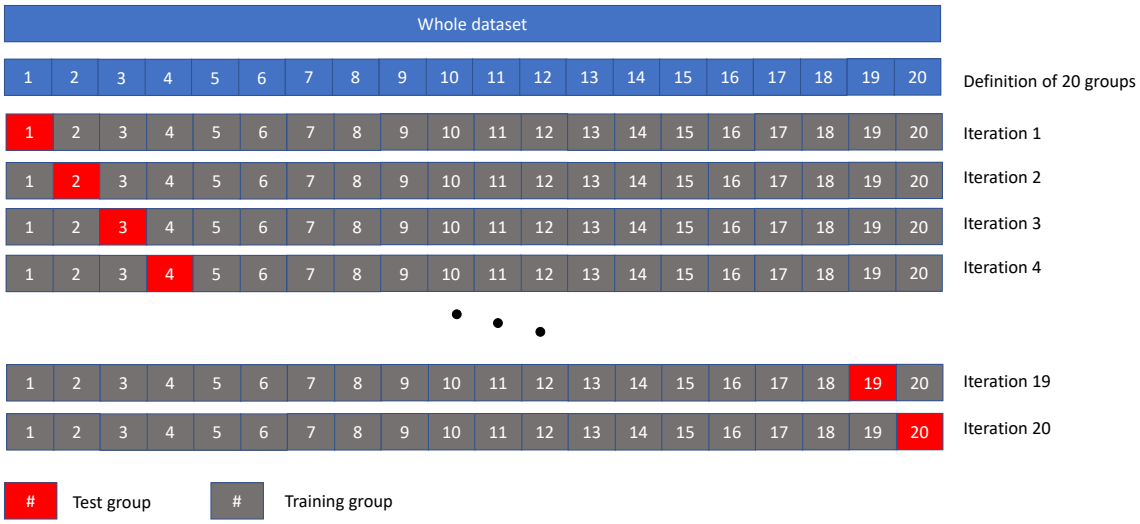

Figure 3.11: Scheme of the 20-fold cross-validation procedure used in scenario A.

3.2.4.2 Scenario B

Scenario B has been adopted with the objective to train and test an RFM for the assessment of continuous time series values by considering the need to apply the model for the assessment of future productive trends of cows under different climatic conditions. In this scenario, the dataset of each cow was divided into two groups: the initial 80% of the data were used for training while the last 20% were used to test the model accuracy and reliability (see Figure 3.12). In this case, for each cow, a continuous series of daily milk yields was obtained from the model and compared to the real one.

Figure 3.12: Scheme of the scenario B.

3.2.4.3 Scenario C

Scenario C was obtained, starting from the scenario B, under the hypothesis that during the time, new available data are added in the training phase to improve the predictive capability of the model. This scenario would simulate the application of a RFM for the prediction of future events in a short period, i.e., 5 days, with time series also taking into account the increase in knowledge of the model that the new available data can provide. Starting from the condition of scenario B, in scenario C the RFM model is trained continuously by introducing one more day and it is adopted for the prediction of the milk yields of 5 days forward (see Figure 3.13).

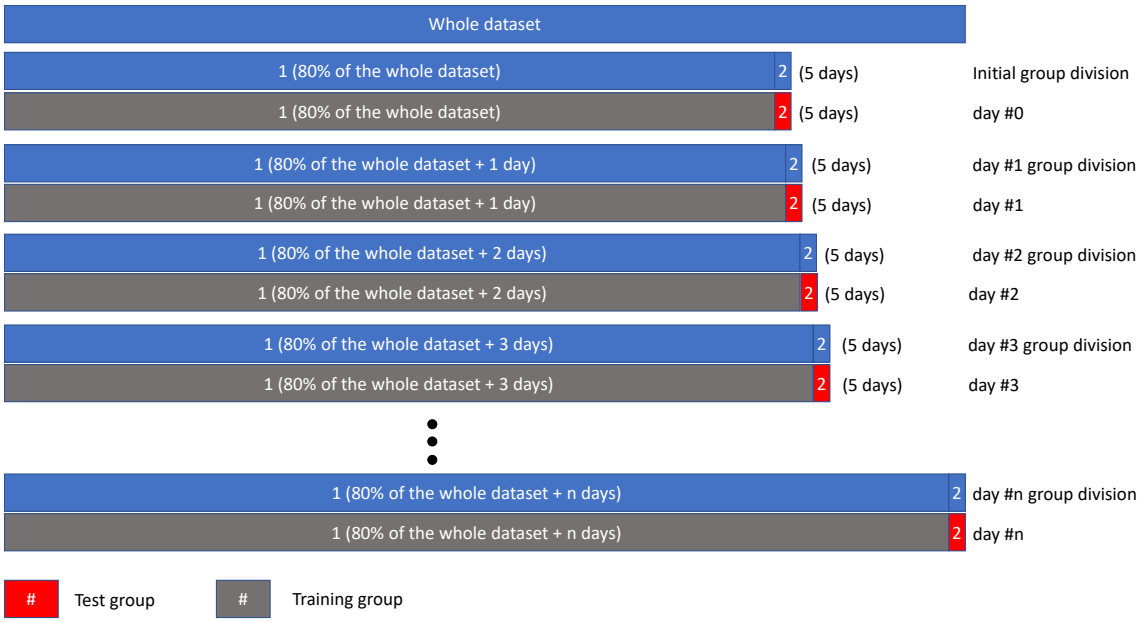

Figure 3.13: Scheme of the scenario C.

3.2.5 General Results

As discussed in the previous section, the statistical model assumed in the paper considers the daily average THI as a representative parameter of the barn thermo-hygrometric conditions. The daily milk yield of a single cow was assessed by RFM including the climatic effects of the actual day (i.e., the "day 0") and those of the past five days (i.e., days -1, -2, -3, -4 and -5). In this way, the model can also consider the heat load duration and the cumulative effects of consecutive days on inducing animal heat stress. A preliminary correlation analysis has been performed with the aim to evaluate the delay between daily yield and climatic conditions. For the 91 cows considered in the study, the Pearson correlation coefficient (PCC) has been established between the milk yield and THI0, THI_{-1}, THI_{-2}, THI_{-3}, THI_{-4} and THI_{-5}. The PCC values are reported in Figure 3.14 for the various days. The values reported are the average on the 91 animals and the average $\pm$ standard deviation (St. Dev.). The trends showed that a weak negative correlation, similar for the different days, exists but is not possible to establish the day with the highest correlation as the different days have similar PCC values. This is because, in the herd, two different cow groups exist. In fact, about 60% of the cows were more sensitive to THI_0 and THI_{-1}, while the other 40% have daily yields more affected by the THI_{-2} to THI_{-5} and, for this group, it is evident that heat stress causes effects with a delay of $3-5$ days.

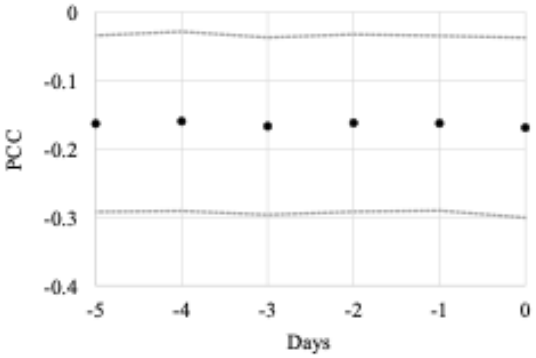

Figure 3.14: Correlation between milk yield and climatic data: Pearson correlation coefficient: average values and average ± St. Dev. values.

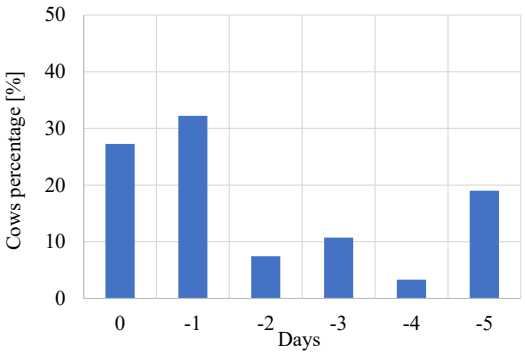

Figure 3.15: Correlation between milk yield and climatic data: percentage distribution of cows vs. day with highest negative effects on milk production (i.e., the day with lower PCC for the single animal)

Figure 3.15 shows the animal percentage vs. day with highest yield decrease. Summarizing, in order to be able to catch the daily milk yield of every single cow, the climatic data up to 5 days before the day of interest have been introduced in the numerical model as independent variables.

3.2.5.1 Goodness of model

Cow level

Firstly, the responses of the RFMs, applied in Scenario A for regression purposes, are reported at the single cow level. In this regard, and for the sake of brevity, an extended description of the results has been reported only for two cows, randomly selected in the herd in order to provide the general validity of the results. The two animals are #226 and #243 (the codes adopted by the farmer have been maintained). They have 360 and 543 test-days, respectively.
To establish the goodness-of-fit of the models, the trends of the relative error Er on the daily milk yield are shown in ascending order in Figure 3.16a, while the prediction accuracies are showed in Figure 3.16b. The minimum and the maximum Er are about 40% to 70% respectively, but most of the daily predictions are characterized by a high accuracy. In fact, for animal #226, about 58% and 28% (i.e., 210 and 100 out of 360 values) of the predictions provided good and very good accuracies,

respectively. Similarly, for animal #243, about 57% and 27% (i.e., 307 and 148 out of 543 values) of the predictions had good and very good accuracies, respectively. For the two cows, the average accuracies (median standard deviation) are 88.64% 14.31% and 87.20% 12.58%. Moreover, from Figure 3.16c, it is evident that the residuals have normal distribution centered on the zero value. In fact, Er for the sum of the daily yields for the test days is very low, i.e., +0.29% and 0.64%, for the two animals described here. This is an important aspect, as it confirms that the trained RFMs can assess, for each single animal, the daily milk production with a good accuracy and with a predicted yield trend, on average, close to the real yield trend.

