Milestone Lead	UNITUS
Milestone due date	2023/09/30
Status	FINAL
Version	V1.0
Project	SEBASTIEN

Co-financed by the Connecting Europe Facility of the European Union

Agreement: INEA/CEF/ICT/A2020/2373580 **Action**: 2020-IT-IA-0234

DOCUMENT INFORMATION

Title	Milestone 4 - Report on data synthesis
Agreement	INEA/CEF/ICT/A2020/2373580
Action	2020-IT-IA-0234
Creator	Marco Milanesi (UNITUS)
Milestone Description	Report on the data synthesis methodologies
Means of verification	Report on the data synthesis methodologies shared with the Agency
Contributors	Giovanni Vignali, Daniele Pietrucci, Federica Gabbianelli, Giovanni Chillemi (UNITUS), Alfredo Reder, Alessandro D'Anca (CMCC), Elisa Somenzi, Paolo Ajmone Marsan (UCSC)
Requested deadline	M18
Reviewer	Mario Barbato (UCSC), Marco Mancini (CMCC)

Summary

1. Executive summary

This milestone aims to describe the methodologies applied to achieve the SEBASTIEN services.
In particular, here we provide an overview on the problems faced and of the solutions implemented. We provide context on data, methodologies and analyses along with reporting the preliminary results obtained in developing the prediction models for the four SEBASTIEN services. The four models will generate the indexes and indicators agreed with the stakeholders, which will ultimately be at the forefront to overcome some of the upcoming challenges facing the livestock sector.
In particular, in service 1.a machine learning (ML) models were developed to estimate the short and long period effect of climate on milk characteristics (yield, fat and protein percentage). Climatic data, together with historical production data were used to achieve this goal. In service 2, Temperature-Humidity Index (THI) is estimated inside the barns on the basis of the external environmental conditions. Inside THI data on reference farms were used to train a ML model; the model is used to predict short- and long-term THI inside the barn values on a much larger number of farms. In service 3, regression models were applied to estimate the pasture biomass (fresh and dry matter) using Sentinel2 satellite data. Finally, in service 4, a ML model was developed to predict the future bluetongue spread in the Sardinia region.

2. Introduction

SEBASTIEN's objectives are combining useful existing data resources using models from classical statistical inference to Machine Learning (ML) or Deep Learning (DL) approaches to obtain indicators (quantitative and qualitative) to monitor and detect the effect of climate/environmental stresses affecting livestock systems. For this reason, four main services have been developed. Here the details about the methods applied are described.
To develop the four SEBASTIEN's services, methods to elaborate, correct and combine the climatic, territorial and animal data collected were explored and pipelines were created, with the objective to obtain indicators and indices useful for the stakeholders. Several empirical approaches along with statistical and mathematical methods (e.g., regression, clustering, machine learning, etc.) were tested to obtain the prediction models used in the four SEBASTIEN's services.

In particular, the methods we present here were applied to:

- data quality control and preparation;
- identification and testing of the prediction models;
- prediction model application in the services.

3. Data and indicators used

To develop the SEBASTIEN services, large High Value Datasets coming from multi-sources and multi-thematic portals were used. The details about the dataset used are reported in the D2.2 "List of suitable data sources and of newly acquired data" and M3 "Brief report on the data sources identified".

The objective is to derive indicators and indices used in the support services. These are reported in D2.1 "List of indicators/indices to be proposed to stakeholders". Those indicators and indices describe environmental, territorial characteristic and animal parameters that could affect animal welfare and could have repercussions on livestock reproduction, productivity, health, and mortality and, as a consequence, animal welfare, production efficiency and environmental impact.

As an example, the THI (Temperature Humidity Index) is an index extensively used in SEBASTIEN services (bioclimate indicators n. 44 - Annex A D2.1). THI is a bioclimatic index that combines the simultaneous effect of temperature and relative humidity (RH) and is used to characterize heat stress in livestock, particularly in cattle [[1,2]. The formula here applied is:

$$THI = (1.8 \times T + 32) - (0.55 - 0.55 \times RH) \times [(1.8 \times T + 32) - 58]$$

where T is the ambient temperature expressed in Celsius degrees (the term *1.8 × AT + 32* accounts for the conversion of temperature data from Celsius degrees to Fahrenheit degrees); RH is the relative humidity.

THI is one of the most worldwide used indexes to set comfort, stress, and life-threatening environmental conditions due to heat stress in livestocks [3]. Heat stress has detrimental effects on animal health and, consequently, on productivity [4]. As the majority of studies on heat stress in livestock have focused mainly on temperature and relative humidity, THI can be considered as a single value representing the combined effects of such variables associated with thermal stress.

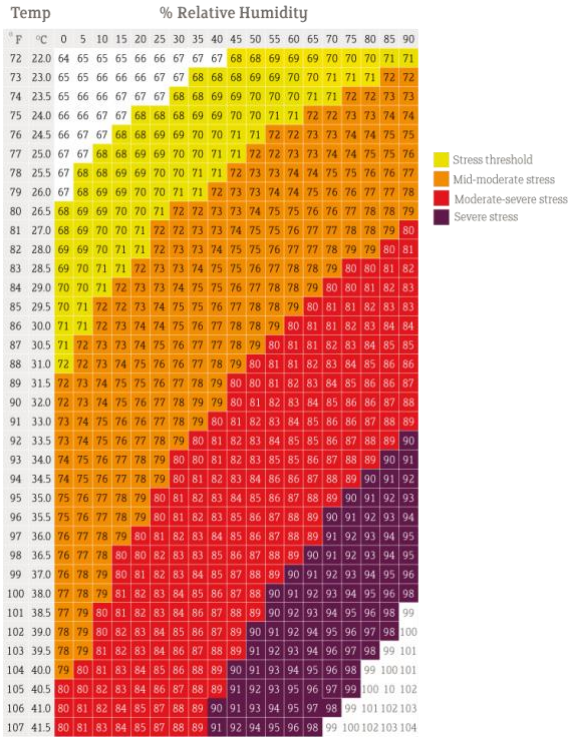

Figure 1 *- THI table and relationship with temperature and relative humidity. Colors represent THI values ranging from comfort (white) to life threatening (purple) for dairy cattle.*

Figure 1 shows how the THI (and so the heat stress) increases depending on temperature and relative humidity: THI indicates a stress (from low to severe) starting from the value of 67 in dairy cows. As mentioned before, THI values are related to heat stress and the consequent risk of undesired consequences, for example reduction of milk yield.

4. Methods

The input data used to develop the following methods may be divided into “target” variables (i.e., data to predict) and features (i.e. data used to predict the target variable). Historical data (i.e.,

variables and features available) were used in the training, to implement the 'algorithm selection' pipeline, whereas real-time, forecast or projection data were used in the inference, to generate the desired indices.

Before proceeding with the algorithm detection, the input data need to be controlled (i.e., quality control), transformed and harmonized. Hence, different layers of analyses were implemented to prepare the input data. Moreover, different pipelines, based on the diverse nature of the data, were implemented to remove incorrect and inconsistent data.

First of all, the raw input data were automatically checked for inconsistency, missing data and typos, such as presence of characters when only numbers were expected. Then, the quality control of the feature, based on its own values or associated features, was applied. The quality control was also based on previous information, retrieved from the literature.

In some cases, the input data need to be transformed or pre-processed before use. These transformations are required to improve an algorithm prediction. Specific pipelines for specific types of data and targets were implemented and applied, e.g.: we implemented a pipeline to minimize autocorrelation among variables (features) and avoid model overfitting.

Based on the final scope of the analyses, specific input data of different nature (i.e., climatic data, productive data, remote sensing, etc.) and origin are merged and analyzed together. The aim is to obtain a validated model, by identifying the best algorithm, optimizing its parameters and automatically selecting the most important features. Automatic pipelines were implemented accordingly.

The pipelines mentioned in this document were implemented using Python, R or bash.

In the four services, we tested classical statistical models, such as linear regression, and machine learning approaches. With the latter, a workflow already implemented in the Highlander project, was improved in SEBASTIEN and applied. In particular, a function to remove autocorrelate features at the beginning of the pipeline was added.

Details about the procedures for each service will be provided in the next sections.

4.1. Machine Learning workflow

The search of the best Machine Learning algorithm family is performed using the H2O.ai AutoML [5] and scikit-learn [6] modules, from Python. H2O tests different Machine Learning algorithms (e.g., Random Forest, XGBOOST, Gradient Boosting Machine - GBM), identifying the best one using statistical metrics, for example, MAE (Mean Absolute Error). Then, the optimized algorithm hyper-parameters were identified using a grid search. Depending on the target variable nature, different metrics to evaluate the best algorithm or the number of features to keep will be used (for example, MAE).

Using the best algorithm and parameters previously identified, the complete dataset is analyzed to select the most informative features associated with the target variable according to the 'feature importance' metric. Finally, the algorithm is running the training only using the previously identified subset of features.

In addition, the SHAP algorithm (SHapley Additive exPlanations) was implemented to allow the explainability of each feature in the classification. This is a recent approach that relies on the Shapely values, first introduced in game theory. The SHAP values are evaluated using a permutation approach for each feature for all the samples in the test set. They can be imagined as the marginal contribution of each feature to the prediction. If a feature presents an “high” SHAP value, it is involved with a positive contribution in explaining the target variable, whereas a “low” SHAP value reflects a negative contribution. This metric is summarized in a ‘SHAP summary plot’, which aids the understanding of each feature's impact on the ML model.

5. Services implementation

In the following sections, some of the methods used and applied will be described, together with the data used and the indicators calculated.

5.1. Service 1

The general scope of this service is to support livestock farms to contrast climate change, in a short- and long-term frame. The service was split into three sub-services, with the objective to respond to stakeholders needs. More details are reported in the next sub-chapters.

