AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Deliverable 4.4

D4.4: AI-PROFICIENT approach for XAI

WP 4: Human-machine interfaces, explainable AI and shop-floor feedback

T4.4: Explainable and transparent AI decision making

Version: 3.0

Dissemination Level: PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

Table of Contents

List of Figures

List of Tables

Disclaimer

This document contains description of the AI-PROFICIENT project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AI-PROFICIENT consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

AI-PROFICIENT has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

Title: D4.4 AI-PROFICIENT approach for XAI

Lead Beneficiary	Institute Mihajlo Pupin (IMP)
Due Date	January 31st 2023
Submission Date	31.01.2023.
Status	[✓ Final] Preliminary Draft
Description	Description of the methodology and results for XAI solutions
Authors	Dea Pujic Marc Anderson Eduardo Gilabert Laritza Limia Fernandez
Type	Report
Review Status	Draft WP Leader accepted [✓ PC + TL accepted]
Action Requested	To be revised by partners For approval by the WP leader For approval by the Project Coordinator & Technical Leaders [✓ For acknowledgement by partners]

VERSION	ACTION	OWNER	DATE
v0.1	Table of contents defined	IMP	30/09/2022
v0.2	Inputs provided by partners	IMP, TEK, IBE, UL	23/12/2022
v1.0	Report prepared for internal review by TF (WP4 leader)	IMP	28/12/2022
v2.0	Report updated in accordance with the WP leader review,	IMP, TEK, IBE, UL, TF	16/01/2023

	prepared for PMT review		
v3.0	Final report	IMP, TEK, IBE	27/01/2023

Executive Summary

The Deliverable D4.4 is a public document of AI-PROFICIENT project delivered in the context of WP4, Task 4.4: Explainable and transparent AI decision making. It is envisioned as the bridge between the technical and ethical parts of the project. Namely, in order to increase the penetration and acceptance of the artificial intelligence in the production process, inclusion and improvement of the explainability and transparency is crucial. Therefore, this task was focused on providing XAI within AI-PROFICIENT project.

The report covers current state-of-the-art analysis related to different XAI approaches – exploitation of semantic technologies, transparent machine learning models and post-hoc explainability techniques. Many of those have been utilized for the development of three XAI services explained in this report – surrogate explainable data-driven model, post-hoc explainable analysis module and auditability system. For all of those, the methodology, results, integration and application were explained. In total, XAI services will be present in, at least three, and likely four use cases, as follows: CONTI2, CONTI5, CONTI10 and INEOS3.

This deliverable has its confidential version, as well. Namely, part of the data analysis, results interpretation and presentation are not available for the public, due to data protection of the pilots. Hence, together with the progress report delivered at the end of project, confidential version of this report will be provided to the consortium, project officer and external experts in order to be able to fully evaluate the work done within T4.4. The main difference between the public and the confidential version of this report is level of details presented. For example, when analyzing the importance of different input features to the target variable, in public version all features were anonymized, whilst in the confidential one they were given through pseudonym *feature x.* Nevertheless, all the approaches and developed services have been reported in both versions.

1 Introduction

XAI is a subfield of artificial intelligence (AI) that focuses on developing systems that can provide explanations for their decisions and actions. This includes both the internal workings of the AI system and the reasoning behind its decisions. XAI systems are designed to be transparent and interpretable, which can help to increase trust in AI and improve the user experience. By providing explanations for the decisions made by AI systems, users can better understand how the system works and why it made a particular decision, which can help to build trust and confidence in the system.

XAI has applications in a variety of industries, including healthcare, finance, education, law, manufacturing, etc. In healthcare, these systems can be used to provide explanations for medical diagnoses or treatment recommendations. In finance, they can be used to provide explanations for investment decisions or credit approvals, whilst in education, their utilization could be for providing explanations for academic recommendations or personalized learning plans. Finally, in this particular case, this report presents XAI application in manufacturing and chemical industry, with the main goal of improving the product quality characteristics and increasing the AI system acceptance. In addition to the application of XAI technologies as shop floor assistants, AI-PROFICIENT offers XAI solution for the data scientists and developers in the same domain, in order to facilitate their work and enable growth of AI systems in the production process.