Analysis of Variability within the Herd

The comparison of the results for a single cow (cow level) can be used to define the variability within the herd, so considering the cow-to-cow variability. In this regard, the median accuracies of the daily yields, of each cow, are showed in Figure 3.17 vs. the data numerousness (i.e., the test-day number of each cow, different from cow to cow). The figure highlights that for the 91 cows considered in the study, having data numerousness higher than 100 days, the median accuracy for the different animals ranges between 63% and 92%.
For the sake of completeness, the figure also depicts the media accuracy of the cows not considered in the work (i.e., cows with less than 100-day dataset numerousness). It is rather clear that a reduced number of events could represent a problematic aspect for the training phase of the RFM and, for this, only 91 animal datasets have been considered robust for the purpose of the work. As far as the cows' datasets bigger than 100 days are concerned, the accuracy values do not increase with the numerousness of dataset, and this leads to the belief that even by enlarging the yield dataset, the average accuracy is not likely to increase significantly. This uncertainty is probably difficult to remove since it can be attributed to the variability in the cow's response, which is not only governed by environmental conditions, but other animal welfare factors could contribute.

Then, in Figure 3.18, the median accuracies ± standard deviation of the 91 cows with datasets bigger than 100 test days are reported in ascending order. The average (out of the cows) median accuracy of the

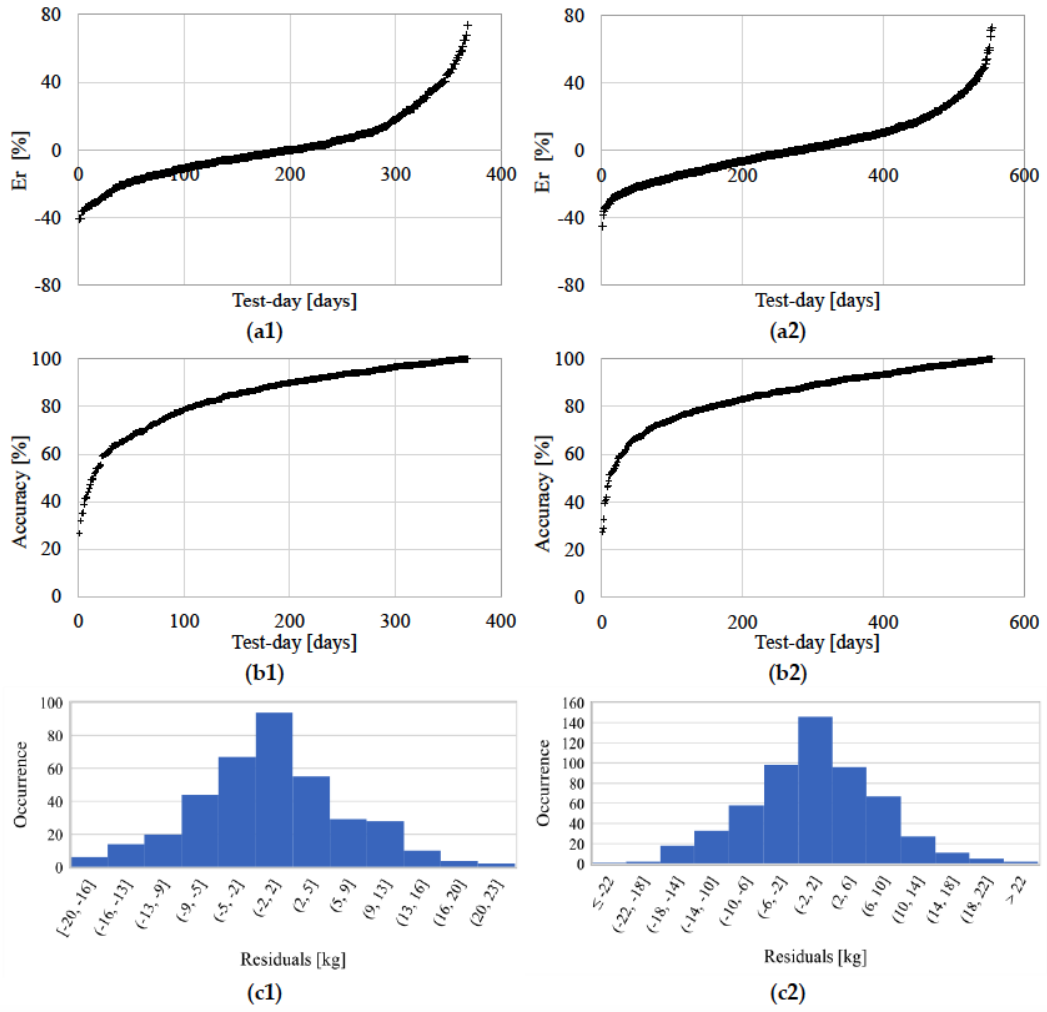

Figure 3.16: Main results obtained for scenario A for two representative cows where (1) represents animal #226 and (2) animal #243. (a) Relative error Er of the daily yield in ascending order; (b) accuracies of the daily yield predictions in ascending order; (c) histogram representation of the residuals of the predictions.

predictions is equal to 79.26%, whereas the standard deviation of the median accuracy is 5.33%. Finally, the analysis of the importance score (IS) of the different features is reported in the following. The IS of the variables represent the key aspect of the RFM since, by means of the IS it is possible to hierarchize the features of the statistical model by attributing different scores to the various independent variables. In Figure 3.19, the boxplot diagram of the different ISs is reported for the dataset containing the 91 investigated cows. Moreover, the most important values of the diagram are summarized in Table 3.1, collecting, for each independent variable (feature), the minimum value, the maximum value, the median value, the standard deviation value, and the coefficient

of variation (CoV) value obtained for the IS. In Table 3.1, as expected, it appears to be clear as the DIM has the highest score, with a median IS = 0.29. Then, THI_0, i.e., the average THI of the day to predict (median IS = 0.13), whereas the other features (THI_{-1}–THI_{-5}) have comparable median ISs, ranging from 0.090 to 0.099. The minimum and maximum values recorded for the different features circumscribe large ranges, confirming the high cow-to-cow variability of the RFMs. The features with the highest median IS values are, at the same time, those associated with highest CoV values.
Lastly, Figures 3.20 - 3.21 report the trends of median and CoV values of each IS disaggregated into single-parity and multiparity cows. From these preliminary results, it seems that the ISs are not dependent, in terms of median values, on the parity number and in general they are quite homogeneous and representative of the whole herd. On the other hand, the presence of cows with different parity numbers may increase the cow-to-cow variability of the IS values of the multiparity group.

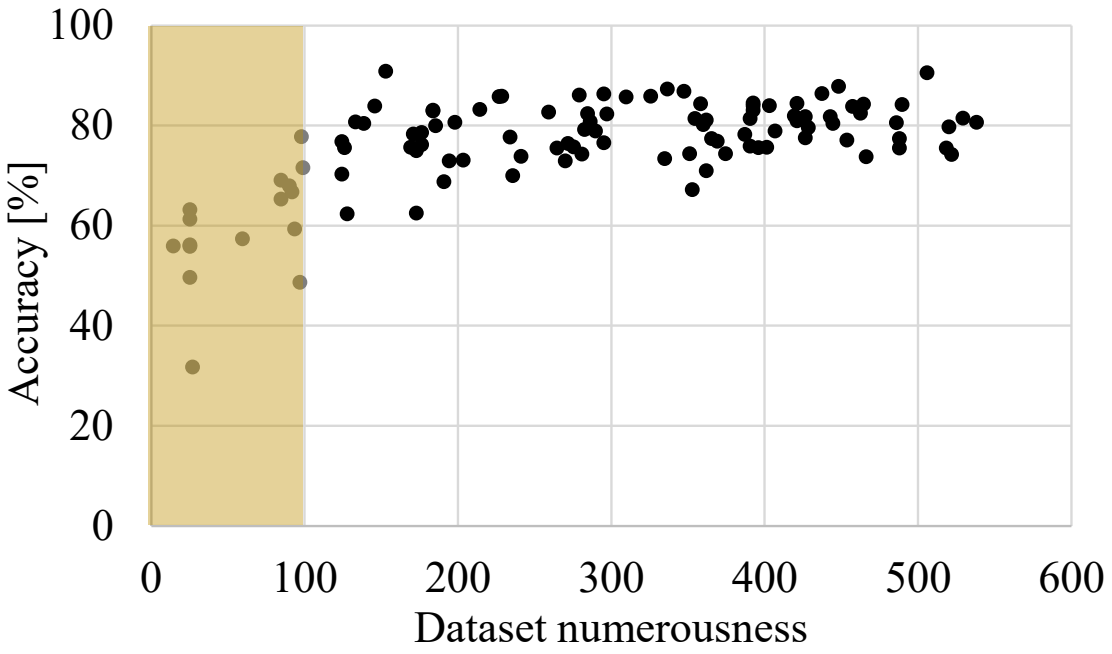

Figure 3.17: Main results obtained for scenario A for the 91 cows of the study: median accuracy for each cow vs. dataset numerousness.

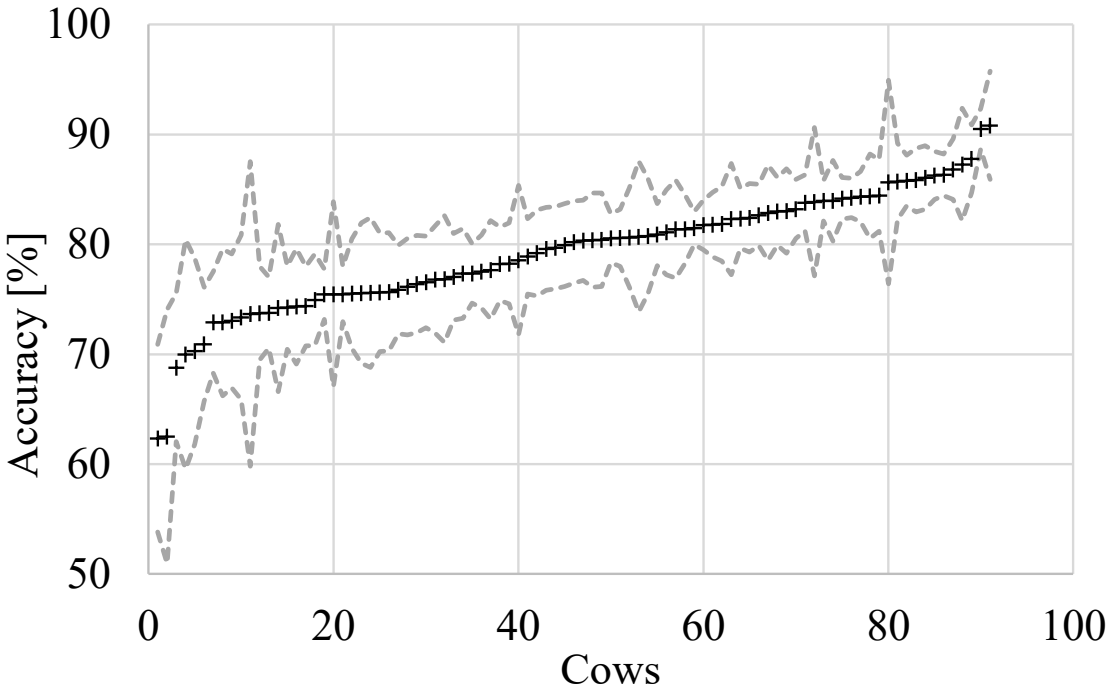

Figure 3.18: Main results obtained for scenario A for the 91 cows of the study: median accuracy ± standard deviation for each cow sorted ascendingly.