5.1.1. Service 1.a

5.1.1.1. Introduction

Climate change will have a relevant impact on several aspects of human life and activities (e.g. agriculture, tourism). Moreover, climate change will generate more frequent extreme climatic events (e.g., hotter temperature, drought, heavy rains, etc.) and will impact our environment on several levels. Livestock is and will increasingly be under pressure since the extreme event will affect their welfare and productivity. To help farmers to minimize the consequences of heat stress, the creation of a model to predict the effects on livestock of climate variation at short and long term can be an important asset to adjust the current and future farm management. The objective of the service 1.a, is to create a ML model of heat stress effects on livestock that can be applied to short term weather forecast (2 days prevision, using COSMO-2I data - D2.2, WC 6) and long term climate projection (VHR-PRO - D2.2, WC 11) [7]. The data from the last one can be used to develop a roadmap to mitigate the effects of climate change, at decision makers level.

5.1.1.2. Data used in the model implementation

In this service, to implement the Machine Learning prediction model we used “animal-based” and bioclimatic data.

About the animal data, we used:

- production data as a predictor, in particular milk yield, protein, and fat percentage from “Pezzata Rossa Italiana”. The data were retrieved through the LEO project from AIA partner and the breed association (ANAPRI) [D2.2, from AW_1 to AW_50];
- phenotypic data are also associated with additional information (e.g., day in milk, number of lactations, age, number of functional controls, etc.). These data are used to correct the observed value and obtain only the residual part affected by the environment [D2.2, from

AW_1 to AW_50]. Moreover, the estimated breeding values (EBVs) were also used to take into account the genetic value; this information was provided by ANAPRI.

About the climatic data, different resources were used:

- single climatic variables, such as temperature, relative humidity, wind speed, etc… from the VHR-REA dataset [8] [D2.2, WC 10];
- climatic indices (Temperature Humidity index - THI) from the VHR-REA dataset [8] for the external conditions (pasture) [D2.2, WC 10]; for the internal conditions the data were retrieved from Service 2 results (see below).

A pilot dataset was used to implement the methods: production data from 1990 to 2020 from Friuli-Venezia Giulia region were obtained from ANAPRI ("Pezzata Rossa Italiana" Breeder Association) and LEO project (from AIA) with a total of 2,511,947 Functional Control (FC). The dataset was controlled to remove outliers, and data with missing information. In particular, days in milk (DIM) between 5 to 400, parity lower or equal to 9, animals older than 22 months, number of FC per lactation between 5 and 14 were kept. To better address the role of climatic data in production (i.e., short and long term effects), the climatic effect was evaluated up to thirty days up to the FC. This allowed a fine evaluation of short and long-term effects on milk production and quality.

5.1.1.3. Model developed

Different models were developed testing single climatic variables and THI for external conditions (pasture) and THI for internal ones. For the sake of simplicity, we present an example of the results obtained using the single climatic variables. However, the same pipeline was used also for the THI (external and internal conditions).

The first analysis step consists in applying a multiple linear mixed model to each phenotype to correct the values from fixed and random effects. In particular the features DIM (as a class of 15 days), age in months, parity (from 1 to 6, and the remaining were included in the class 7), and IDAS EBV ("Indice Duplice Attitudine Sostenibile"- Sustainable Double Purpose Index; to control the genetic value of each animal) were used as fixed effect; animal identification (to take into account the data repetition) and farm identification (to take into account the farm management) were included as random effects.
The objective is to obtain the residual values that include the error of the model plus the environmental effect. The last one is the one we are interested in evaluating using the ML model hereafter presented.

At the same time, the climatic variables were analyzed. In particular, a correlation matrix was used to evaluate autocorrelation among variables. In many cases, the result for the correlation between the same variable for different days is a tight correlation (Figure 2). Correlations among different climatic variables were also evaluated, and were also observed in this strong correlation, for example between the maximum and average temperature. For this reason, some climatic

variables were removed from the dataset (i.e., average temperature) and for the remaining statistical transformation (for example, mean or sum) were used to reduce the high correlation among days. The scope is to avoid bias in the estimations and avoid underestimating the model error.

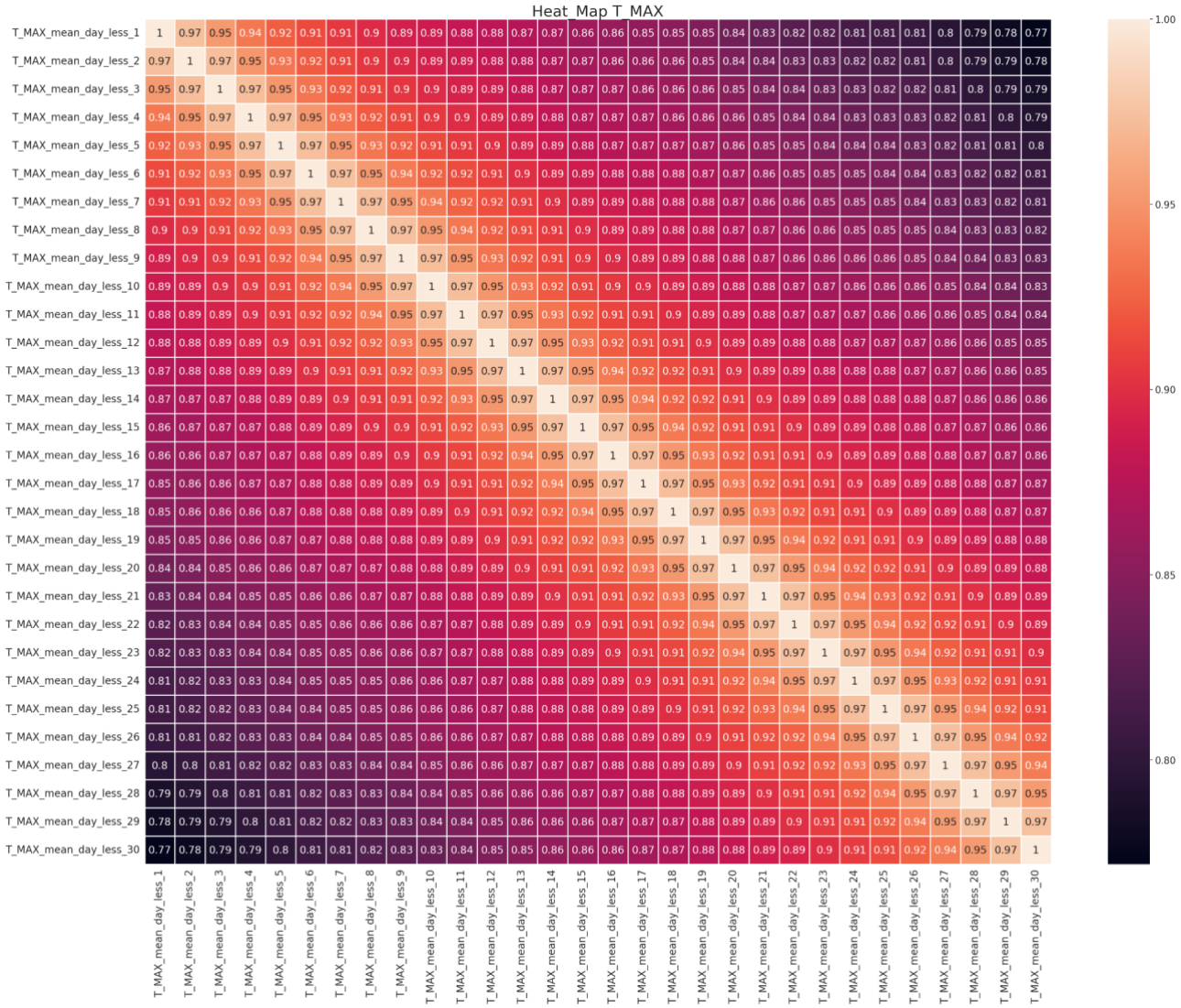

***Figure 2**. Correlation table of the Maximum temperature values among the data from 1 day before the FC and 30 days before the FC. High correlations are represented with brighter colors, instead low correlations are represented with darker colors.*

Data were regrouped according to a range specific for each climate variable, and the sum the values were used, in this case. In detail, the gap is 5 days for the temperature, 3 days for the relative humidity, 2 days for the wind speed and 2 days for the cloud coverage. The data from the precipitation aren't regrouped. The list of the variables used in this analysis is reported in Table 1.

***Table 1**. Feature used in the ML model.*

Variable	**Feature**
Cloud Coverage	'somma_CLCT_mean_1-2', 'somma_CLCT_mean_3-4', 'somma_CLCT_mean_5-6', 'somma_CLCT_mean_7-8', 'somma_CLCT_mean_9-10', 'somma_CLCT_mean_11-12', 'somma_CLCT_mean_13-14', 'somma_CLCT_mean_15-16', 'somma_CLCT_mean_17-18', 'somma_CLCT_mean_19-20', 'somma_CLCT_mean_21-22', 'somma_CLCT_mean_23-24', 'somma_CLCT_mean_25-26', 'somma_CLCT_mean_27-28', 'somma_CLCT_mean_29-30',
Wind Speed	'somma_WS_KMH_1-2', 'somma_WS_KMH_3-4', 'somma_WS_KMH_5-6', 'somma_WS_KMH_7-8', 'somma_WS_KMH_9-10', 'somma_WS_KMH_11-12', 'somma_WS_KMH_13-14', 'somma_WS_KMH_15-16', 'somma_WS_KMH_17-18', 'somma_WS_KMH_19-20', 'somma_WS_KMH_21-22', 'somma_WS_KMH_23-24', 'somma_WS_KMH_25-26', 'somma_WS_KMH_27-28', 'somma_WS_KMH_29-30',
Maximum relative humidity	'somma_RH_MAX_1-3', 'somma_RH_MAX_4-6', 'somma_RH_MAX_7-9', 'somma_RH_MAX_10-12', 'somma_RH_MAX_13-15', 'somma_RH_MAX_16-18', 'somma_RH_MAX_19-21', 'somma_RH_MAX_22-24', 'somma_RH_MAX_25-27', 'somma_RH_MAX_28-30',
Minimum relative humidity	'somma_RH_MIN_1-3', 'somma_RH_MIN_4-6', 'somma_RH_MIN_7-9', 'somma_RH_MIN_10-12', 'somma_RH_MIN_13-15', 'somma_RH_MIN_16-18', 'somma_RH_MIN_19-21', 'somma_RH_MIN_22-24', 'somma_RH_MIN_25-27', 'somma_RH_MIN_28-30'
Maximum temperature	'somma_T_MAX_1-5', 'somma_T_MAX_6-10', 'somma_T_MAX_11-15', 'somma_T_MAX_16-20', 'somma_T_MAX_21-25', 'somma_T_MAX_26-30'
Minimum temperature	'somma_T_MIN_1-5', 'somma_T_MIN_6-10', 'somma_T_MIN_11-15', 'somma_T_MIN_16-20', 'somma_T_MIN_21-25', 'somma_T_MIN_26-30'
Daily accumulated precipitation	'TOT_PREC_mean_day_less_1', 'TOT_PREC_mean_day_less_2', 'TOT_PREC_mean_day_less_3', 'TOT_PREC_mean_day_less_4', 'TOT_PREC_mean_day_less_5', 'TOT_PREC_mean_day_less_6', 'TOT_PREC_mean_day_less_7', 'TOT_PREC_mean_day_less_8', 'TOT_PREC_mean_day_less_9', 'TOT_PREC_mean_day_less_10', 'TOT_PREC_mean_day_less_11', 'TOT_PREC_mean_day_less_12', 'TOT_PREC_mean_day_less_13', 'TOT_PREC_mean_day_less_14', 'TOT_PREC_mean_day_less_15', 'TOT_PREC_mean_day_less_16', 'TOT_PREC_mean_day_less_17', 'TOT_PREC_mean_day_less_18', 'TOT_PREC_mean_day_less_19', 'TOT_PREC_mean_day_less_20', 'TOT_PREC_mean_day_less_21', 'TOT_PREC_mean_day_less_22', 'TOT_PREC_mean_day_less_23', 'TOT_PREC_mean_day_less_24', 'TOT_PREC_mean_day_less_25', 'TOT_PREC_mean_day_less_26', 'TOT_PREC_mean_day_less_27', 'TOT_PREC_mean_day_less_28', 'TOT_PREC_mean_day_less_29', 'TOT_PREC_mean_day_less_30'