The presentation of the previously explained work is covered by this report. Relevant state-of-the-art review related to the beneficial XAI methods is to be presented and followed by the explanation of the AI-PROFICIENT XAI approaches developed for two pilot sites – Continental plant in France and INEOS in Germany. The corresponding results of the developed models are given, different approaches are benchmarked against each other and various uses cases and application of XAI are represented. Moreover, technical details regarding the integration and practical application of the described models are given. Since there are ongoing debates about the ethical implications of XAI, including issues related to bias and accountability, it is suggested for developers and users of XAI systems to carefully consider different ethical issues and take steps to mitigate any negative impacts. Hence, T4.4 was recognized as the main bridge between the ethical and technical part of AI-PROFICIENT project and overview of the ethical aspects, recommendations and corresponding taken actions for preventing ethical issues are part of this report.

1.1 Description of the task (IMP)

The goal of this task was to improve the explainability and transparency of the artificial intelligence (AI) systems, which are important for establishing trust in AI and enabling effective human-machine collaboration. Hence, this task utilized different approaches in order to achieved predefined goals: 1) collaboration with task T5.2 for creating **RDF semantic data model** was established 2) development and deployment of the **explainable** data driven **models** 3) development of **unwrap black-box models** for introducing explainability to the non-explainable machine learning (ML) models. Finally, it was essential to establish close collaboration with the ethical team, so that achieving all these goals was carried out respecting ethics by design approach.

1.2 List of services (IMP)

This introductory section, except defining scope and goals of this task and report, is devoted to linking work from T4.4 with the previous work within WP1. Hence, this subsection is devoted to the models developed in T4.4 with the once previously defined in D1.5 and their corresponding use case applications.

Within D1.5, with the goal of conceptualizing AI-PROFICIENT platform architecture, a list of foreseen AI services has been established. It was from out of twelve different services, with Explainable and transparent decision making service (S_ETD) standing out as the one corresponding to this particular task. Nevertheless, with the further advancement of the use cases from October 2021, when this report was submitted, there was a need to group different services in one unit and services that are results of this task exceed the description of just S_ETD. Hence, in the following Table 1, a new list of the developed services is be given, and the corresponding mapping the previously services in D1.5. The first one, surrogate explainable data-driven model (SDDM) is envisioned to be a support to generative optimization service (S_GHO) from T3.4, and is supposed to estimate expected value of the particular product characteristics and provide the most probable cause for that decision – the most influential process parameters. Hence, this final service, explained in more details within section 3.1.1 is a combination of the three foreseen services – *predictive production quality assurance service* (S_PRE), *root-cause identification service* (S_ROO) and *explainable and transparent decision making* (S_ETD). When post-hoc explainable analysis module is considered (PEAA), it will be utilized for root cause anomaly analysis for a certain recent historical data, as given in section 3.1.2. Hence it will combine *root-cause identification service* (S_ROO) and *explainable and transparent decision making* (S_ETD). Finally, additional service, which was not envisioned in the original planning, was developed – auditability system. It is not intended directly for the plants, but for data scientists and developers preparing services for the plant. Nevertheless, it will enable improved explainability and transparency, and hence, could be categorized as *explainable and transparent decision making* (S_ETD).

To sum it up, even with the modifications of the service strict definitions, no functionalities were lost, but the pipeline within the use case was more deeply defined and restructured. Hence, all previously promised results, and even new ones are achieved through the developed XAI modules.

Table 1 - Mapping of the final XAI service and the list of services given in D1.5

Service name	Abbrev.	S_DIA	S_PRE	S_ROO	S_EAR	S_ETD
Surrogate explainable data-driven model	SDDM		+	+		+
Post-hoc explainable analysis module	PEAA			+		+
Auditability system	AS					+

Apart from the mapping between the envisioned and developed services, in this subsection analysis of the use cases which were considered within this task will be given. In D1.3 in Figure 84 a summary of the expected enrolment of tasks per use cases are given. In that table, T4.4 was foreseen to be present in the following use cases - CONTI5, CONTI10, INEOS3 and potentially CONTI2 and CONTI7. Services are present on CONTI2, CONTI5, CONTI10, and will likely be in INEOS3 once validation phase is ended, which is not the case in the moment when this report is being written. A break down on the service level related to the specific use cases is given in Table 2.