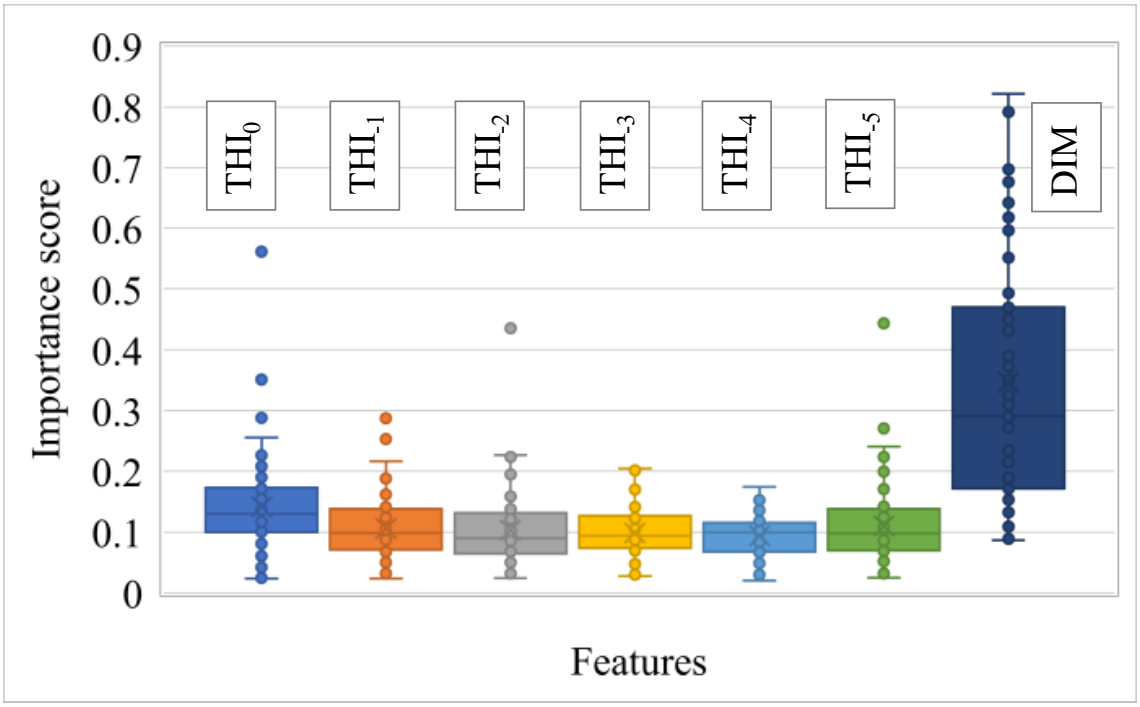

Figure 3.19: Boxplot diagram of the importance score of the different features for the whole dataset in scenario A.

	THI_0	THI_{-1}	THI_{-2}	THI_{-3}	THI_{-4}	THI_{-5}	DIM
Min	0.023	0.024	0.024	0.028	0.020	0.025	0.086
Max	0.561	0.287	0.435	0.204	0.175	0.444	0.821
Median	0.130	0.099	0.090	0.094	0.099	0.098	0.290
St. Dev.	0.075	0.049	0.056	0.039	0.034	0.061	0.197
CoV [%]	57.692	46.483	54.057	39.206	36.103	55.003	56.850

Table 3.1: Minimum value, maximum value, median value, standard deviation value, and coefficient of variation (CoV) value of the ISs obtained for the different independent variables calculated for the whole dataset in scenario A.

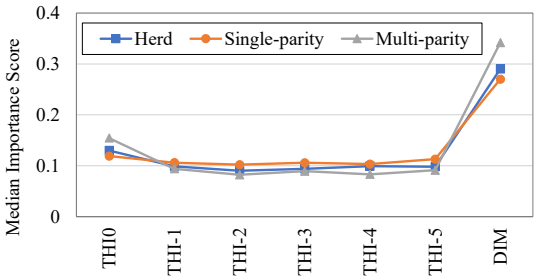

Figure 3.20: Disaggregation of the IS values between single-parity and multiparity cows: Median value of each IS

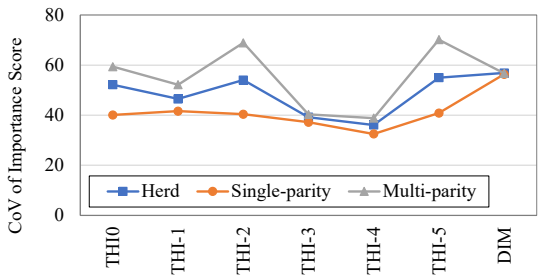

Figure 3.21: Disaggregation of the IS values between single-parity and multiparity cows: CoV value of each IS.

3.2.5.2 Milk Yield Predictions (Scenario B and Scenario C)

Scenario B

In this scenario, the Random Forest model was used to assess future milk yields. The same database as for scenario A has been used, even if with different division between training and testing. As far as the average accuracy related to single cow is concerned, it has very similar results to those obtained for the same animal in scenario A. For the sake of a general comparison, the histogram distribution of the ratio AccB/AccA, i.e., the ratio between the average accuracy obtained in scenarios B and A, for the same cow, is depicted in Figure 3.22. For 59 cows out of 91, i.e., 65% of the analyzed animals, the ratio ranges from 0.9 to 1.1, and for 93% of the cows it ranges from 0.7 to 1.3. Thus, the accuracies of the predictions of scenario A and scenario B appear to be very similar and the RFMs provide comparable precision levels. In scenario B, the average (out of the cows) median accuracy of the predictions is equal

to 81.91% (equivalent to Er = 18%), whereas the standard deviation of the median accuracy is 13.02%. confirming a generally good accuracy, even if it is slightly more scattered than scenario A.

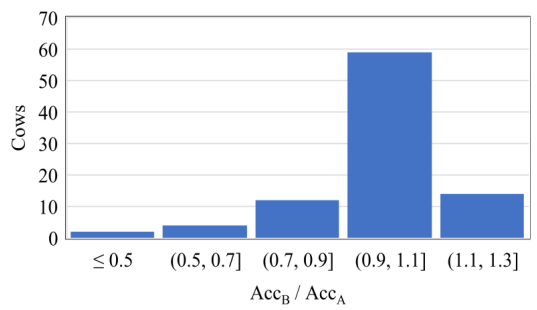

Figure 3.22: Performance indicators for scenario B: Histogram distribution of the ratio between the average accuracy obtained in scenarios B and A for the same cow.

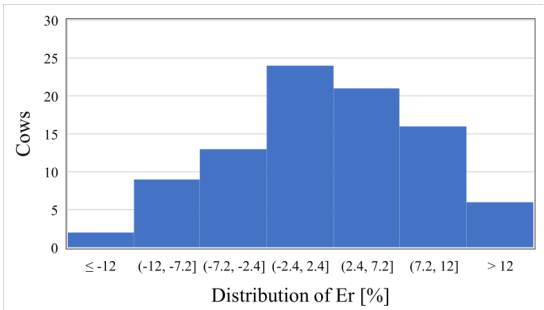

Figure 3.23: Performance indicators for scenario B: distribution of the error Er on the sum of the daily yields over the test days.

As a confirmation of the good accuracy of the models, Figure 3.23 displays the distribution of Er on the sum of the daily yields over the period of tests. For 80% of the animals, the Er value is included in the range ±10%, with an average value of the cows equal to 1.85%. This means that, if we sum the daily yield of each cow for the test days (68 days on average), the relative error in the assessment of the total milk production is lower than 2%.
Lastly, the boxplot diagram of the different ISs is reported in Figure 3.24 for the whole dataset containing the 91 investigated cows for scenario B and the most representative values of the diagram are collected in Table 3.2.

The DIM has the highest importance scores, with a median IS = 0.29. Then, THI_0, i.e., the average THI of the day to predict, has a median of IS = 0.13, whereas the other features (THI_{-1}–THI_{-5}) have comparable median ISs ranging from 0.093 to 0.11. Moreover, the feature with highest median IS value (DIM) is affected by the highest variability in the IS values (i.e., highest values of CoV).