Animal phenotype (target variable - residual form the linear model) and climatic (features) data were merged into a unique dataset to be used in the subsequent ML analyses. The Machine learning workflow before presented was applied. This workflow uses h2o.ai, an Open Source, Distributed, Fast & Scalable Machine Learning Platform. The first step of the ML pipeline is to find the best family algorithm. The best model selected for all the three phenotypes is a Gradient Boost Machine (GBM). As an example, we are reporting the best first 10 models for milk yield analyses (Table 2).

***Table 2**. Best model found when testing different models for milk yield.*

RANK	model ID	RMSE*	MSE*	MAE*	Mean residual deviance
1	*GBM_grid_1_AutoML_2_20230831_91425_model_15*	3,84	14,7	2,91	14,7
2	*GBM_grid_1_AutoML_2_20230831_91425_model_5*	3,84	14,7	2,91	14,7
3	*GBM_grid_1_AutoML_2_20230831_91425_model_10*	3,84	14,7	2,91	14,7
4	*GBM_grid_1_AutoML_2_20230831_91425_model_4*	3,84	14,7	2,91	14,7
5	*XRT_1_AutoML_2_20230831_91425*	3,84	14,8	2,91	14,8
6	*GBM_grid_1_AutoML_2_20230831_91425_model_8*	3,84	14,8	2,91	14,8
7	*DRF_1_AutoML_2_20230831_91425*	3,84	14,8	2,91	14,8
8	*GBM_grid_1_AutoML_2_20230831_91425_model_26*	3,85	14,8	2,92	14,8
9	*GBM_grid_1_AutoML_2_20230831_91425_model_1*	3,85	14,8	2,92	14,8
10	*GBM_1_AutoML_2_20230831_91425*	3,85	14,8	2,92	14,8

* Root Mean Square Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used to evaluate different model accuracy.

Once we have found the best algorithm for each phenotype, a grid search was performed to optimize the algorithm hyperparameter. This ML model with the optimized hyperparameter was used to obtain the feature importance as shown in Table 3.

***Table 3**. Top 10 most important features for milk yield*

Feature	**Proportion of importance**
somma_T_MIN_1-5	0.0423
somma_T_MAX_1-5	0.0269
somma_T_MIN_26-30	0.0177
somma_WS_KMH_5-6	0.0157
somma_WS_KMH_11-12	0.0156
somma_WS_KMH_3-4	0.0152
somma_WS_KMH_1-2	0.0152
somma_WS_KMH_15-16	0.0148
somma_WS_KMH_7-8	0.0148
somma_CLCT_mean_5-6	0.0144

Based on the features order, the following step consists in finding the optimal number of features to use in the ML model. The MAE (mean absolute error) was used with the objective to minimize it (Figure 3). In this case, the first 4 features were identified as the most important ones.

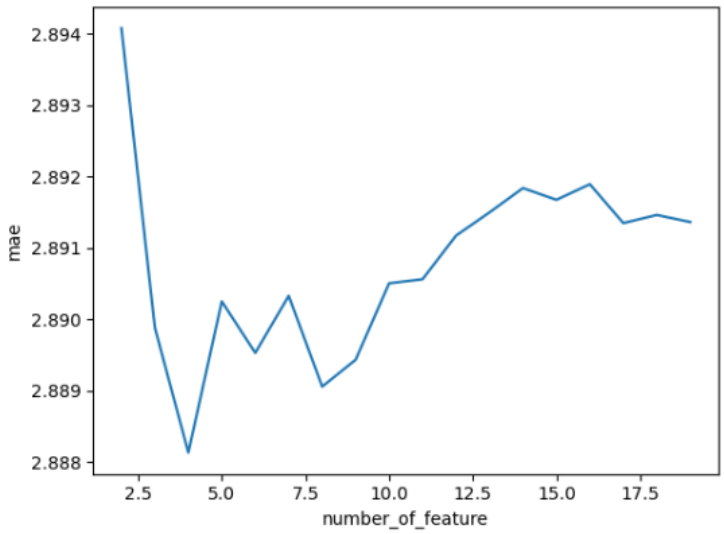

***Figure 3**. Identification of the most important using the MAE to evaluate each model in milk yield*

Once the optimal number of features is identified, the ML model was trained to obtain the prediction model. In Table 4 are reported the information about the ML analyses for milk yield, fat and protein with the climatic variable selected for the ML model, and in Table 5 the features selected.

***Table 4**. Identification and evaluation of the best Machine Learning model using the climatic variable as values.*

Feature	Algorithm	Proxy	RMSE*	MAE*	R-squared*	Number of features**
Milk yield	Gradient Boosting Machine	Production	2.8963	2.6797	0.1979	4
Fat	Gradient Boosting Machine	Milk quality	0,3988	0.3778	0.1836	6
Protein	Gradient Boosting Machine	Milk quality	0.2101	0.1517	0.2304	7

* Root Mean Square Error (RMSE), R-squared and Mean Absolute Error (MAE) are used to predict the model accuracy. They provide an estimate of the typical magnitude of prediction errors. Lower values indicate better model performance.
** Number of features selected for the model.

***Table 5**. Feature (climatic variable) selected for each phenotypes.*

Variable	Feature selected
Milk yield	somma_T_MIN_1-5, somma_T_MAX_1-5, somma_T_MIN_26-30, somma_WS_KMH_5-6
Fat	somma_T_MAX_1-5, somma_T_MAX_6-10, somma_WS_KMH_3-4, somma_WS_KMH_15-16, somma_WS_KMH_21-22, somma_WS_KMH_1-2
Protein	somma_T_MAX_1-5, somma_T_MIN_1-5, somma_WS_KMH_27-28, somma_WS_KMH_7-8, somma_WS_KMH_1-2, somma_WS_KMH_29-30, somma_WS_KMH_19-20

Once the algorithm performance was evaluated, the SHAP analysis was performed to identify and explain the most important variables. The SHAP plot (Figure 4), explains which contribution each feature (in this case, the most important features involved in predicting the milk yield content) gives to the plot. For example, high values of the minimum value of "Somma_T_min_1-5" days before the functional control have a high negative contribution to the prediction, while low values have a high positive contribution.

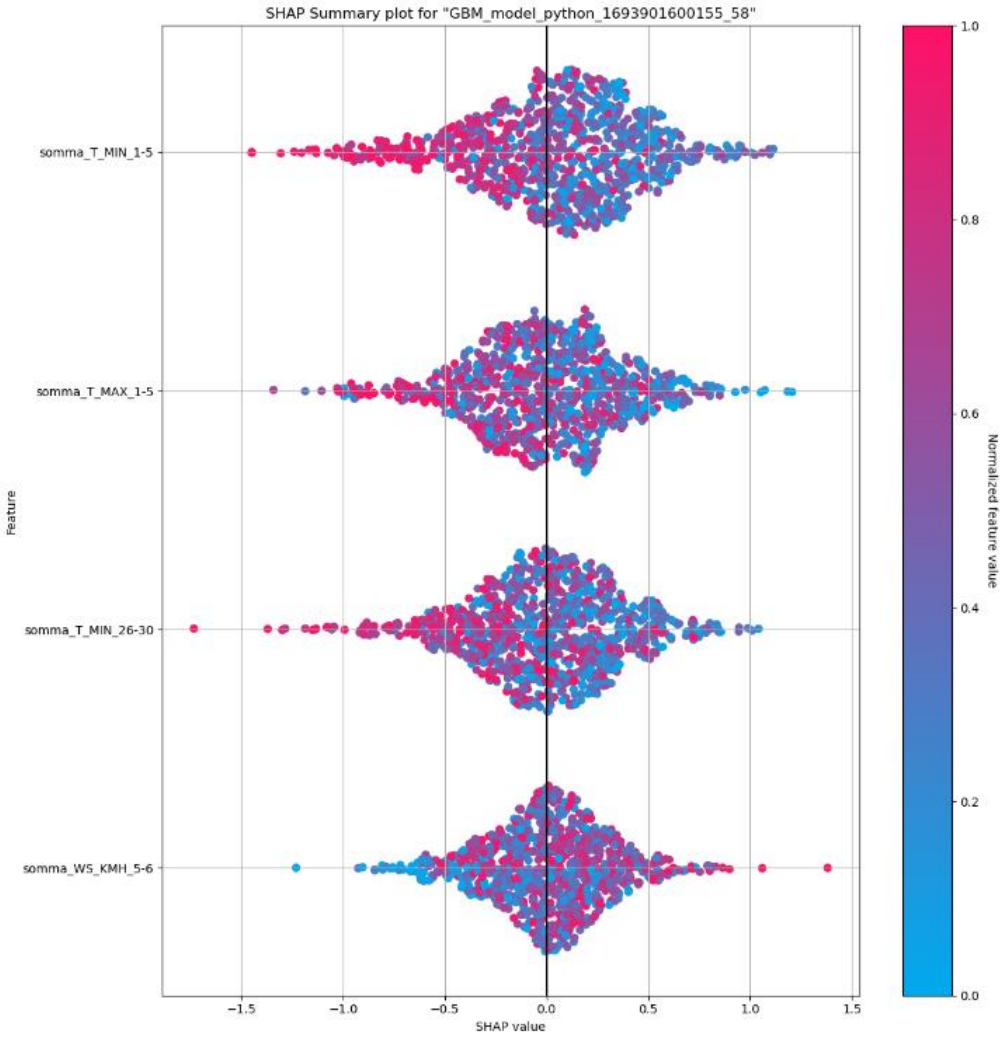

Figure 4. *Identification of the most important features involved in the prediction of the target variable "milk yield production". The variables are reported on the y-axis and are sorted to the most important (at the top*