Table 2 - XAI services presence on a use case

Service name	Abbrev.	CONTI2	CONTI5	CONTI10	INEOS3
Surrogate explainable data-driven model	SDDM	+		+	expected, to be confirmed after UC validation
Post-hoc explainable analysis module	PEAA			+	
Auditability system	AS	+	+		

1.3 General functional requirements (IMP)

In this final subsection of the introduction, contribution of each XAI service to the functionalities that were promised to be provided by AI-PROFICIENT project within D1.4 will be given. As it could be observed from Table 3 contributions corresponds to the mapping given in Table 2, meaning that T4.4 contributes to three functionalities – predictive production quality assurance, root-cause identification and explainable and transparent decision making.

Table 3 - Contribution of XAI service to functional requirements from D1.4

Service name	**Abbrev.**	**DIA**	**PRE**	**ROO**	**EAR**	**ETD**
Surrogate explainable data-driven model	SDDM		+	+		+
Post-hoc explainable analysis module	PEAA			+		+
Auditability system	AS					+

2 State of the art analysis

Even though the maturity of artificial intelligence (AI) technologies is rather advanced nowadays, according to McKinsey[1], their adoption, deployment and application is not as wide as might be expected. This could be attributed to many barriers including cultural, economic and technical [1], [2], as well as social barriers, where the lack of trust of potential end-users in AI systems is remarkable [3], [4]. As a matter of fact, there are many concerns that lead to this lack of trust such as potential safety issues that may lead to harm humans [5], [6] and biases towards the penalization of certain social groups [7]–[9]. However, this lack of trust, if carefully managed, can be overcome thus contributing to the acceptance of AI systems [2].

AI trustworthiness can be defined as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" [10]. There are many factors that affect this lack of trust [11], [12], including explainability. This factor has been addressed by so-called explainable artificial intelligence (XAI), which refers to the "techniques that enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" [13]. XAI was intensively studied from the 1970s to the 1990s [14], although a resurgence of the topic has been seen recently due to the current technological advancements in the various fields of AI [15].

Explainable artificial intelligence (XAI) is a subfield of artificial intelligence (AI) that focuses on developing systems that can provide explanations for their decisions and actions and that is becoming increasingly popular these days. XAI includes both the internal workings of the AI system and the reasoning behind its decisions. These systems are designed to be transparent and interpretable, which can help to increase trust in AI and improve the user experience. By providing explanations for the decisions made by AI systems, users can better understand how the system works and why it made a particular decision. This can help to build trust and confidence in the system. Moreover, XAI is beneficial for preventing negative impact of AI, such as bias and discrimination. Namely, XAI systems can help to identify and mitigate these issues by providing explanations for the decisions made by AI systems. Furthermore, they improve user experience by providing explanations for the decisions made by AI since they are helping users understand how the system works and why it made a particular decision. Finally, some industries and governments are, even, starting to require AI systems to be transparent and explainable in order to ensure that they are fair and accountable, which is further driving the development of XAI systems.

Overall, the increasing complexity of AI systems, the need for transparency and accountability, the desire to improve the user experience, and the regulatory environment are all contributing to the increasing popularity of XAI systems. In general, as stated by [16], humans are hesitant to adopt techniques that are not directly interpretable, tractable and trustworthy. Hence, this report is focused on the analysis of the XAI approaches and their application in production process.

2.1 Data-driven Explainable AI Modeling approaches (IMP)

The first considered significant group of approaches are models and data-driven explainable techniques. As categorized by [17], these models could be separated into two main groups:

1. **Transparent machine learning models** – the once that are understandable themselves and that do not require any additional method in order to explain the decisions or suggestions given by the considered ML models; examples of transparent machine learning models are linear regression, decision trees, rule-based models, Bayesian models, etc.
2. **Post-hoc explainability techniques for ML models** - On contrary to the transparent ML models, post-hoc explainability or unwrapping black box models techniques are approaches designed to create transparency and explainability to the non-explainable models. Namely, it is

[1] https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain

common understanding that with the increase of the ML model accuracy, its complexity increases, as well, but its interpretability decreases, as given in Figure 1 from [2]. Hence, it is crucial to have methods which would support the most performable ML models, such as deep neural networks, in order to ensure their long term acceptance amongst different users in application domains. Moreover, in the same figure, it could be observed that with the increase of interpretability by non-interpretable models, additional improvement on the model performance side could be achieved. Namely, by detecting the most influential factors on the decision making process and/or improved understanding on the decision making process, model designers and developers could redesign models and adapt them, so that performance increase is achieved.

