Del. 3.2 – Report on analytical and ML/AI procedures for the indicators/indices database ingestion

Deliverable Lead	UNITUS
Deliverable due date	2024/06/30
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	Deliverable 3.2
Agreement	INEA/CEF/ICT/A2020/2373580
Action	2020-IT-IA-0234
Creator	UNITUS
Deliverable Description	Report on analytical and ML/AI procedures for the indicators/indices database ingestion
Contributors	Marco Milanesi, Federica Gabbianelli, Daniele PIetrucci (UNITUS), Alessandro D'Anca, Gabriele Tramonte (CMCC)
Requested deadline	M30
Reviewer	Cinzia Caroli (CINECA), Elena Rubinacci (NEATEC)

Summary

1. Executive summary

The purpose of this deliverable is to present the analytical procedures (also ML/AI) that produced the indicators and indices stored into the SEBASTIEN Data Lake and on which the four services of the SEBASTIEN Portal are based. This report is closely related to Milestone 9 (Scientific report on final indicators) and Milestone 4 (Report on data synthesis) which complements it.

The document details all the pipelines applied in the SEBASTIEN project. The results obtained are exploited in the Open Data and Services Exploitation Web Portal, the access point to data, services, and informative content. Alongside the Web Portal, a Mobile App was developed that provides the same information. The released Services' Platform will contribute to making the livestock sector more environmentally (esp. climate) while also socio-economically sustainable and valuable, balancing the needs and priorities of different stakeholders.

A total of 4 services are provided, subsequently divided into various areas.

Service 1 consists of three sub-services: 1a, 1b and 1c.

In service 1a, predictions are made to evaluate the effect of climate on productive values (milk yield, fat and protein content). The effect is assessed inside and outside the barn, and for short and long periods.

In service 1b, the Temperature Humidity Index (THI) is used to assess thermal stress threshold levels in different bovine species. Finally, service 1c includes climatic variable effect in estimating breeding values estimation (in collaboration with ANAPRI - Associazione Nazionale Allevatori Razza Pezzata Rossa Italiana).

Service 2 involves the ability to predict the THI inside farmers' barns, for short and long periods.

Service 3 concerns the possibility of predicting the amount of biomass present in pastures, in order to allow the farmer to manage the grazing.

Finally, service 4 pertains to animal health. This last service is divided into services 4a and 4b. Service 4a involves the ability to predict the number of somatic cells present in milk, a proxy of mastitis, adopting the same pipeline used for Service 1a. Service 4b assesses the possibility of developing Blue Tongue disease in Sardinia, in a long-term overview.

Predictions are made in the "short term", "long term", or both "long and short term." In the "short term", predictions provide daily information to farmers, allowing them to take mitigation measures. In the "long term", predictions cover the coming decades from 2021 to 2050. These predictions are useful for administrators to plan long-term interventions.

2. Pipelines from SEBASTIEN data lake

2.1. Machine Learning models for estimating the effect of climate change on livestock productivity and health

The pipeline is applied in the *Services 1a* (long and short period), *Service 4a* (long and short period) and *Service 4b* (long period).

The pipeline was applied only once for "long period" evaluations, and daily, in a routine, for "short period" evaluations.

The pipeline uses as input only climatic data. For the "short period" calculation, the data are daily downloaded from Mistral portal (https://www.mistralportal.it/), using dedicated APIs. For the "long period", historical and future projection data from the Highlander project (https://highlanderproject.eu/) were used. For the indoor evaluation, the results from *Service 2* pipeline are used.

The specific Machine Learning (ML) models developed for each service were applied. The details are reported in the M4 document. Firstly, the input dataset with the specific features for each ML model is created. As already mentioned, for the "short period" analysis, the model is applied every day. Conversely, for the "long period" analysis, the model was applied only once. For the "long period" analysis, the data stored is the anomalies (i.e. difference) between the results obtained in the "past" period (1989-2018) and in the "future" period (2021-2050).

The results from the ML models are NetCDF files, stored in the Data Lake.

The data is exposed from the Data Lake via specific APIs, to allow access and visualization in the Web Portal and in the Mobile App..

2.2. Model to estimate livestock adaptation

The pipeline is applied in the *Service 1b* (long period) each time a user utilizes the service.

The pipeline uses climatic data as input, in particular the future projection data from the Highlander project (for "external") and results from *Service 2* pipeline (for "internal"). The THI (Temperature Humidity Index) is used. The data is stored in the Data Lake as NetCDF.

Thresholds to define different levels of stress (associated to different colors) in different bovine breeds or bovine groups were previously identified.

The pipeline identifies, for a specific THI value, the level of stress (i.e. color) for each group and reports it.

2.3. Machine Learning models for estimating the THI inside the barns

The pipeline is applied in the *Service 2* (long and short period). The pipeline was applied only once for “long period” evaluations, and daily, in a routine, for “short period” evaluations.

The pipeline uses as input data only climatic data. For the “short period” calculation, the data are downloaded daily from Mistral portal, using an API. For the “long period”, historical and future projection data from the Highlander project were used.

The details of the specific ML model applied in this pipeline are reported in the M4 document.

Firstly, the input dataset with the specific features for the ML model is created. As already mentioned, for the “short period” analysis the model is applied every day, instead, for the “long period” analysis, the model was applied only once. For the “long period” analysis, the data stored is the anomalies (i.e. difference) between the results obtained in the “past” period (1989-2018) and “future” period (2021-2070) in terms of THI variation inside barns and increased stress days.

The results from the ML models are NetCDF files, stored in the Data Lake.

The data is exposed from the Data Lake via specific APIs, to allow access and visualization in the Web Portal and in the Mobile App.

2.4. Model to estimate pasture characteristics

The pipeline is applied in the *Service 3* (short period).

The pipeline is applied every 10 days, when Sentinel2 satellite data is updated. Substantially, the Sentinel2 data are updated in a timeframe of 10 days, so, the pipeline automatically downloads the data, to provide up-to-date forecasts to the breeders.

The pipeline uses as input only Sentinel2 satellite data, downloaded from Copernicus portal, using the dedicated API. Once the data are downloaded, the model, which relies on a regression function, is applied to provide the estimation of the available fresh and dry matter biomass in the pasture. Details are reported in the M4 document.

The results from the ML models are NetCDF files, stored in the Data Lake.

The data is exposed from the Data Lake via specific APIs, to allow access and visualization in the Web Portal and in the Mobile App.

2.5. Pipeline to manage IoT sensors data

This pipeline is applied to the data coming from the “animal” and “environment” sensors already described in M4.

The pipeline receives hourly IoT sensor data in JSON format from an MQTT broker. The data are handled to remove the unnecessary information, stored in the Data Lake in NetCDF format and then visualized in the Portal, upon users’ request, using a specific dashboard. Graphics of the data of the last 24 hours are also available in the dashboard.

3. Conclusion

Here a report on analytical procedures for the indicators and indices database ingestion regarding the four SEBASTIEN's services was presented. The pipelines applied to the SEBASTIEN project estimate various parameters such as the effect of climate change on livestock, livestock adaptation, the THI inside the barns, pasture characteristics, using different models, also ML/AI ones. Moreover, a pipeline to manage IoT sensor data (animal and environmental) was also described.

All data produced by these pipelines is exploited in the Web Portal and in the Mobile App providing farmer-facing services for easier livestock and pasture management, enabling prompt action.