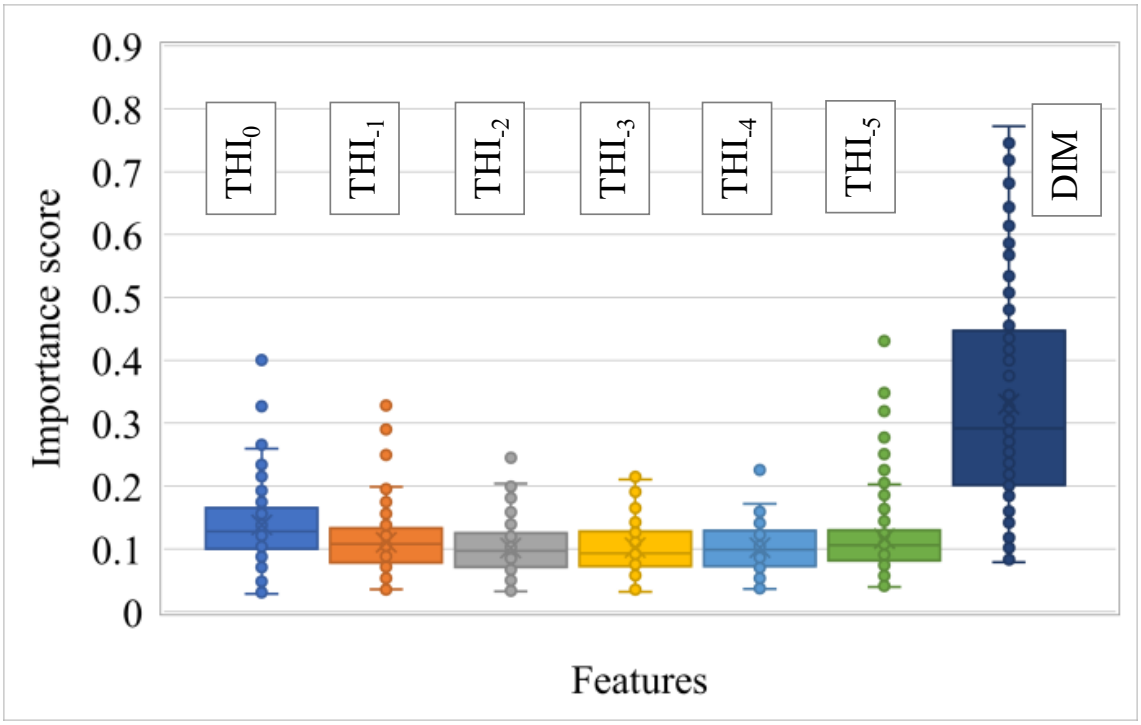

Figure 3.24: Boxplot diagram of the importance score of the different features for the whole dataset in scenario B.

	THI_0	THI_{-1}	THI_{-2}	THI_{-3}	THI_{-4}	THI_{-5}	DIM
Min	0.028	0.035	0.032	0.032	0.036	0.039	0.079
Max	0.416	0.328	0.245	0.230	0.238	0.431	0.772
Median	0.128	0.108	0.097	0.093	0.099	0.106	0.291
St. Dev.	0.065	0.045	0.039	0.041	0.039	0.057	0.172
CoV [%]	50.915	41.595	40.595	43.999	39.314	54.174	58.964

Table 3.2: Minimum value, maximum value, median value, standard deviation value, and coefficient of variation (CoV) value of the ISs obtained for the different independent variables calculated for the whole dataset in scenario B.

Scenario C

In scenario C, the new available data can improve the model predictive capability since new data are introduced in the training phase. In this scenario, milk yield predictions for a short-term period of 5 days have been evaluated with the measured values for each cow. Then, the average relative error on the five daily yields (Er_5) has been evaluated for each subcase obtained by pushing forward the training phase 1 day at a time, as represented in Figure 3.13. The added value of this scenario is that it can monitor the evolution, over the time, of the performance indicators by giving further purposes to the RFMs developed here. In this way, it is possible to establish the effects of the new data introduced in the training dataset and evaluate the trend of Er_5 due to the training window increase. As a representative example, Figure 3.25 displays, for a generic cow (#248), the evolution of the values of the ratio Er_{5B}/Er_{5C}, between the values of Er_5 calculated in an analogous way for the scenarios B and C, respectively. As the figure shows, the values are rather scattered, but the general tendency of the trend is to increase with respect to the value recorded at the beginning of the series (i.e., for a value of the training dataset increase equal to 0, the ratio must be equal to 1). The positive effect during the time can be evaluated as a whole and qualitatively by the slope (m) of the best fitting linear equation. If the value of m is considerably higher/lower than 0, it means that the model is improving/worsening its accuracy. Instead, values of m around 0 indicate a substantial stability of the accuracy of the predictions. This parameter, even if very intuitive, does not have a physical explanation and cannot be related in a simple way to an increase in accuracy. Therefore, for practical reasons, the increase in the predictive capability of the models has been numerically evaluated in terms of median ratio Er_{5B}/Er_{5C} for each cow (see Figure 3.26). Globally, the median ratio for the different cows goes from 1.02 to 3.35, with a mean value for the 91 cows equal to 1.64. Then, the augment in knowledge of the models, as expected, can increase, in a significant way, the accuracy of the predictions.

A further interesting aspect comes from the analysis of the trends of the ISs along the time. Figures 3.27-3.28 display the trends obtained from each IS for two animals, i.e., #26 and #85, having different dataset sizes.

Cow #26 has milk yield data covering 180 days while cow #85 has data from 484 days. For both cows, the IS values are rather stable in the

monitored period. Similar considerations can be drawn for the other investigated animals even if their results are not reported here for brevity.

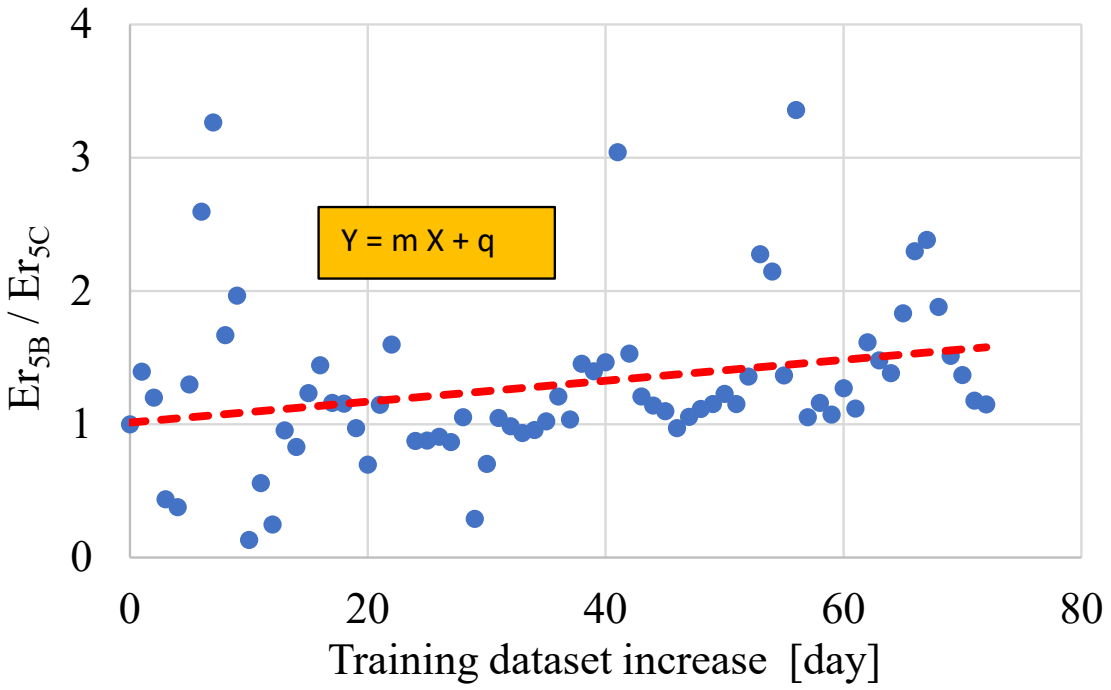

Figure 3.25: Analysis of the results on the ratio Er_{5B}/Er_{5C}: Evolution of the ratio for the cow #248 with training dataset increase.

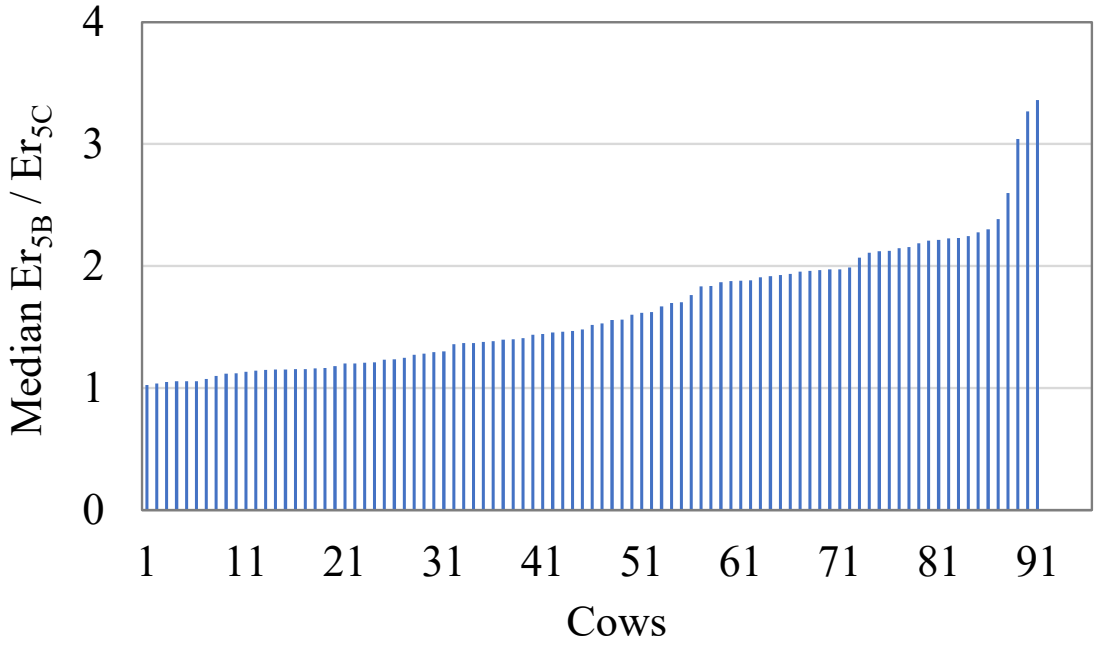

Figure 3.26: Analysis of the results on the ratio Er_{5B}/Er_{5C}: Median ratios in ascending order for the 91 cows.