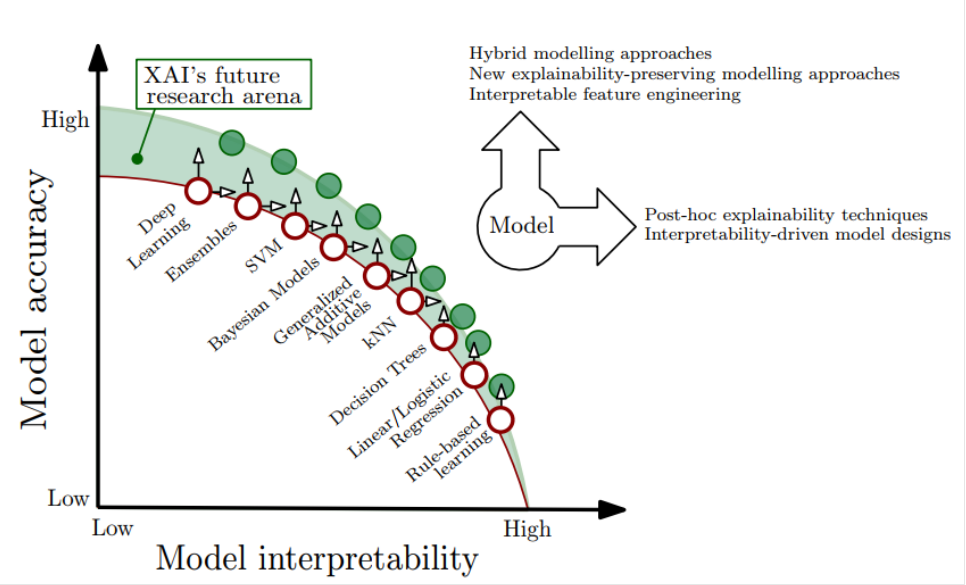

XAI's future research arena
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning

Figure 1 - Dependency of model interpretability from model accuracy

Various post-hoc explainability technologies are present, and could separate into two groups, depending on the models for which they are applicable [17]:

- **Model agnostic approaches** can be applied to any ML, regardless of its internal processing or representation. These approaches do not require any specific knowledge about how the ML model works or what it is designed to do, which makes them widely applicable to a variety of ML models.
    - **Local Interpretable Model-Agnostic Explanations (LIME)** presented in [18] is designed to approximate the complex non-interpretable model in the small proximity with the linear interpretable one, which is further used for interpreting complex model relations. This linearization is carried out by generating synthetic perturb input data, which are just slightly modified that the initial one.