of the graph), to the less important (on the lower part of the graph). On the x-axis, the SHAP value is reported. Each dot represents a sample used in the test set. Each sample is colored according to the corresponding normalized feature value.

The dataset used in this experiment has a huge number of points, which is the combination between the features (almost 100) and phenotypic data (more than 2.5 million). This dataset (reaching almost 0.25 billion points) could be considered for the size as a big data dataset [9] . In this case, ML approaches are more efficient than classical ones, because they are already designed for big data datasets. Moreover, we expect a nonlinear relationship among the data, with hidden structure which is hard to be revealed using traditional linear models [10].

5.1.2. Service 1.b

5.1.2.1. Introduction

Cattle are highly susceptible to heat stress, a condition that occurs when their core body temperature rises above the threshold at which they can effectively dissipate heat. Heat stress not only compromises the health and welfare of the animals but also leads to reduced feed intake, lower milk production, decreased fertility, and even mortality in severe cases [11]. Different cattle breeds exhibit varying heat tolerance levels, with some adapted to hot climates while others requiring more intensive management. The THI is a crucial parameter in livestock management, assessing the impact of temperature and humidity on animal comfort and often used as a proxy to identify heat stress in cattle [12].

Considering this, in service 1.b we developed a tool integrating THI data predictions, for external (pasture) and internal conditions, and breed-specific heat tolerance information, helping farmers make informed breed selection decisions in anticipation of the increasing of THI average values in following decades.

5.1.2.2. Data used

To identify THI tolerance thresholds for dairy and beef cattle, as well as specific breeds, a comprehensive literature review was conducted. This review involved an extensive search of available research papers, studies, and publications pertaining to THI tolerance levels in cattle. According to the literature revised the thresholds reported in Table 6 have been identified.

Table 6*. THI thresholds for bovine, according to literature*

Name	Group	Thr* no stress	Thr moderate stress	Thr high stress	Thr extreme stress	Reference
Beef	*purpose*	THI <72	72 >= THI < 82	82 >= THI < 94	THI >=94	[13]

Dairy	*purpose*	THI <72	72 >= THI < 79	79 >= THI < 89	THI >=89	[13]
Holstein	*breed*	THI <72	72 >= THI < 79		THI >=79	[14]
Jersey	*breed*	THI <72	72 >= THI < 90		THI >=90	[14]
Brown Swiss	*breed*	THI <72	72 >= THI < 83	83>= THI < 89	THI >=89	[15]

* The: threshold

At each geographical location, for each breed for which information on THI tolerance thresholds was available, we created a color-coded system ranging from green to red, where green means a THI value in which the breed is not threatened by heat (Table 6 - the no stress) and red meaning the breed is exposed to severe heat stress (Table 6 - extreme stress). This system is designed to enable users to determine whether a chosen breed will be suitable for farming in the future decades in a specifically selected geographic location based on projected future THI values.

5.1.3. Service 1.c

5.1.3.1. Introduction

The scope of this service is to identify animals more genetically resilient to adverse environmental conditions which could transmit this characteristic to future generations. To reach this objective, a collaboration has been set up with the Italian Red Pied cattle Association (ANAPRI), to estimate stress-resilience EBVs applying ANAPRI models to the subset of phenotypic data collected in stressful environmental conditions, in terms of THI (inside the barn). In this way animals can be ranked according to this new EBV that will possibly not match the exact ranking with current EBVs. Farmers and breeding centers will therefore have complementary information on the genetic potential of sires and dams in average environmental conditions (current EBV regularly calculated by ANAPRI) and under stress condition (stress-resilience EBVs). The latter will be used by those farmers that like to anticipate the effects of climate change, breeding animals in their farm towards robustness and resilience.

5.1.3.2. Data used in the service implementation

The analyses and the results from Service 1.a, using THI values from inside the barns (see service 2), were used to detect the most important THI variable and relative threshold to select only the phenotypes collected in days above the selected THI threshold (i.e., stressful days), from the total of FC.

5.1.3.3. Model developed

The subset of the phenotypic data will be used by ANAPRI to calculate the EBVs using the same model they routinely use. In particular, the IDAS ("Indice Doppia Attitudine Sostenibile" - Sustainable Double Purpose Index) index (https://www.anapri.eu/it/indice-idas.html) will be calculated. The index calculation is a prerogative by law of the breeders' associations. For this reason, details about the model used cannot be disclosed, the EBV calculation will be done by ANAPRI and the results will be available through a Sebastien link to ANAPRI website.

5.2. Service 2

5.2.1. Introduction

This service is devoted to the evaluation of the THI computed inside stables. The results will be used in service 1.a to evaluate the heat stress of the cattle inside stables.

More in detail, the approach used to evaluate the THI is through the application of an AI procedure able to implement the relationship between some input parameters and the related THI value. Specifically, the considered input drivers are: i) the latitude of the stable, ii) the altitude of the stable, iii) the external THI considered at the location (lat, lon) nearest to the stable. More parameters will be taken into account in the next period of the project with the aim to optimize the accuracy of the resulting internal THI.

5.2.2. Service 2a

The first subservice of the Service 2 has the objective to evaluate the variation of THI inside a stable for the next two days, with a temporal resolution of 1 hour. Given the latitude and altitude of the stable and the external THI value, the developed ML approach is able to extract the possible THI inside the stable.

5.2.2.1. Data used in the model implementation

First of all, two stages can be distinguished in the ML workflow adopted for Service 2: training phase and inference phase.

For the training phase, data concerning 677 different stables scattered throughout Italy were employed. Specifically, each stable was hourly monitored by a control unit which measured the internal temperature and relative humidity needed to compute the THI. These data about the THI index internal to stables have a nearly hourly temporal resolution and were exploited as the ground truth of the ML models [D2.2, AW 2].

Moreover, for developing the ML learning approach, it was necessary to collect data concerning the THI index external to these 677 stables. For this purpose, hourly data derived from the ERA5-Land reanalysis were employed. Specifically, 2m temperature and 2m dewpoint-temperature (turned into relative humidity) were the variables downloaded from ERA5-Land (horizontal resolution ~ 9 Km), then statistically downscaled on the COSMO-2I grid, which has a horizontal resolution of about 2.2 Km. Therefore, THI index on the COSMO-2I grid was finally computed.

Since during the training phase the ML model must learn an inherent mapping between the THI index external to stables (from now on called 'external THI') and the one internal to stables (from now on called 'internal THI'), different preprocessing steps were necessary. Indeed, for each stable it was associated, as external THI value, the one referred to the closest point on the COSMO-2I grid by taking into account the geographical coordinates (longitude, latitude) of the stable. Therefore, a geographical matching was performed between each stable location and a COSMO-2I grid point. Instead, regarding the internal THI index data, preprocessing steps included an outlier detection analysis based on the Z-score method and a temporal alignment that was needed to exactly fit data to the hourly temporal resolution of the external data.

Finally, internal data were combined with the external ones by ensuring a matching of the records based on the stable ID, the date, and the hour to provide a unified dataset.

5.2.2.2. Model developed

Once trained on supervised data, the ML model should have learnt the relationship between the external THI, latitude, altitude of the stable and the internal THI index in order to produce, during the inference phase, an estimate of the internal THI value for an unseen stable. Thus, during the training phase the ML model takes as input the external THI index, as well as the stable latitude and altitude, whereas the target variable is the internal THI. Eventually, other useful features could be added among the predictors of the ML model.

After shuffling the unified dataset, a subset of 30000 records was considered for the training phase and it was split into the training set (80%) and the test set (20%). The search of the best ML algorithm family was performed by exploiting the H2O.ai AutoML [5], a Python module which tests different ML algorithms (such as Random Forest, XGBOOST, Gradient Boosting Machine – GBM) and provides their performance in term of different metrics. Since the target is a continuous variable, this case study is configured as a regression task and the metric used to evaluate the performance of the ML model is the Root Mean Squared Error (RMSE). The results showed that the family of ML algorithms which provided the best accuracy was GBM with a RMSE equal to 4.23, as reported in Table 7. This result could be improved by adding new predictors and increasing the training set.