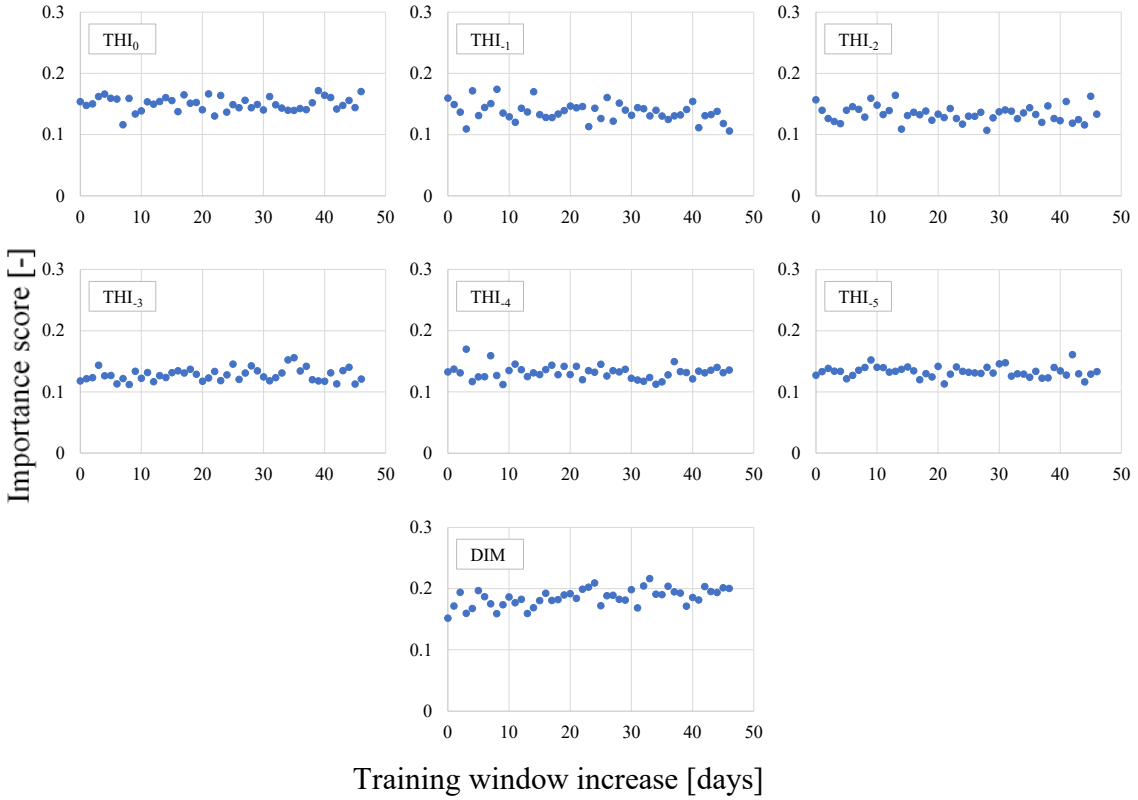

Figure 3.27: Trends of each importance score vs. training window increase for cows #26.

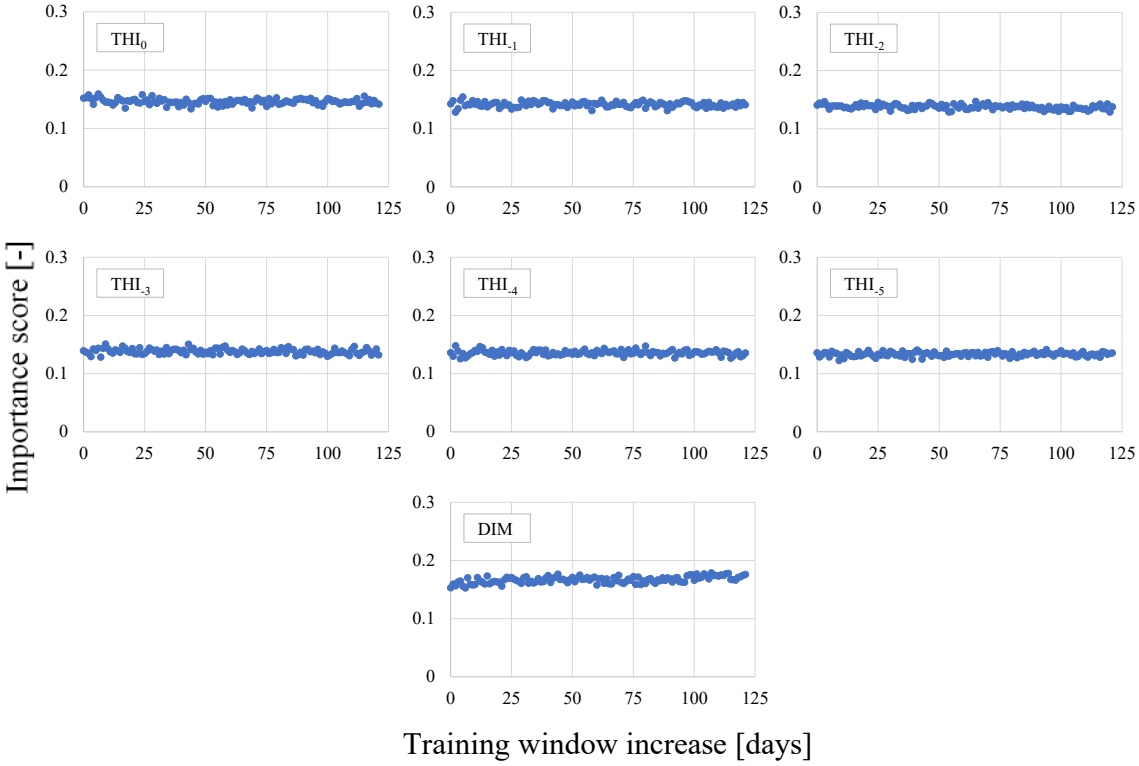

Figure 3.28: Trends of each importance score vs. training window increase for cows #85.

3.2.6 Comparison of Italian and German data results

This section describes the results obtained by applying the model to Italian and German data, the latter coming from the Educational and Experimental Institution for Animal Breeding and Husbandry-LVAT, Groß Kreutz, Brandenburg, Germany. The barn is a free-stall dairy barn with dimensions of 36 m · 18 m, that keeps 51 Holstein Friesian cows.

To make the comparison statistically more realistic, the original german dataset, wider than the italian one, was sampled with the scope of having the same number of cows for the two countries.

In Figure 3.29, the spreading of mean and median error are described. For the Italian case both the errors are greater than the corresponding ones of the german case.

Figure 3.30 shows the histograms of the mean accuracy for the Italian and German case.

The values are more spread in the Italian case, while the maximum is

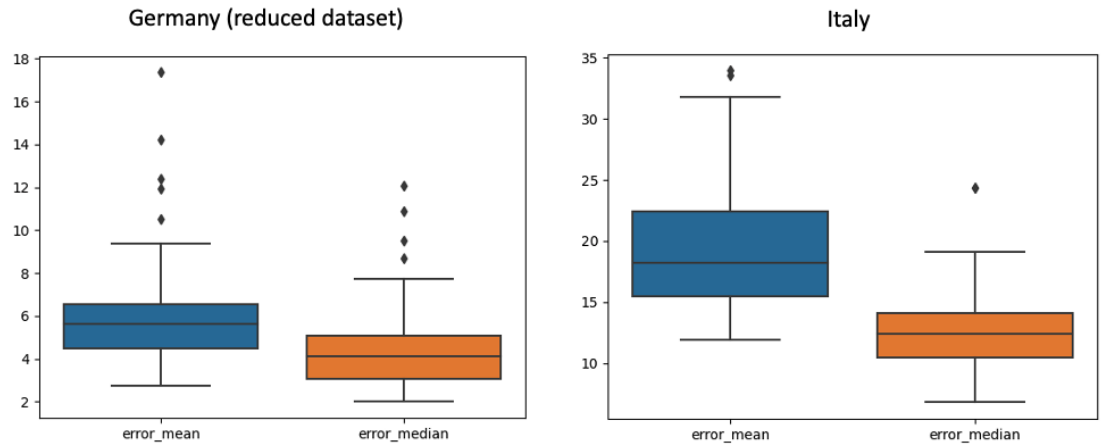

Figure 3.29: RF model applied to Italian and German Data: boxplots of median ed mean errors.

approximately and more spiked in the German one, probably because in the German case the animals react in a more predictable way.
Furthermore, for the German case the most frequent value is around 95, while for the Italian case is between 80 an 85.
Looking at Figure 3.31, it is possible to notice that the importance given by the model to the different features is very similar in the Italian and German case.

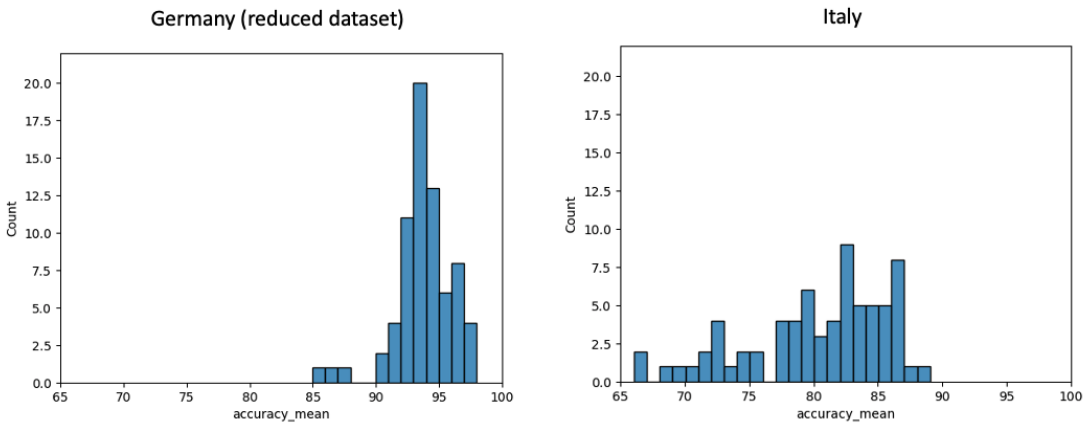

Figure 3.30: RF model applied to Italian and German Data: feature importance.