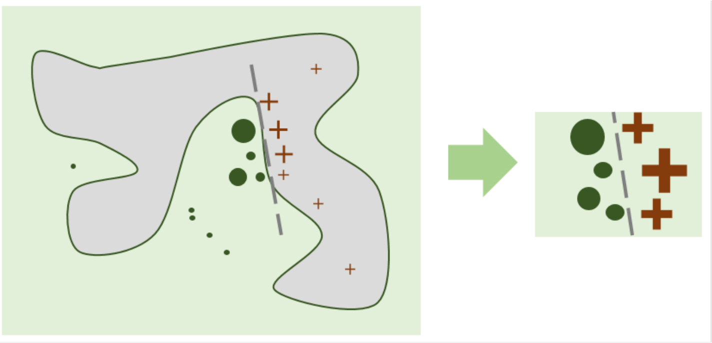


Figure 2 - LIME approach

  - **SHapley Additive exPlanations** (SHAP) presented in [19] suggests application of game theory for introducing explainability to non-explainable ML models. The basic idea behind SHAP is to explain the decision made by an ML model by quantifying the contribution of each input feature to the prediction. This is done using Shapley values, which are a concept from game theory that measure the value of a player's contribution to the overall outcome of a game. In the context of ML explanations, Shapley values are used to measure the contribution of each input feature to the prediction made by the ML model.
  - **Quantitative Input Influence** (QII) [20] is a measure of the importance or influence of an input feature on the output of a machine learning (ML) model. QII measures how much the output of the ML model changes when the value of the input feature is changed. This can help to identify which features are the most important for the ML model to make a particular prediction. QII can be calculated using various methods, such as sensitivity analysis, which involves perturbing the input data and observing how the output changes as a result, or by using gradient-based methods, which measure the change in the output as the input feature is changed. QII can be used to provide explanations for the decisions made by ML models and to identify which features are most important in influencing the model's predictions. QII is often used in combination with other techniques, such as feature importance or Shapley values, to provide more detailed and accurate explanations for the decisions made by ML models.
  - **Automatic STRucture IDentification method** (ASTRID) presented in [21] is the approach designed to discover the subset of the most relevant input features, such that the accuracy of a classifier trained with this subset of features is not significantly different from the accuracy of a classifier trained on the full set of features.
- **Model specific approaches** are techniques that are designed to explain the decisions made by specific machine learning (ML) models. These approaches are tailored to the particular characteristics and structure of the ML model and may not be applicable to other types of models. One of the key benefits of model specific approaches is that they can provide more detailed and accurate explanations for the decisions made by the ML model. However, this benefit comes at the cost of flexibility, as model specific approaches are only applicable to the specific ML model for which they were designed.
  For example, some model specific approaches may rely on knowledge about the internal representation of the ML model, such as the weights and biases of a neural network. Other model specific approaches may make use of the structure of the ML model, such as the decision tree structure of a random forest. By exploiting these characteristics of the ML model, model specific approaches can provide more detailed and accurate explanations for the decisions made by the model.
  However, it is important to note that model specific approaches are only applicable to the specific ML model for which they were designed. If the ML model is changed or replaced, the model specific approach may no longer be applicable or effective. As a result, it may be necessary to develop a new model specific approach for the new ML model. Overall, model specific approaches can be useful for explaining the decisions made by specific ML models, but they may not be as flexible as model agnostic approaches.
  - The first group of the XAI approaches for interpreting neural networks are the once which are **utilizing explainable ML models**, such as decision trees or rule based model, for decomposing neural networks. For example, **DeepRED** [22] is a model specific XAI approach developed for improving explainability of deep neural networks by creating decision tree for its representation. Moreover, **KT method** presented in [23] is designed to generate rules based on the neural network model. Unfortunately, this approach could become highly computationally expensive. Moreover, since in real life use, trees or rulesets for representing ML model become huge and complex, their explainability is quite low, as their usage.
  - Another group of approaches are the once utilizing **backward propagation for generating post-hoc explanations**. Backpropagation works by propagating the error gradient backwards through the layers of the deep network model, starting from the output layer and working towards the input layer. This allows the model

to adjust the weights and biases of the network to minimize the error between the predicted output and the true output. To generate post-hoc explanations using backpropagation, the algorithm can be modified to propagate the influence of the input features backwards through the network instead of the error gradient. This can help to identify the input features that had the greatest influence on the model's prediction. One way to visualize the results of this process is to create a heat map that shows the importance of each input feature for a particular prediction. The heat map can be created by multiplying the input feature values by the influence of the feature on the prediction and summing the results for each feature. The resulting values can be used to create a color-coded map that shows the relative importance of each feature. One of the representatives is **Layer-wise Relevance Propagation** (LRP) [24], a local description model for extracting relevant input features, which could be seen as deep Taylor decomposition [25], as suggested by [26, p.]. Moreover, **Deep Learning Important Features Technique** (DeepLIFT) [27], [28] is also based on backpropagation. It is based on the idea that the contribution of each input feature to the model's prediction can be quantified and propagated backwards through the layers of the model. DeepLIFT works by calculating the change in the output of the model that is caused by a change in the value of an input feature. This change in the output is referred to as the "lift" of the feature, and it is used to quantify the influence of the feature on the model's prediction. The lift of a feature is calculated by comparing the output of the model with and without the feature, and it is adjusted for any changes in the output that are caused by other features. Furthermore, **SmoothGrad** [29] works by adding Gaussian noise to the input data and calculating the change in the output of the ML model as a result. By repeating this process multiple times and averaging the results, SmoothGrad can identify the input features that had the greatest influence on the model's prediction. Additionally, **Integrated Gradients** [30] are envisioned to calculate the difference in the output of the ML model between the input data and a reference point, and then multiplying this difference by the gradient of the model's output with respect to the input data. The resulting values are then integrated over the input data to produce a set of importance scores for each input feature.

2.2 Semantic technologies (TEK)

Explainability is necessary but far from sufficient for achieving the desired trustworthiness in AI systems. In order to do so, not only should the developed AI systems be explainable, but also accountable [31], [32]. As a matter of fact, the ability to hold them accountable by explaining their inner workings, their results and the causes of failures to users, regulators and citizens, is critical to achieve trust [33]. Accountability can be defined as the ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met [32]. This means that with an accountable AI system, the causes that lead to a given decision can be discovered, even if its underlying model’s details are not fully known or must be kept secret. In other words, the person, group or company in charge of the AI system should be able to answer questions that are related, not only to the obtained outputs (e.g., what the output result is or when the output is generated), but also to the AI procedures that led to such outputs (e.g., which data set(s) are used to train the AI system or how well the AI system performs in terms of accuracy).