Table 7. *Best model found when testing different models for estimating internal THI.*

Model ID	RMSE	MSE	MAE
GBM_3_AutoML_1	4.22998	17.8927	3.17434

GBM_4_AutoML_1	4.23564	17.9407	3.17343
GBM_2_AutoML_1	4.26234	18.1675	3.20364
XGBoost_grid_1_AutoML_1_model_3	4.27493	18.2751	3.19477
GBM_5_AutoML_1	4.2808	18.3253	3.21918
XGBoost_3_AutoML_1	4.28128	18.3294	3.20814
XGBoost_grid_1_AutoML_1_model_2	4.29637	18.4588	3.18392
XGBoost_grid_1_AutoML_1_model_1	4.32608	18.7149	3.2173
XGBoost_2_AutoML_1	4.3374	18.8131	3.23283
GBM_grid_1_AutoML_1_model_2	4.37078	19.1037	3.30344
XRT_1_AutoML_1	4.37884	19.1743	3.29377
GBM_1_AutoML_1	4.38807	19.2552	3.3071
GBM_grid_1_AutoML_1_model_1	4.38812	19.2556	3.34482
DRF_1_AutoML_1	4.39	19.2721	3.3041
XGBoost_1_AutoML_1	4.40224	19.3797	3.26472
DeepLearning_grid_2_AutoML_1_model_1	4.88621	23.875	3.71833
DeepLearning_grid_1_AutoML_1_model_1	4.91193	24.1271	3.75584
DeepLearning_1_AutoML_1	4.93045	24.3093	3.77118
GLM_1_AutoML_1	5.23151	27.3687	4.09632

5.2.3. Service 2b

The second subservice will support users by assessing the potential variations of THI inside a stable due to climate change for near- and long-time horizons under IPCC-RCP4.5 and RCP8.5 scenarios. The data will display as expected changes between 30-years future periods and a specific baseline (1981-2010). By tuning different parameters (e.g., types of conditioning systems and stable arrangements), users could plan adaptation actions and design the optimal setup of new stables in a climate-proof concept, reducing then the stress due to the expected future increases in temperature on animals. The possible THI inside the stable is evaluated by exploiting the same ML approach developed for Service 2a.

5.2.3.1. Data used in the model implementation

Climate projection used to develop the second subservice derives from VHR-PRO_IT (Very High-Resolution PROjections for Italy; [7]), an open access hourly climate projection with a resolution of ≃2.2 km from 1981 up to 2070, covering the Italian peninsula and some neighboring areas. VHR-PRO_IT was produced within the Highlander project (https://highlanderproject.eu/) by dynamically downscaling the Italy 8km-CM climate projection (spatial resolution ≃8 km; output frequency = 6 h; driven CMIP5 GCM = CMCC-CM) with the Regional Climate Model COSMO-CLM. Its global forcing is the historical experiment driven by the observed natural and anthropogenic atmospheric

composition for 1981–2005 and the RCP4.5 and RCP8.5 greenhouse gas concentration trajectories for 2006–2070.

5.2.3.2. Model developed

The model used in service 2a will be also applied in service 2b. In these terms, this model is able to identify a relationship between the input variables (latitude, altitude and external THI) with the internal THI of the stable. This relationship is valid regardless of the time at which the input variables are considered. Consequently, by applying the considered model it is possible to find the corresponding value of internal THI also considering as input variables climatic values (external THI) extracted from future climate projections. Data will be presented as expected changes between 30-years future periods and a specific baseline (1981-2010). Such an approach allows us to provide expected variations with respect to the baseline, avoiding bias correction procedures that at the hourly resolution and for a huge amount of grid points is very expensive.

5.3. Service 3

5.3.1. Introduction

The management of an extensive farm is not easy: it is not possible to monitor the animals constantly and the feed availability (in terms of quantity and quality) is not easy to observe. To monitor the surrounding environments of extensive livestock farming, satellite data, under different spectral ranges and bands, are used to detect vegetation structure, status, and contents. This will allow the user to schedule and update grazing availability and detect possible overgrazing effects. In this service, satellite data were combined in a statistical model with pasture field data, to evaluate pasture productivity and characteristics.

5.3.2. Data used in the model implementation

Pasture data were collected in two different farms in the Lazio region. The pasture management system used is called “rational pasture”, where the animals are moved in different sub-area following the grass growth. This gives the possibility to have, at the same time, areas without grass cover and areas with optimal grass cover. In each sampling day, three areas for low, medium, and high levels of NDVI Sentinel2 index (a proxy of the total Sentinel2 data) were collected. In this way, the expectation is to have a distribution of value along the time. Fresh and dry matter biomass were collected, and laboratory evaluation (i.e., fiber characteristics, lignin, protein, fat, etc.) were done. Climate and topographic data were also considered important to correct shading/radiation/background effects. However, to our knowledge, they were not normally included in the models. For this reason, as a next step, we will include this information as fixed effects in the regression model or features in the ML models that we're going to implement.

The Sentinel2 satellite data (D2.2, AW 64) were used, in particular:

- bands: B2, B3, B4, B5, B6, B7, B8, B8A;
- indexes: NDVI (Normalized Difference MIR/NIR Normalized Difference Vegetation Index), NDWI (Normalized Difference Water Index), EVI (Enhanced Vegetation Index), GLI (Green leaf index), SAVI (Soil Adjusted Vegetation Index), GCI (Green Chlorophyll Vegetation Index), RGR (Simple Ratio Red/Green Red-Green Ratio), SIPI (Structure Insensitive Pigment Index), ARVI (Atmospherically Resistant Vegetation Index), NBRI (Normalized Burned Ratio Index).

5.3.3. Model developed

The prediction was tested starting from approaches already reported in bibliography. For example, we tested classical statistical approaches, such as linear regression as reported by Primi and colleagues [16]. Machine learning approaches were reported in literature, such as random forest by Bretas and colleagues [17]; however we decide, at this moment, to not apply them to this data because the number of data collected in the field are not enough to justify a ML approach.

For each phenotype (fresh and dry matter), all the single bands/index, all the bands, all the indexes and all the bands + indexes were tested in multiple linear regression models. For models with multiple fixed effects, only the significant ones were maintained in the final model. For each model, MAE, r-square, AIC (Akaike information criterion), and overall p-value of the model were saved. The model with higher r-square (i.e., more efficient prediction) was saved and will be used in the prediction. Here the results for the estimation of fresh biomass (Table 8) and dry matter biomass (Table 9).

***Table 8**: Model results for fresh (tq) biomass for hectare. In red the selected model.*

Phenotype	***Sentinel2***	***r2***	***AIC***	***MAE***	***p-value***	***Model****
Grass(tq/ha)	B2	0,2725	233,0451	1,1319	4,90E-06	-
Grass(tq/ha)	B3	0,2416	235,7511	1,1707	1,92E-05	-
Grass(tq/ha)	B4	0,2924	231,2447	1,1443	1,99E-06	-
Grass(tq/ha)	B5	0,2179	237,7525	1,1962	5,28E-05	-
Grass(tq/ha)	B6	0,1674	241,8195	1,1575	4,21E-04	-
Grass(tq/ha)	B7	0,2153	237,9671	1,1286	5,88E-05	-
Grass(tq/ha)	B8	0,2245	237,2016	1,1108	3,99E-05	-

Grass(tq/ha)	B8A	0,2038	238,9132	1,1217	9,52E-05	-
Grass(tq/ha)	NDVI	0,3621	224,506	1,046	6,91E-08	-
Grass(tq/ha)	NDWI	0,4005	220,472	1,0024	9,35E-09	-
Grass(tq/ha)	EVI	0,3339	227,3159	1,028	2,79E-07	-
Grass(tq/ha)	GLI	0,2621	233,9729	1,0838	7,82E-06	-
Grass(tq/ha)	SAVI	0,3504	225,6842	1,022	1,24E-07	-
Grass(tq/ha)	GCI	0,3028	230,2832	1,0275	1,23E-06	-
Grass(tq/ha)	RGR	0,2855	231,8758	1,0745	2,73E-06	-
Grass(tq/ha)	SIPI	0,2441	235,5374	1,1746	1,72E-05	-
Grass(tq/ha)	ARVI	0,3609	224,6284	1,0266	7,34E-08	-
Grass(tq/ha)	NBRI	0,3421	226,5084	1,0255	1,87E-07	-
Grass(tq/ha)	band	0,3392	227,7587	1,0365	9,89E-07	B2|B7
Grass(tq/ha)	index	0,4005	220,472	1,0024	9,35E-09	NDWI
Grass(tq/ha)	*band+index*	*0,4668*	*219,1987*	*0,8802*	*4,02E-07*	*B2|B3|B8|NDVI|NDWI|GLI|GCI|RGR*

* In the case of more than one band/index was used, here were reported the significant ones. For models with only one band/index used, the "-" is reported.

***Table 9**: Model results for dry matter (ss) biomass for hectare. In red the selected model.*

Phenotype	***Sentinel2***	***r2***	***AIC***	***MAE***	***p-value***	***Model****
Grass(ss/ha)	B2	0,0416	103,5844	0,3867	5,35E-02	-
Grass(ss/ha)	B3	0,0561	102,564	0,3802	3,00E-02	-
Grass(ss/ha)	B4	0,0342	104,1045	0,3863	7,25E-02	-
Grass(ss/ha)	B5	0,0347	104,0645	0,383	7,08E-02	-
Grass(ss/ha)	B6	-0,0121	107,2405	0,4022	6,49E-01	-
Grass(ss/ha)	B7	-0,004	106,7009	0,4011	3,94E-01	-
Grass(ss/ha)	B8	0,0031	106,2221	0,4006	2,76E-01	-

Grass(ss/ha)	B8A	0,0006	106,3941	0,401	3,12E-01	-
Grass(ss/ha)	NDVI	0,0375	103,8751	0,3863	6,34E-02	-
Grass(ss/ha)	NDWI	0,0626	102,1006	0,3775	2,32E-02	-
Grass(ss/ha)	EVI	0,0226	104,9024	0,3922	1,17E-01	-
Grass(ss/ha)	GLI	-0,0013	106,5202	0,3988	3,43E-01	-
Grass(ss/ha)	SAVI	0,0294	104,4359	0,3895	8,82E-02	-
Grass(ss/ha)	GCI	0,0266	104,6258	0,3867	9,88E-02	-
Grass(ss/ha)	RGR	0,0046	106,1262	0,3967	2,58E-01	-
Grass(ss/ha)	SIPI	0,0355	104,0113	0,3879	6,86E-02	-
Grass(ss/ha)	ARVI	0,0318	104,2701	0,3879	7,99E-02	-
Grass(ss/ha)	NBRI	0,0228	104,8899	0,3934	1,16E-01	-
Grass(ss/ha)	band	0,163	95,4742	0,3532	1,26E-03	B6|B8A
Grass(ss/ha)	index	0,1653	96,2301	0,3557	2,38E-03	NDVI|NDWI|GCI
Grass(ss/ha)	*band+index*	*0,2452*	*94,7825*	*0,3148*	*2,19E-03*	*B2|B4|B6|B8A|NDVI|GLI|GCI|SIPI|ARVI*

* In the case of more than one band/index was used, here were reported the significant ones. For models with only one band/index used, the "-" is reported.