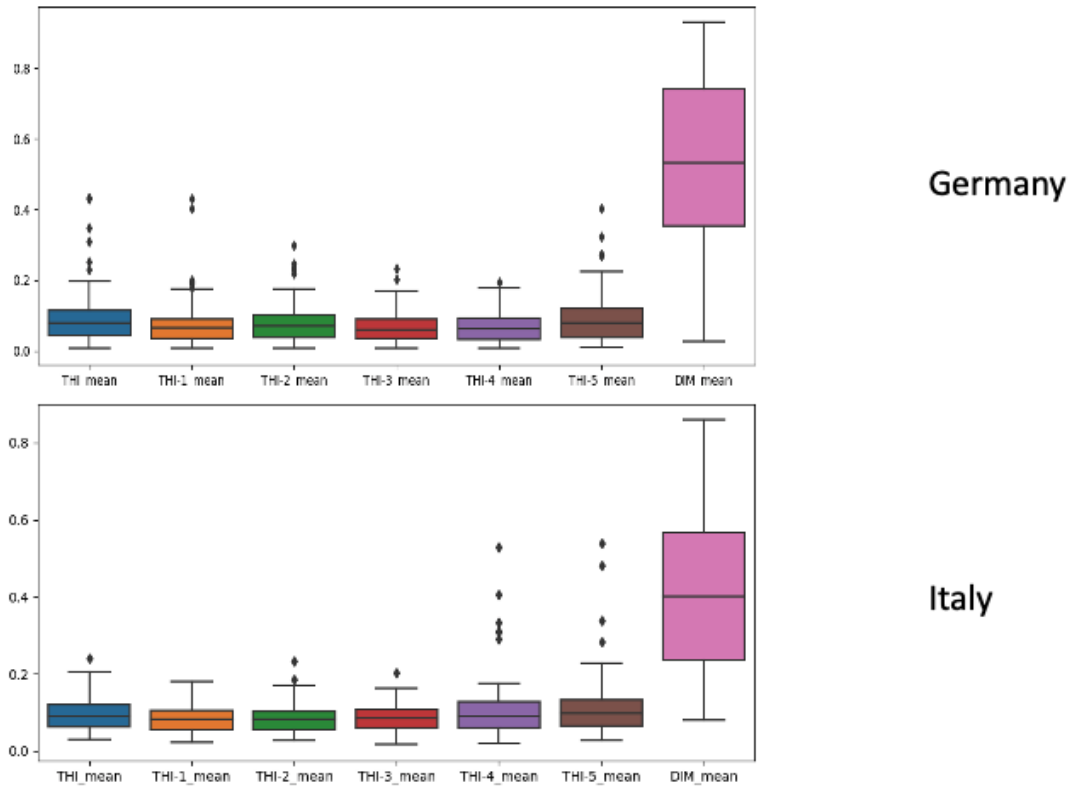

Figure 3.31: Boxplot of the feature's importance for the Italian and German case.

3.3 Robust statistics: multiple Wood fit for anomalies detection

3.3.1 Description of Data

Data analyzed in this sections come from an experimental barn in Potsdam, the same already cited in section 3.2.6.
The only information needed was the daily mean THI, the daily production and the day in milk.

3.3.2 Iterative Wood fit

The datasets related to the AMS and to the microclimatic data were first read using the pandas library and subsequently joined in order to obtain a dataset containing all the information needed for the analysis. The Temperature-Humidity Index (THI) was calculated in the following way (National Research Council, 1971):

$$THI = 0.8 \cdot T + RH(T - 14.4) + 46.4 \tag{3.7}$$

As reported in (Agrusti et al., 2022), for each animal and for each lactation, it is possible to fit the Wood model to obtain the parameters a, b, c:

$$MY(DIM) = a \cdot DIM^b \cdot e^{-c \cdot DIM} \tag{3.8}$$

Where: MY is the daily milk yield [kg/d] and DIM is days in milk. A filter was applied to select only the data where the daily THI did not exceed a threshold of 65, used as a predicator of potential heat stress. A more robust statistics can be obtained by randomly sampling the original amount of data and producing a different Wood fit for each sample. This way, a collection of Wood models is obtained:

$$MY_k(DIM) = a_k \cdot DIM^{b_k} \cdot e^{-c_k \cdot DIM} \tag{3.9}$$

where a_k, b_k, c_k are the parameters of the k-th curve. The sampling and fitting process can be repeated N times, selecting each time a fixed fraction f of the original data. Here, we used N=500 and f=1/10. The obtained family of curves and the corresponding parameters can then be used to define a representative median curve:

$$MY_{median}(DIM) = A \cdot DIM^B \cdot e^{-C \cdot DIM} \tag{3.10}$$

where:

$$\begin{cases} A = median(a_k) \\ B = median(b_k) \\ C = median(c_k) \end{cases} \tag{3.11}$$

The median values of the parameters have been assumed instead of the mean values since they are not affected by outlier values. In some cases, unacceptable curves are obtained, e.g., entailing negative or infinite values or unrealistic trends. For this reason, it has become necessary to perform a selection of only the meaningful curves. Then, residuals were calculated as the difference between the actual milk yield data and values of the median curve. The beam of such curves is then filtered selecting only curves with an initial positive trend. The dispersion of the different values obtained in correspondence of the different curves can be used to define a criterion for the detection of anomaly values, for instance by selecting a proper multiple value of the standard deviation value of defining a confidence interval. This because the values more

distant from the median curve would be considered in the method as anomaly points.
Figure 3.32 describes the flow chart of the Iterative Wood fit method, from data collection to the detection of the anomalies in daily milk yield.

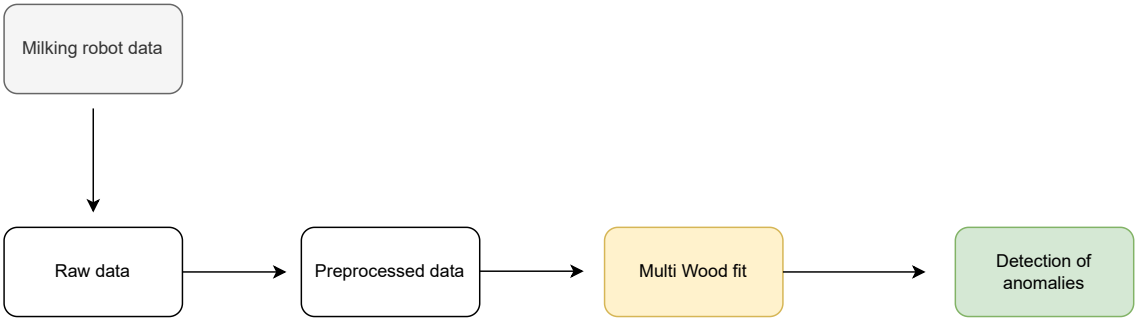

Figure 3.32: Flow chart of Iterative Wood fit method for detection of anomalies.

3.3.3 Standard deviation as threshold for anomalies' severity

The method proposed here allows the introduction of a lactation model that is robust with respect to statistical fluctuations and automatically creates an acceptability range linked to the dispersion of the curves belonging to the beam. The standard deviation and its multiples can be used to find a threshold for the residuals in order to determine whether any given value is an anomaly. All points out of the beam were considered "anomalies", Note that positive residuals were also considered. While positive residuals cannot be attributed to heat stress or other adverse effects, they can be informative and could serve as indicators of change in the physiological condition of a cow.

3.3.4 Anomalies' detection applications

The use of a multiple fit on different partial datasets obtained after the sampling operation, led to 500 Wood curves for each lactation. In Figure 3.33, the curves deemed physically acceptable are shown for a lactation cycle of one sample animal.

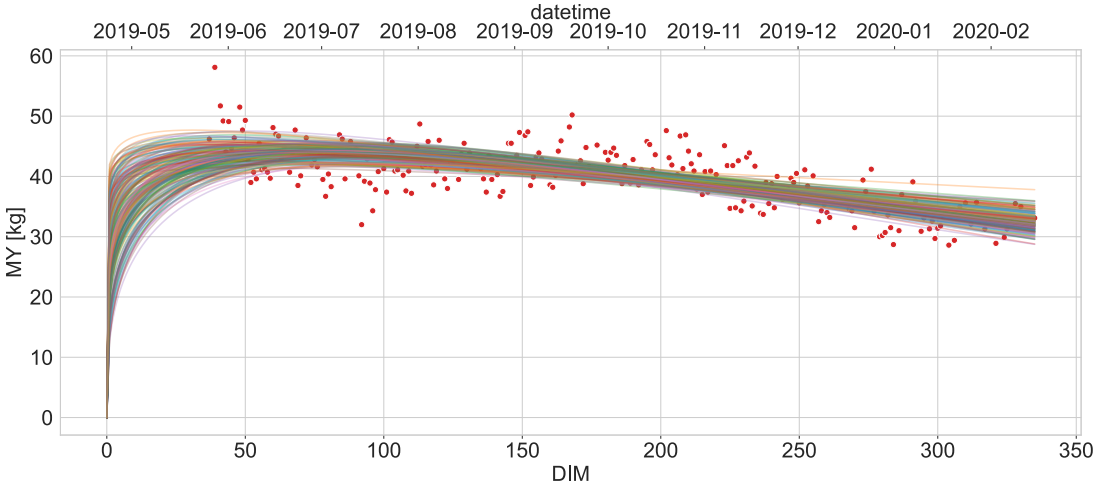

Figure 3.33: : Wood curves with zootechnical significance for one lactation of a cow considered in the study.

Figure 3.34 shows, in solid blue, the median curve obtained by computing the median values of the parameters of the fit curves of Figure 3.33. Then it shows the 95% confidence interval expressed by two standard deviations from the median curve. Blue points are considered within normal range (could be attributed to the expected normal dispersion), while orange points, outside the 95% range have been considered as anomaly points.

As seen in Figure 3.35, residuals can be negative or positive, meaning in the latter case that milk yield can exceed the expected value. Therefore, in order to detect net production deficit in consecutive intervals, the daily residuals must be accumulated. In Figure 3.36 the cumulative curves of the expected and real milk yield trends are shown and overlapped to the trend line of their differences (the solid blue line). It is interesting to note that this differences (solid blue line in Figure 3.36) reach the value in correspondence with a DIM value equal to 90 days, about corresponding to the days with a production peak in the lactation curve. This is a recurrent condition with reference to the group of cows analysed, meaning the model was able to predict with high accuracy the cumulative milk yield in the first 90 days corresponding to the most productive stage of the lactation.