However, the information needed to answer these questions is hardly ever accessible in a straightforward way. This information is scattered across multiple files, repositories and systems, and in the worst-case scenario, is not even registered. That means that, if the person, group or company in charge of the AI system wanted to answer the aforementioned questions, it would be very time consuming, as it would be needed to be an expert or have the help of experts in different frameworks, systems, data models, repositories and query languages. As a matter of fact, the regular performance of these accountancy tasks would be infeasible.

Therefore, it seems reasonable to consider that the adequate representation of data, processes and workflows involved in AI systems could contribute to make them accountable in an easier and systematic manner. There are a variety of technologies that offer conceptual modelling capabilities to describe a domain of interest, but only ontologies combine this feature with Web compliance, formality and reasoning capabilities [34].

Since AI is a field that comprises a variety of fields ranging from natural language processing to knowledge representation [35], this work focuses on a specific branch: machine learning (ML). Namely, an ontology-based approach is proposed towards achieving the accountability of ML systems.

Although the usage of Semantic Technologies towards the achievement of Trustworthy AI has been researched in the literature, their full potential is yet to be exploited.

[36] provides a literature-based overview of the usage of Semantic Technologies alongside ML methods in order to facilitate their explainability. According to the reviewed literature, the main role of the Semantic Technologies is, on the one hand, to make Neural Networks explainable, and on the other, to create explainable embeddings with knowledge graphs. As for the domains of application, the healthcare domain has attracted a lot of attention, although they are also present in the entertainment or commercial field.

[37] presents an approach for creating more understandable post-hoc explanations of decision tree algorithms. In this approach, ontologies that model the concerned domain knowledge are used in the process of generating such explanations. Results showed that decision trees generated with the support of domain ontologies are more understandable than those generated without them. The downside of this approach is that the used ontologies are manually created ad-hoc for each problem, which definitely hinders their usability.

[38] proposes an explanation ontology that can be used by designers to support the generation of different explanation types into their AI-enabled systems. Nine different explanation types are identified, each with different needs, and the proposed ontology can encode them as OWL (Web Ontology Language) restrictions. This provides a means for system designers to translate their user requirements gathered from user studies to explanations that can be generated by their systems.

Doctor XAI is presented in [39], an ontology-based approach for producing post-hoc explanations of black-box sequential data classification methods. The application of the approach is focused on the medical domain, but since the method is agnostic with regards to the black box model, the possible applications cover several scenarios where a sequence of events linked to ontology concepts can be identified, including an online market basket analysis or Wikipedia user behavior forecast.

[40] proposes an ontology-based knowledge representation and reasoning framework for human-centered transfer learning explanations. This approach exploits the reasoning capabilities offered by the Semantic Technologies and makes use of external knowledge bases to infer different kinds of human understandable explanatory evidence, allowing common users without ML expertise to have a good insight of transfer learning explanations. In [41], semantic reasoning and ML have been combined for explaining the rationale of classification predictions in an informative manner to human users, which is expected to in turn strengthen the trust relationship between human decision makers and intelligent systems making the prediction.

Knowledge graphs can provide an explainable layer that may act in an effective way to interpret the black box answers given by neural models [42], [43], which have been proved to be useful in the field of conversational agents [44] and recommender systems [45]. Furthermore, the existing approaches, limitations and opportunities for knowledge graphs in XAI are analyzed in [46]. In this article, knowledge graphs are envisioned to bring XAI to the right level of semantics and interpretability, supporting explanations that may overtake existing limitations in different AI fields, ranging from computer vision to natural language processing.

In [47], it is stated that semantic representations for explainability can evolve from existing representations for provenance and context. Therefore, the strengths of the Semantic Web, coupled with ML methods, will be a significant contributor to hybrid explainable AI systems. To the author’s knowledge, so far, the main focus of the usage of Semantic Technologies has been placed on explainability, although accountability is considered a key requirement that should be met to achieve trustworthy AI systems [17], [48]. [49] makes a first contribution on the usage of ontologies to support the accountability of ML systems, proposing a method to know which predictive model was responsible for making a given forecast, but also, to understand where such forecasts come from, that is, what their underlying rationale is. However, many fundamental aspects that could contribute to making the ML systems accountable remain unaddressed, such as the description of the procedure followed to develop the predictive models.