The comparison between observed and predicted values for fresh grass using the regression model selected is reported in Figure 5.

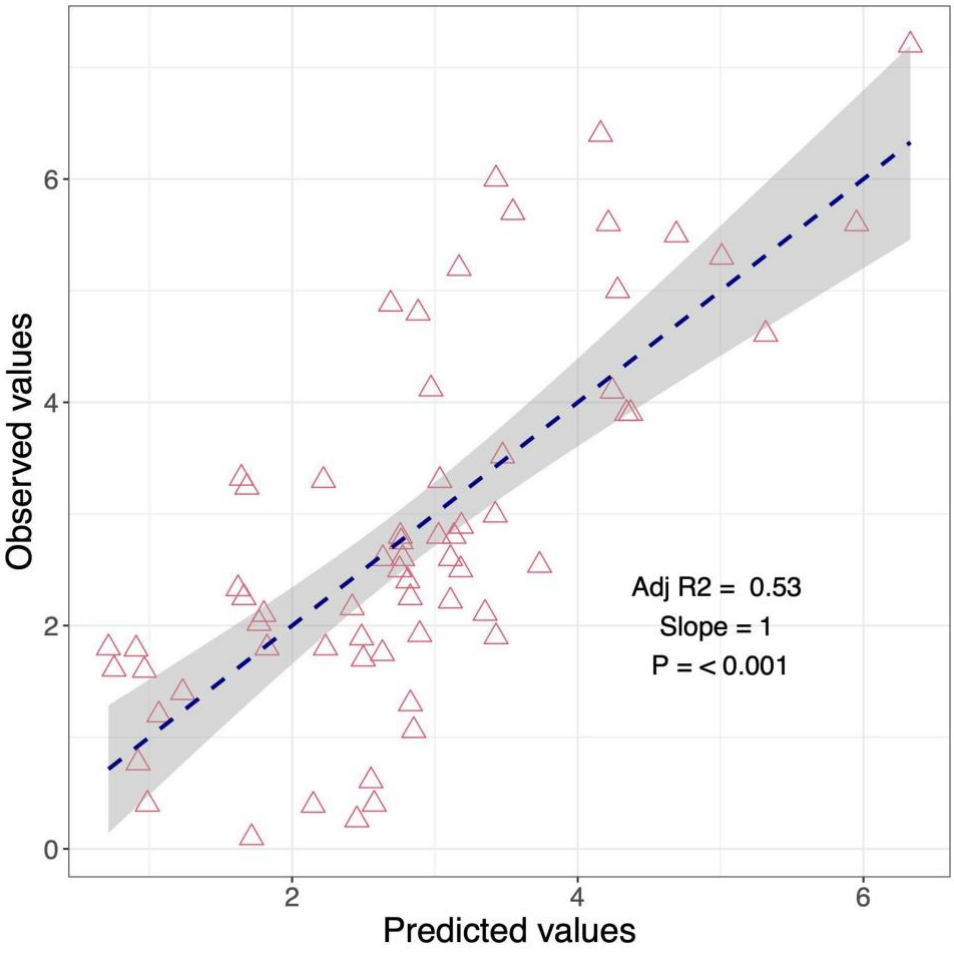

***Figure 5**: Observed versus predicted values in the model here developed to estimate the fresh grass biomass per hectare.*

At this moment, the r-squared values obtained in both the models are lower than expected (see the r-squared reported in [16], as an example), probably due to the lower number of field samples collected. As already mentioned, including other effects, such as climatic and topography ones, could improve the model accuracy.

5.4. Service 4

The Service 4 will help farmers and decision makers to contrast the spread of diseases from parasites (i.e., bluetongue in sheep) or health conditions (i.e. mastitis in bovine). Literature information about the environmental, climatic, farm management, geographical conditions, which can potentially favor or trigger them, will be used to obtain prediction models. Climate projections will give support in projecting possible shifts, in the future, of favorable conditions for parasites and diseases. The results of this service will be risk maps for parasites and diseases' spreads.

5.4.1. Service 4.a

In this service, Machine Learning approaches were used to produce risk maps of parasites and diseases spread, combining abiotic and biotic factors.
About the Sardinia case study, where bluetongue, a vector-borne disease transmitted by species of *Culicoides midges*, were modeled based on a logistic multilevel mixed model [18], the Machine Learning pipeline (chapter 4.1) was applied to improve the already existing prediction model.

5.4.1.1. Data used in the model implementation

To create the dataset, three different kinds of data (information related to the farm; climatic information; environmental information) were collected. The data were organized to create a dataset, which was then used to create the Machine Learning model.

The first type of data concerned information related to the farm located in Sardinia, such as latitude, longitude, and a unique farm ID. This information was provided by the Experimental Zooprophylactic Institute (IZS - "Istituto Zooprofilattico") of Sardinia. Furthermore, the number of animals present in the stable for each farm was reported, with the date of confirmed clinal suspects of bluetongue and the number of infected animals. The IZS also provided information regarding the vaccination practice, indicating the date of the vaccination for each farm. This date was used to understand if the animals were vaccinated before the clinical cases of bluetongue. If the date of vaccination precedes that of the clinical case, then the animals were considered vaccinated. Otherwise, they were not considered vaccinated. From this data, the probability of developing the disease was calculated by dividing the number of animals with confirmed pathology by the total number of animals on the farm. The cases report data from the year 2013. We used this year since there were numerous bluetongue cases in Sardinia and, consequently, was considered interesting data for developing the first ML model. In total, information was collected on 5600 companies.

The information on companies was used to obtain two other types of data, both essential in determining the life cycle of *Culicoides*, the vector of bluetongue. The types of data integrated with companies are: 1) climatic information and 2) environmental information. These types of data are crucial in understanding the environmental factors that can influence the spread of the disease.

Regarding the climatic information, these were organized in NetCDF files and downloaded from Highlander DDS, as already reported in previous services. In particular, the data derive from the **VHR-REA_IT** dataset, with a spatial resolution of 2.2 km, and only data for the Sardinia region were used [8]. The following information was downloaded: 1) mean temperature, 2) minimum temperature, 3) maximum temperature, 4) mean value of relative humidity, 5) maximum value of relative humidity, 6) minimum value of relative humidity, 7) cloud coverage, 8) precipitation, 9) wind speed and 10) solar radiation . Since the life cycle of *Culicoides* can be as long as two months [19], we decided to collect climatic information near the farm up to 60 days before the clinical case. For example, if a clinical case occurred on 06/30/2013, then climatic information was collected and analyzed up to 60 days before the clinical case (02/05/2013). The grid has a spatial resolution of 2.2

km, so the climatic variables were computed as the mean across the four nearest points to the farm. We used the latitude and longitude data associated with each farm to perform this computation. The mean was considered a good approximation of the climatic value near the farm because the *Culicoides* can move around 500 meters around the farm [18]. Each climatic variable was computed by evaluating its mean value in a range of 5 days, except for the precipitation, which was evaluated using the accumulated sum in a range of 5 days. To avoid the collinearity between the climatic variables, we computed the Pearson correlation coefficient among each pair of variables. Furthermore, high collinear variables were removed using the Variance Influence Factor, as previously shown by Cappai et al. [18].

Finally, the information related to the environmental characteristics of the environment across the farm were used. This information is organized in shape files that were downloaded from the geoportal website of the Sardinia region (http://webgis2.regione.sardegna.it/download). As previously mentioned, the *Culicoides* vector can fly in a range of 500 meters around the farm. Therefore, in some cases, it was easy to associate the environmental information of some farm, such as in the case of Figure 6A. In this case, the farm is located inside a polygon, which defines the environmental characteristics of the environment. However, in other cases, such the one reported in Figure 6B, the farm was characterized by several polygons in a range of 500 meters. In the latter case, the environmental characteristic more frequent inside this range was associated with the farm. This strategy was also performed by Cappai et al. [18].

***Figure 6**. Example of company localization in two different situations. In **A)** the farm is located, within a radius of 500 meters, within a polygon. Therefore, the environmental characteristics of the polygon are associated with the company. In **B)** the farm is located in different polygons within a radius of 500 meters. Therefore, the environmental characteristic associated with the company will be the most frequent one among the polygons within the radius of 500 meters.*

Finally, the three datasets (management, climatic and environmental) were merged. The dataset was used to perform ML analyses.

5.4.1.2. Model developed

The Machine Learning model was developed considering the probability of developing the disease as the target variable. The climatic, environmental, and management variables were used as predictors.

The approach used to create the ML models is described in 4.1. Firstly, the best family of ML algorithms was searched to identify the best approach to perform the predictions using these data. The following algorithms were tested: 1) Distributed Random Forest; 2) Generalized Linear Model; 3) XGBoost and 4) Gradient Boosting Machine. Using the autoML function in h2o package, 1080 models were tested. For each algorithm family, different models were trained and tested, by using different parameters. The results showed that two families provided the best accuracy: the XGBoost and GBM. We evaluated the accuracy by considering the MAE. An example of the first 10 models selected from the H2O library is reported in Table 10.