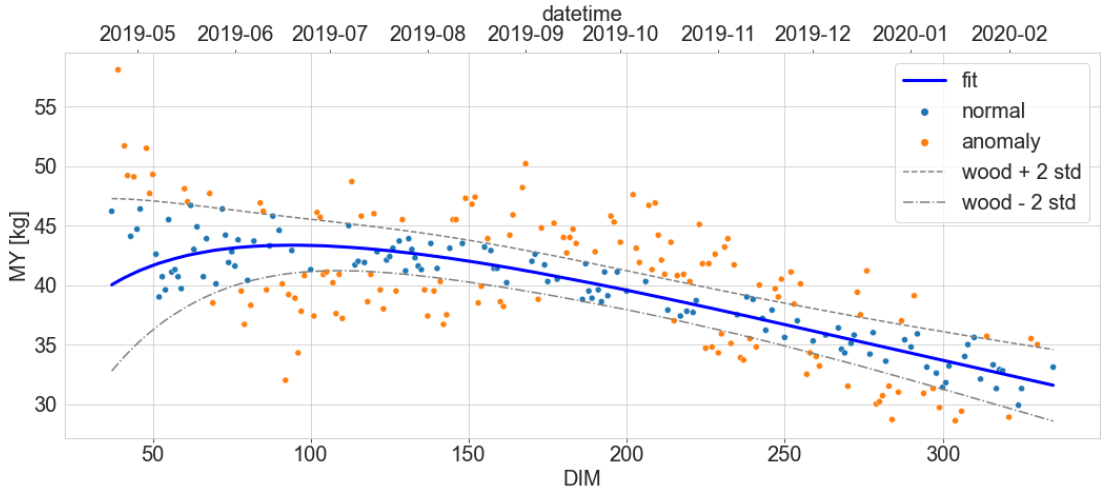

Figure 3.34: Anomaly detection in the plane DIM Vs MY.

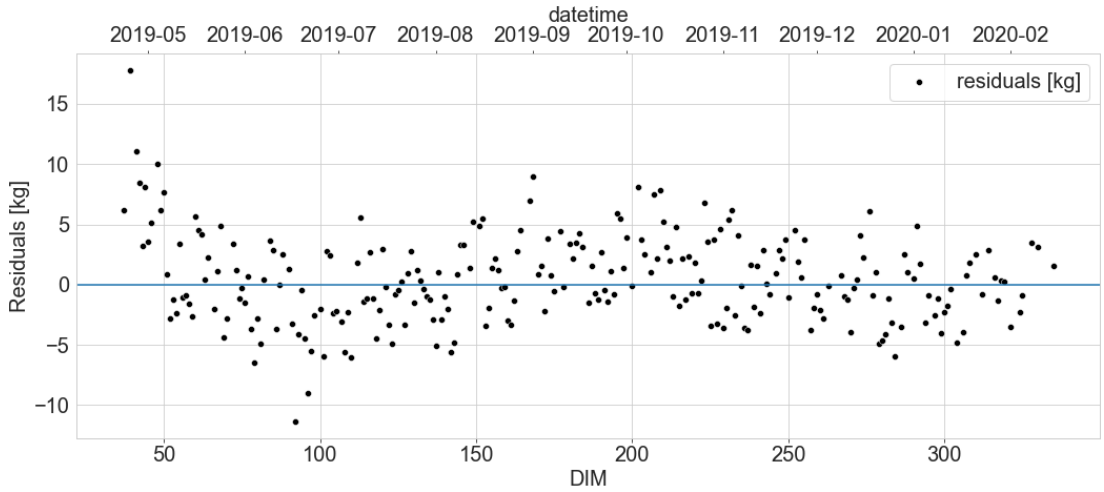

Figure 3.35: Scatterplot of the residuals of a fit curve.

Figure 3.37 shows the trend of the average difference between the expected and real cumulated milk yield for the entire group. As above anticipated the minimum values can be observed in correspondence to DIM ranges characterized by high values of the lactation curve (see also Fig. 3.34). This analysis has been performed on all the cows counting at least 100 consecutive days in milking and at least 8 valid fit curves.

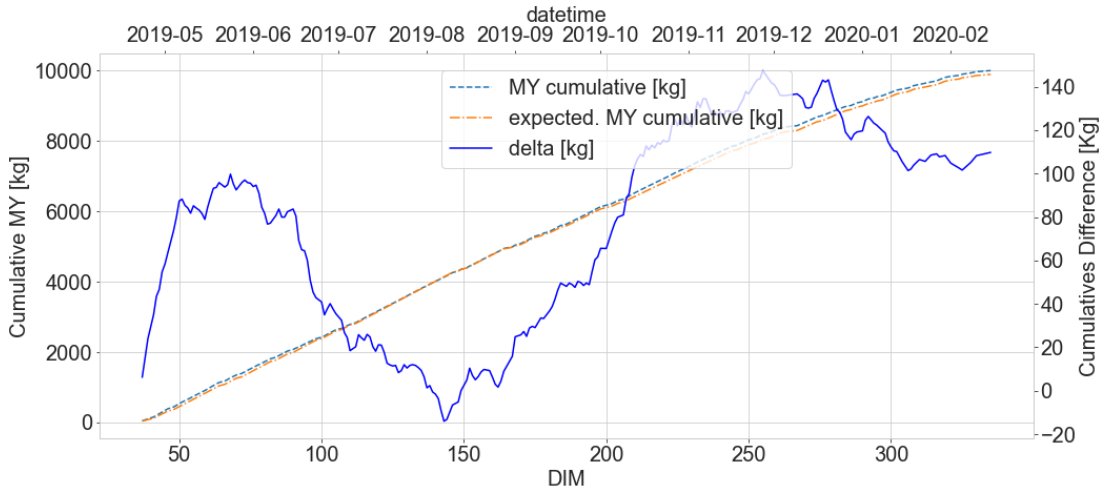

Figure 3.36: Cumulative curves of the expected and real Milk Yield (dashed and dash-dotted lines, respectively) and the corresponding differences (continuous blue line). In the legend delta represents the curve of the differences between the real and expected values.

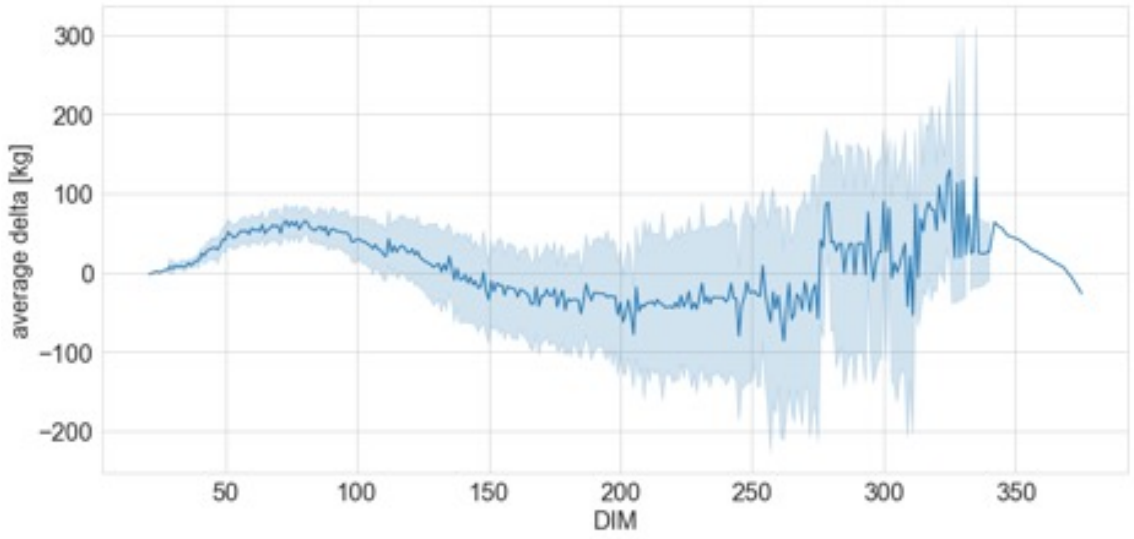

Figure 3.37: Average delta between expected and real cumulated milk yield. The light-blue coloured interval represents the 95% confidence interval.

Chapter 4

Discussion of results

The results presented in chapter 3 deserve an appropriate comment starting from the usefulness of these in the technical field and arriving at the limits and problems connected to them.
Statistical methods and machine learning today represent the avant-garde of research, in the field of Big Data and computer Vision, as they allow to simulate the learning and use of acquired knowledge to predict, classify and estimate. However, it is right to recognize that these methods, even if they return objective results that cannot be influenced by a human opinion, can lead to deception. For this reason, for example, in the most delicate areas such as the medical one, the automatic recognition of pathologies, organs and undesired objects and the prediction of problems based on vital parameters, is always accompanied by the opinion of an expert doctor, who can confirm or deny as provided by the algorithms.
In the section 3.2.3 the use of the Random Forest algorithm for the prediction of the daily milking yield of the single animal was introduced. The algorithm is able to make this prediction by looking at other examples of the cow's response to the microclimatic conditions of the barn in the previous 5 days. Certainly the forecast accuracy is strongly dependent on the days of training of the algorithm. It follows that in the first days, the prediction is almost entirely random and becomes more exact in the second part of lactation.
On the other hand, a point in favor of the algorithm is its ability to work with a limited number of data and not to need an extremely large number of examples before reaching an acceptable accuracy.
The model should therefore be understood as a guide and a tool for the farmer, who with his experience knows how to recognize the truthfulness of the forecast and take the necessary measures in advance.
The algorithm was used to develop an application, also in collaboration

with stakeholders interested in the commercialization of proper software.
In section 3.3.2 introduces the multi fit of wood, as a tool to identify a variable acceptability range of daily production, and to give rise to the introduction of the concept of anomaly. The method relies on robust statistics and the objectivity of parameters such as standard deviation to find the lower and upper limits of the acceptability beam around the median fit. On the one hand, it is good to underline the robustness of the method, which allows, by means of resampling, to have a more stable wood model less influenced by production outliers; on the other hand, it is useful to highlight that the identified method introduces an objective criterion for the identification of anomalies which unfortunately has a purely theoretical and academic implication for the moment, as the anomalies of an animal can only be highlighted at the end of lactation.
A more practical utility of the method can be identified by looking at a subsequent use of the anomalies as examples for the training of a machine / deep learning algorithm that can recognize production irregularities on the basis of these examples.
This would be the missing piece that has hitherto prevented the development of such a forecasting tool.
Last but not least, it is necessary to highlight a problem concerning the analysis of data in the zootechnical field and which does not depend on the method of analysis used. The data produced by robots or milking stalls are in fact generated in different formats depending on the brands of the tools. Furthermore, the measured variables also change from model to model and therefore often from barn to barn. This makes the analysis of the data much more complicated, since data need to be as homogeneous as possible.
The heterogeneity of the data also inevitably produces a great waste of money in the installation of sensors that will be completely ignored at the end of the process, as they measure and monitor variables not present in all the stables.