Table 10. *The table shows the first 10 results of the selection of algorithms tested by H2O. The Rank column reports the order of the model, while the "model_id" column reports the algorithm's name indexed by H2O. RMSE, MSE, MAE, and Mean Residual Deviance are reported for each model. The models are ordered in ascending order of MAE.*

RANK	model id	RMSE	MSE	MAE	Mean Residual Deviance
1	XGBoost_lr_search_selection_AutoML_1_20230912_181008_select_grid_model_3	21,44	459,78	16,25	459,78
2	GBM_grid_1_AutoML_1_20230912_181008_model_15	21,64	468,39	16,42	468,39
3	XGBoost_grid_1_AutoML_1_20230912_181008_model_56	21,77	474,05	16,43	474,05
4	GBM_grid_1_AutoML_1_20230912_181008_model_380	21,63	467,96	16,43	467,96
5	GBM_grid_1_AutoML_1_20230912_181008_model_528	21,63	468,0	16,43	468,05
6	XGBoost_grid_1_AutoML_1_20230912_181008_model_530	21,59	466,37	16,43	466,37

7	GBM_grid_1_AutoML_1_20230912_181008_model_585	21,62	467,81	16,44	467,81
8	GBM_grid_1_AutoML_1_20230912_181008_model_504	21,67	469,67	16,45	469,67
9	XGBoost_grid_1_AutoML_1_20230912_181008_model_205	21,77	474,08	16,46	474,08
10	XGBoost_grid_1_AutoML_1_20230912_181008_model_520	21,72	471,75	16,47	471,75

Even though the model with the lowest MAE is XGBoost, we decided to use the GBM algorithm, which simultaneously elaborates numeric (i.e., climatic variables) and categorical (i.e., environmental information). This allowed us to interpret many of the environmental variables, which turn out to be categories. We performed a Grid Search to tune the parameters of the algorithm. Using the h2o packages, the grid search was performed for 24 hours and 1400 different GBM models were evaluated. The parameters model with the lowest MAE were selected. All the operations were performed using a k-fold (k=5) cross-validation (80% of the data were used as training set, while 20% were used as test set). The importance of the variables was obtained, and the model was trained again iteratively. The model was trained using the first n-th important features, sorted by the most to the last important. An example of the first 10 variables are reported in Table 11.

Table 11. *The table reports the importance of the top 10 variables identified by the GBM algorithm. Each variable's name and importance (relative and scaled) are reported. The percentage (on a scale from 0 to 1) of the variable's importance is also provided. Finally, a brief description of the variable is given.*

Rank	Variable name	Relative importance	Scalde importance	Percentage	Variable description
1	TMAX_2M_DAILY_MEAN_minus_0_4	4861498,5	1	0,116	Climatic variable. Maximum temperature from 0 to 4 days before the clinical case date.
2	Soil_map_Sardinia_Region	3856497,5	0,79	0,092	Environmental variable. Soil typology of the Sardinia Region.
3	Land_use_Sardinia_Region	3343237,3	0,68	0,080	Environmental variable. Land use of the Sardinia Region.
4	T_2M_DAILY_MEAN_minus_0_4	1948211,9	0,40	0,046	Climatic variable. Average temperature from 0 to 4 days before the clinical case date.
5	Number_of_animals	1696118,1	0,34	0,040	Company information. Number of animals present on

					the farm.
6	RH_MIN_2M_DAILY_MEAN_minus_35_39	1482198,6	0,30	0,035	Climatic variable. Minimum humidity from 35 to 39 days before the clinical case date.
7	ASOB_S_DAILY_MEAN_minus_15_19	1423922,6	0,29	0,034	Climatic variable. Solar radiation from 15 to 19 days before the clinical case date.
8	ASOB_S_DAILY_MEAN_minus_30_34	1108334,1	0,22	0,026	Climatic variable. Solar radiation from 30 to 34 days before the clinical case date.
9	U_10M_DAILY_MEAN_minus_0_4	982941,3	0,20	0,023	Climatic variable. "U component" of the wind from 0 to 4 days before the clinical case date.
10	U_10M_DAILY_MEAN_minus_15_19	921032,5	0,18	0,022	Climatic variable. "U component" of the wind from 15 to 19 days before the clinical case date.

The trend of MAE as a function of the number of variables used is reported in Figure 7. A model with 43 variables provided the same MAE of the full model (67 variables). Therefore, we removed the 24 least important and uninformative variables from the model.

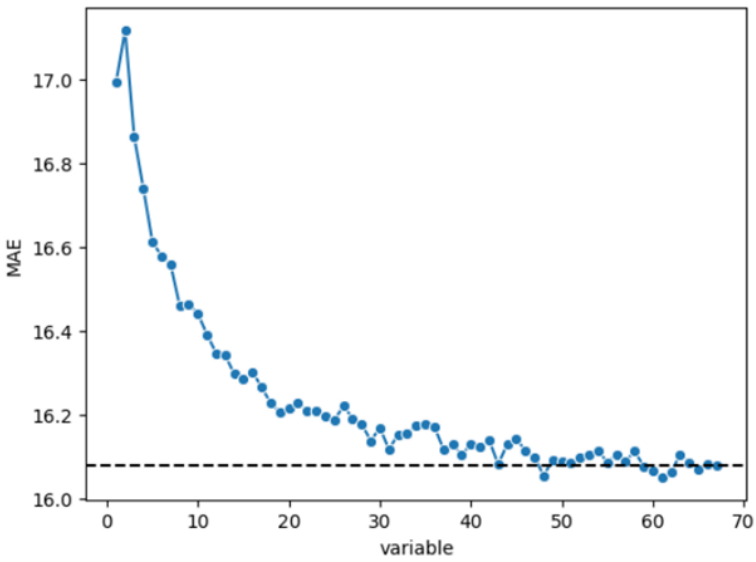

Figure 7. *The following figure shows the trend of MAE on the y-axis, while the number of variables used to train and test the model is shown on the x-axis. The model was trained in several iterations with increasing features. The variable added at each iteration was selected by observing their importance with the total model. The graph shows that using 43 variables, a comparable MAE is obtained to that obtained with the complete model.*

The selected variables are the following:

- 2 environmental variables: soil typology and land use;
- 1 variable related to farm information: number of livestock present;

- 3 variables related to the average maximum temperature recorded in the following time frames before the clinical case: from 0 to 4 days, from 20 to 24 days, and from 35 to 39 days;
- 4 variables related to the average temperature recorded in the following time frames before the clinical case: from 0 to 4 days, from 10 to 14 days, from 25 to 49 days, and from 45 to 49 days;
- 6 variables related to the average minimum temperature recorded in the following time frames before the clinical case: from 0 to 4 days, from 5 to 9 days, from 15 to 19 days, from 40 to 44 days, from 50 to 54 days, and from 55 to 59 days;
- 5 variables related to the average maximum humidity recorded in the following time frames before the clinical case: from 15 to 19 days, from 30 to 34 days, from 35 to 39 days, from 45 to 49 days, and from 55 to 59 days;
- 3 variables related to the average humidity recorded in the following time frames before the clinical case: from 0 to 4 days, from 15 to 19 days, and from 35 to 39 days;
- 3 variables related to the average minimum humidity recorded in the following time frames before the clinical case: from 0 to 4 days, from 25 to 29 days, and from 35 to 39 days;
- 6 variables related to the averaged surface net downward shortwave radiation recorded in the following time frames before the clinical case: from 0 to 4 days, from 15 to 19 days, from 30 to 34 days, from 40 to 44 days, from 50 to 54 days, and from 55 to 59 days;
- 6 variables related to the U component of the wind recorded in the following time frames before the clinical case: from 0 to 4 days, from 5 to 9 days, from 10 to 14 days, from 15 to 19 days, from 35 to 39 days, and from 55 to 59 days;
- 2 variables related to the cloud coverage recorded in the following time frames before the clinical case: from 30 to 34 days and from 40 to 44 days;
- 2 variables related to the cumulative sum of precipitation recorded in the following time frames before the clinical case: from 15 to 19 days and 25 to 29 days.

5.4.2. Service 4.b

The objective of service 4.b service is to study the somatic cells (somatic cell count - SCC) that are normally present in milk. SCC is an index used to estimate mammary gland health and milk quality. The number of somatic cells in milk is affected by different factors (e.g. the animal's health, lactation stage, breed) [20] and an increase is associated with changes in the environmental conditions and stress conditions. If mastitis, a mammary gland inflammation, is present, SCC will greatly increase. Today mastitis remains one of the most important diseases for the worldwide dairy industry [21]. Today to reduce the mastitis effect, antibiotics are used. However, there are losses in milk production, an increase in cost and general sanitary problems (antibiotic resistance) associated with this practice. In view of the correlation between mastitis and SCC, the last is

currently used as a proxy to control mastitis. Here, in particular, we evaluate the effect of environmental stressful conditions (heat stress) on SCC.

5.4.2.1. Data used to develop the model

The data and pipelines used in Service 1.a were also used here. The data from the SCC were transformed using a base 10 logarithm to obtain a similar normal distribution of the values.

5.4.2.2. Model developed

The ML model used is a GBM as shown in Table 12.

***Table 12**. Identification and evaluation of the best Machine Learning model to be applied in the Service 4.*

Feature	Algorithm	Proxy	RMSE	MAE	R-squared	Range
SCC	Gradient Boosting Machine	Health	0.4533	0.3468	0.0689	4

The four selected variables are: "*somma_WS_KMH_29-30", "somma_WS_KMH_1-2", "somma_WS_KMH_5-6"*, and *"somma_WS_KMH_15-16".*

In this ML model only the wind speed is reported as the most important climatic feature. In Figure 8 are reported the SHAP results. A short- and long-term effect was identified, showing that stressful conditions have acute but also "chronic" effects. Interestingly, the temperature, which is always present in the other ML model of service 1A, is not present in this model as one of the most important features. This may be associated with the nature of the phenotype (i.e., associated with cattle health rather than production). However, further analyses need to be performed to confirm the obtained results.