Conclusions

We have concluded our discussion about the applications of Big Data Analytic algorithms to PLF data. In this work we have touched several and different topics related to this theme.

The first part of the study offers an analytical interpretation of the internal temperature trend dependent by the cooling coefficient of the barn, allowing in future studies to build a customized model for each barn and to link the internal temperature trend to the external one, optimizing the use of sensors.
Section 3.2 aimed to define and test a Random Forest-based model for the assessment of the daily milk yield at the single cow level. The model has been applied to the data collected in two years, 2016 and 2017, in a dairy farm, located in northern Italy, and collected both productive data from the automatic milking system and environmental data from two thermo-hygrometric sensors.
The statistical model used for the interpretation of the collected data is composed of seven predictors: days in milk of the cow, daily average THI of the day of the assessment and those of past five days.
The results showed that the model can detect the drop in the cow's milk yield due to extreme hot conditions inducing heat stress effects. In fact, the average relative error provided by the model in the predictions, is about 18% with a single daily yield, whereas it becomes just 2% if the total milk production in the test days is considered. The outcomes reported in the study seem to be particularly relevant for three main reasons:

1. the size of the training dataset adopted in the analysis is suitable for the objective of the study;

2. the statistical model assumed in the study seems suitable for the work;

3. the RFM developed by the regression procedure is rather robust and reliable with respect to the type of data.

Then, the results confirm that the obtained RFM can represent a reliable and viable tool for the evaluation of future productive scenarios of dairy cows in the presence of heat stress effects. This could help to develop and improve decision support for farmers to increase both milk yield and animal welfare and, on the other hand, to reduce the resources needed, so to increase the sustainability of the dairy sector.
Furthermore, since currently there is no systematic and statistically robust method for detecting production deviations from caused by various factors, such as environmental stress, the statistical method developed in section 3.3.2 offers a robust way to identify production anomalies in the lactation period of individual cows on the basis of a multiple fit. Moreover, the use of a multiple of the standard deviation to define the acceptability range of the daily milk yield can leads to the introduction of a variable threshold, which could be used for production anomaly detection.
The anomalies identified with this method can be both positive and negative with respect to the range of acceptable values. This feature makes it clearly visible a further prospective of the use of this method: its application to an even greater number of cows and lactations will allow to collect an increasing number of anomalies. This can help a machine learning model, which is the subject of an ongoing study, in its training phase, making its forecasting performances more stable. This approach can be also used to classify daily production data as "normal" or "abnormal".

Acknowledgment

A well-deserved thanks for the writing of this thesis but above all for the three-year work behind it goes to the supervisor Prof. Stefano Benni and to the co-supervisors Prof. Daniel Remondini, Eng. Marco Bovo and Prof. David Nail Manners. Their support was essential to introduce me to the world of precision animal husbandry and agricultural engineering. A special thanks also goes to the whole research group, led by Prof. Patrizia Tassinari and Prof. Daniele Torreggiani, who are always careful to take care of and follow the progress of the doctoral students' activities.
Not least should be remembered the parents, grandparents, uncles and friends, who, encouraging me in the darkest moments, have always cheered for the success of this enterprise from "behind the scenes".
A special thanks goes to my grandfather, who in the last year has reminded me how essential it is to grasp the importance of time passing and of the extraordinary things that we are lucky enough to possess.

Bibliography

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., and Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey.

Agrusti, M., Foroushani, S., Ceccarelli, M., Bovo, M., Torreggiani, D., Tassinari, P., Amon, T., and Benni, S. (2022). Assessment of productive anomalies in dairy cows. *Conference Proceedings, European Conference on Precision Livestock Farming (ECPLF) 2022*.

Benni, S., Pastell, M., Bonora, F., Tassinari, P., and Torreggiani, D. (2020). A generalised addictive model to characterise dairy cows' responses to heat stress. *ANIMAL*, 14:418–424.

Berckmans, D. (2014). Precision livestock farming technologies for welfare management in intensive livestock systems. *OIE Revue Scientifique et Technique*, 33(1).

Bonora, F., Benni, S., Barbaresi, A., Tassinari, P., and Torreggiani, D. (2018a). A cluster-graph model for herd characterisation in dairy farms equipped with an automatic milking system. *Biosystems Engineering*, 167.

Bonora, F., Pastell, M., Benni, S., Tassinari, P., and Torreggiani, D. (2018b). ICT monitoring and mathematical modelling of dairy cows performances in hot climate conditions: a study case in Po valley (Italy). *Agricultural Engineering International: CIGR Journal*, 20(Special issue: Animal Housing in Hot Climate):1–12.

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., and Tassinari, P. (2021). Random Forest Modelling of Milk Yield of Dairy Cows under Heat Stress Conditions. *Animals*, 11(5).

Bovo, M., Benni, S., Barbaresi, A., Santolini, E., Agrusti, M., Torreggiani, D., and Tassinari, P. (2020). A Smart Monitoring System for

a Future Smarter Dairy Farming. In *2020 IEEE International Workshop on Metrology for Agriculture and Forestry, MetroAgriFor 2020 - Proceedings*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2).

Breiman, L. (2001). Random Forest. Machine Learning 45. *Machine Learning*.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Boca Raton, routledge edition.

Cockburn, M. (2020). Review: Application and prospective discussion of machine learning for the management of dairy farms.

Cowley, F. C., Barber, D. G., Houlihan, A. V., and Poppi, D. P. (2015). Immediate and residual effects of heat stress and restricted intake on milk protein and casein composition and energy metabolism. *Journal of dairy science*, 98(4):2356–68.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random Forests. In *Ensemble Machine Learning*, pages 157–175. Springer US, Boston, MA.

Dawkins, M. S. (2017). Animal welfare and efficient farming: is conflict inevitable? *Animal Production Science*, 57(2):201–208.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *31st International Conference on Machine Learning, ICML 2014*, volume 2.

Fournel, S., Rousseau, A. N., and Laberge, B. (2017). Rethinking environment control strategy of confined animal housing systems through precision livestock farming. *Biosystems Engineering*, 155:96–123.

Halachmi, I., Guarino, M., Bewley, J., and Pastell, M. (2019). Smart Animal Agriculture: Application of Real-Time Sensors to Improve Animal Well-Being and Production. *Annual review of animal biosciences*, 7:403–425.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer-Verlag New York 2009, New York, second edition.

Heinicke, J., Ibscher, S., Belik, V., and Amon, T. (2019). Cow individual activity response to the accumulation of heat load duration. *Journal of Thermal Biology*.

John, A. J., Clark, C. E., Freeman, M. J., Kerrisk, K. L., Garcia, S. C., and Halachmi, I. (2016). Review: Milking robot utilization, a successful precision livestock farming evolution. *Animal*.

Lovarelli, D., Bacenetti, J., and Guarino, M. (2020a). A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production? *Journal of Cleaner Production*, 262:121409.

Lovarelli, D., Finzi, A., Mattachini, G., and Riva, E. (2020b). A Survey of Dairy Cattle Behavior in Different Barns in Northern Italy. *Animals : an open access journal from MDPI*, 10(4).

Lovarelli, D., Tamburini, A., Mattachini, G., Zucali, M., Riva, E., Provolo, G., and Guarino, M. (2020c). Relating Lying Behavior With Climate, Body Condition Score, and Milk Production in Dairy Cows. *Frontiers in veterinary science*, 7:565415.

National Research Council (1971). *A guide to environmental research on animals*. National Academy of Sciences, Washington.

Ng, A. (1998). Preventing "overfitting" of cross-validation data. *Proceedings of the Fourteenth International Conference on Machine Learning*.

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., and Belfkih, S. (2018). Big Data technologies: A survey.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piwczyński, D., Sitkowska, B., Kolenda, M., Brzozowski, M., Aerts, J., and Schork, P. M. (2020). Forecasting the milk yield of cows on farms equipped with automatic milking system with the use of decision trees. *Animal Science Journal*.

Python Software Foundation (2020). Python.

Rotz, C., Coiner, C., and Soder, K. (2003). Automatic Milking Systems, Farm Size, and Milk Production. *Journal of Dairy Science*, 86(12):4167–4177.

Rowsell, H. C. (1972). A Guide to Environmental Research on Animals. *The Canadian Veterinary Journal*, 13(8):196.

Shinde, P. P. and Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. In *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*. Institute of Electrical and Electronics Engineers Inc.

Shobana, V. and Kumar, N. (2015). Big data - A review. *International Journal of Applied Engineering Research*, 10(55):1294–1298.

Strpić, K., Barbaresi, A., Tinti, F., Bovo, M., Benni, S., Torreggiani, D., Macini, P., and Tassinari, P. (2020). Application of ground heat exchangers in cow barns to enhance milk cooling and water heating and storage. *Energy and Buildings*, 224:110213.

Tassinari, P., Bovo, M., Benni, S., Bonora, F., Barbaresi, A., Santolini, E., Franzoni, S., Daniele, T., Poggi, M., Mammi, L. M. E., Mattoccia, S., and Di Stefano, L. (2021). A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. *Computers and Electronics in Agriculture*, 182(106030):1–15.

Tullo, E., Mattachini, G., Riva, E., Finzi, A., Provolo, G., and Guarino, M. (2019). Effects of Climatic Conditions on the Lying Behavior of a Group of Primiparous Dairy Cows. *Animals : an open access journal from MDPI*, 9(11).