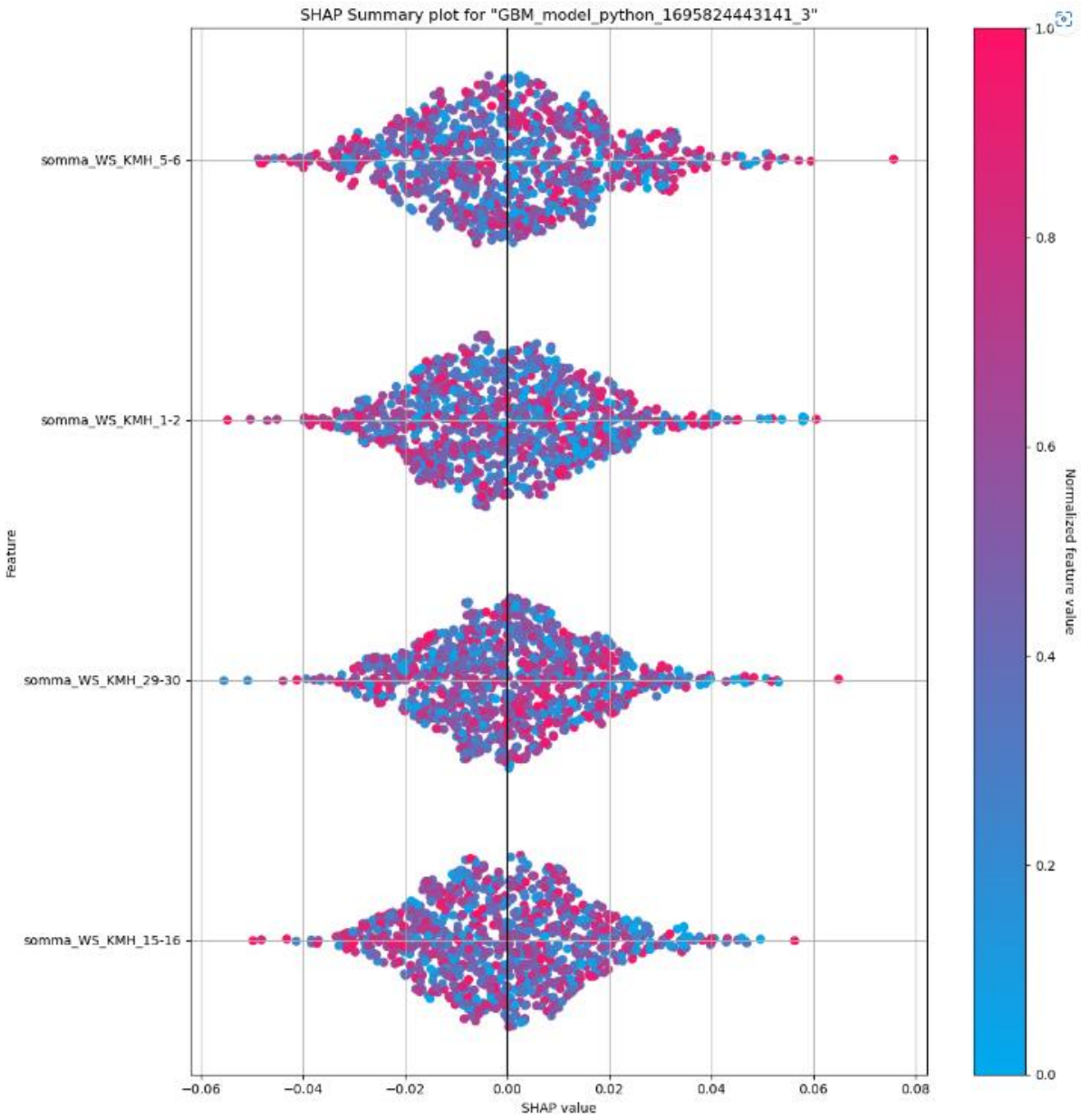

Figure 8. *Identification of the most important features involved in the prediction of the target variable "Somatic cell count". The variables are reported on the y-axis and are sorted to the most important (at the top of the graph), to the less important (on the lower part of the graph). On the x-axis, the SHAP value is reported. Each dot represents a sample used in the test set. Each sample is colored according to the corresponding normalized feature value.*

5.5. Animal and environmental sensors

5.5.1. Introduction

In the SEBASTIEN project, environmental (D2.2, AW 67) and animal (D2.2, AW 66) sensors were developed and used to collect data in real-time related to animals (in barns but also in pasture) and to barns. The sensor monitoring gives important information to identify stressful situations due to, for example, heat stress conditions, poor pasture quality or health problems.

5.5.2. Data acquired and analyses performed

The animal sensor consists of a collar that can collect the information needed for assessing the animal welfare and make them available remotely. In particular, the data collected are movements, ambient temperature, and relative humidity, GNSS position and heart rate.

The environmental sensor consists of a platform that is built to measure the concentration of some gasses in the air, in particular: CO2, H2S, NH3, CH4. PM1, PM2.5, PM10, temperature and relative humidity.

Most of the parameters from animal sensors are representative of animal welfare on their own. However, some elaborations are needed to have clearer and more useful information. For example, about the movement data, by analyzing the movement variation it is possible to understand how much the animal is active. Variation in this could be used to detect anomalies or, in the case of female animals, estrus period. Evaluation of the heart rate could be used itself to detect anomalous behavior for the single animal or for the herd (i.e., if all the animals present an anomalous heart rate the cause could be environmental. Finally, the THI around the animal can be calculated. THI could be also calculated from the environmental sensors. In general, warnings can be set up to alert the breeder to any problems, anomalies, or special conditions associated with a specific animal or environment.

6. Conclusion

Here the methodologies applied to develop the models, which will be used in the SEBASTIEN services were presented. In many cases, innovative Machine Learning approaches were developed to respond to demands that still have no clear answer nowadays in the livestock system. Data already available through previous (i.e., Highlander) and actual (LEO from AIA partner) projects or newly acquired data (i.e., animal and environmental sensors; field data on pastures) were used to train the models. Several types of data have been used, from environmental (i.e., VHR-REA) to satellite (i.e., Sentinel2), and soil profile data. The model produced indexes and indicators agreed with the stakeholders, for example variation in milk production (milk yield, fat and protein percentage) in service 1.a, THI evaluation in service 2 and from the IoT sensors (animal and environmental), and pasture availability in service 3.
Using the methodologies and data here described, the prediction model will be improved in accuracy, to give to the stakeholders the most accurate indexes and indicators.

7. Bibliography

1. Herbut P, Angrecka S, Walczak J. Environmental parameters to assessing of heat stress in dairy cattle-a review. Int J Biometeorol. 2018;62: 2089–2097.

2. Rashamol VP, Sejian V, Pragna P, Lees AM, Bagath M, Krishnan G, et al. Prediction models, assessment methodologies and biotechnological tools to quantify heat stress response in ruminant livestock. Int J Biometeorol. 2019;63: 1265–1281.

3. Bohmanova J, Misztal I, Cole JB. Temperature-Humidity Indices as Indicators of Milk Production Losses due to Heat Stress. J Dairy Sci. 2007;90: 1947–1956.

4. Heat Stress – Temperature-humidity Index. [cited 28 Sep 2023]. Available: https://www.megalac.com/resources-advice/fats-advice/104-heat-stress-temperaturehumidity-index

5. General — H2O 3.42.0.3 documentation. [cited 28 Sep 2023]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/faq/general.html#i-am-writing-an-academic-research-paper-and-i-would-like-to-cite-h2o-in-my-bibliography-how-should-i-do-that

6. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014;8: 14.

7. Raffa M, Adinolfi M, Reder A, Marras GF, Mancini M, Scipione G, et al. Very High Resolution Projections over Italy under different CMIP5 IPCC scenarios. Scientific Data. 2023;10: 1–13.

8. Raffa M, Reder A, Marras GF, Mancini M, Scipione G, Santini M, et al. VHR-REA_IT Dataset: Very High Resolution Dynamical Downscaling of ERA5 Reanalysis over Italy by COSMO-CLM. *Data*. 2021; 6(8):88. https://doi.org/10.3390/data6080088.

9. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13: e1002195.

10. Nayeri S, Sargolzaei M, Tulpan D. A review of traditional and machine learning methods applied to animal breeding. Anim Health Res Rev. 2019;20: 31–46.

11. Polsky L, von Keyserlingk MAG. Invited review: Effects of heat stress on dairy cattle welfare. J Dairy Sci. 2017;100: 8645–8657.

12. Bernabucci U, Biffani S, Buggiotti L, Vitali A, Lacetera N, Nardone A. The effects of heat stress in Italian Holstein dairy cattle. J Dairy Sci. 2014;97: 471–486.

13. Thornton P, Nelson G, Mayberry D, Herrero M. Increases in extreme heat stress in domesticated livestock species during the twenty-first century. Glob Chang Biol. 2021;27: 5762–5772.

14. Smith DL, Smith T, Rude BJ, Ward SH. Short communication: Comparison of the effects of heat stress on milk and component yields and somatic cell score in Holstein and Jersey cows. J Dairy Sci. 2013;96: 3028–3033.

15. Leles JS, Rodrigues ICS, Neto MFV, Neto AMV, da Rocha DR, da Costa ANL, et al. Heat Stress and Body

Temperature in Brown Swiss Cows Raised in Semi-Arid Climate of Ceará State, Brazil. Acta Scientiae Veterinariae. 2017;45: 1–8.

16. From Landsat to leafhoppers: A multidisciplinary approach for sustainable stocking assessment and ecological monitoring in mountain grasslands. Agric Ecosyst Environ. 2016;234: 118–133.

17. Bretas IL, Valente DSM, Silva FF, Chizzotti ML, Paulino MF, D'Áurea AP, et al. Prediction of aboveground biomass and dry-matter content in Brachiaria pastures by combining meteorological data and satellite imagery. Grass Forage Sci. 2021;76: 340–352.

18. Cappai S, Loi F, Coccollone A, Contu M, Capece P, Fiori M, et al. Retrospective analysis of Bluetongue farm risk profile definition, based on biology, farm management practices and climatic data. Prev Vet Med. 2018;155: 75–85.

19. Veronesi E, Venter GJ, Labuschagne K, Mellor PS, Carpenter S. Life-history parameters of Culicoides (Avaritia) imicola Kieffer in the laboratory at different rearing temperatures. Vet Parasitol. 2009;163: 370–373.

20. Milk somatic cells, factors influencing their release, future prospects, and practical utility in dairy animals: An overview. [cited 28 Sep 2023]. Available: https://www.veterinaryworld.org/Vol.11/May-2018/1.html

21. Antibiotic dry cow therapy, somatic cell count, and milk production: Retrospective analysis of the associations in dairy herd recording data using multilevel growth models. Prev Vet Med. 2020;180: 105028.