All this evidence reinforces the discourse that the Semantic Technologies could play a more important role in achieving trustworthy AI systems in general, and in solving the accountability challenge for ML systems in particular.

3 XAI approaches and the results

In this section, methodologies designed for XAI services and their corresponding results are given.

3.1 Data-driven Explainable AI Modeling approaches

3.1.1 Surrogate explainable data-driven model (IMP)

Surrogate explainable data-driven model (SDDM) was designed with the two main goals in mind – to support generative holistic optimization (GHO) and to provide the end user with the most probable cause of a specific product characteristic degradation. Namely, SDDM was envisioned to be used in all use cases where optimization is present and a model digital twin is not available. Hence, SDDM is envisioned in three use cases – CONTI2, CONTI10 and INEOS3, as presented in Table 2. Since validation of INEOS3 use case has not been finished by the time this report is submitted, corresponding models will be developed after the formal end of this task.

SDDM is a data-driven based approach which was envisioned to forecast the value of each relevant product characteristics depending on the measured process parameters and to estimate the most influential factors on those forecasts, as shown in Figure 3. Hence, SDDM consists of two parts – a machine learning (ML) based approach for forecasting purposes and an accompanying XAI model for root cause analysis. Various approaches have been explored and tested for both of these cases, with more details presented through the obtained results.

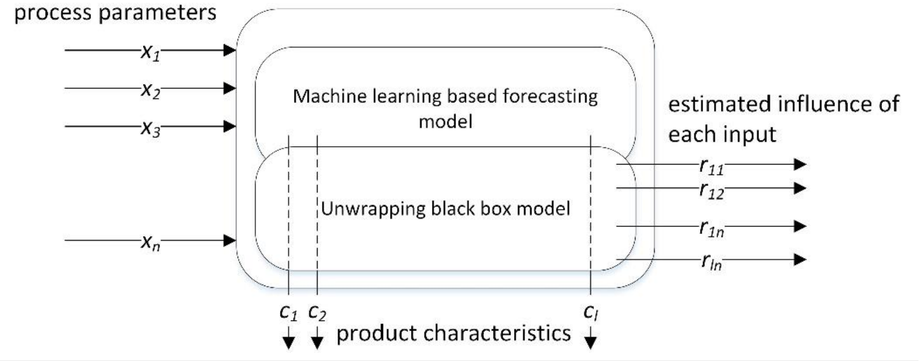


Figure 3 - SDDM design

3.1.1.1 Data preprocessing

As already explained, the SDDM approach was envisioned to be utilized in two Continental and Ineos Cologne use cases. Since INEOS3 UC has not been validated yet, models have been developed using Continental data and, consequently, corresponding data analysis will be presented. To be able to provide precise models, it was necessary for the plant to provide huge amount of historical data. This data set included different process parameters and corresponding product characteristics for four years. Hence, extensive data analysis and preprocessing was an essential first step before the model development. In order to avoid work duplication, a data preprocessing framework was created within Task 3.4, and will be presented in more details in the corresponding report. It offered the most common data cleansing requests – data resampling, data normalization, data filtering in consistence with the production type, etc.

After the first data cleansing, it was necessary to filter the data in accordance with the production regime. Namely, as specified by the plant, various regimes exist such as nominal regime, machine setup, mix change, machine failure, etc. Hence, it was necessary to filter out the data only from the periods which correspond to the use case. For example, in case of CONTI10, that would be nominal production regime.

Moreover, since the production process is continuous, measurements of a specific process parameter and the measured corresponding product characteristics do not have the same timestamps. Namely, a certain time latency is a consequence of the finite time of the production process through the production line. In order to be able to estimate the latency, photocell measurements have been exploited. By analyzing points in time when different process data, product characteristics and photocells measurements changes, it was concluded that latency between the extrusion and end of production is approximately 8 min. After that time, product achieves gaining characteristics. In order to make sure that variations in latency due to the line speed change is covered, for each relevant feature it was decided to use a window of $(latency - 20s, latency + 20s)$ as the input time series. Finally, comparison between the filtered and original signal could be observed in Figure 4. From the left hand side figure, it could be seen that an extremely high amount of the product weight values were zeros, which is expected, since production was not in the nominal regime, even though the set point was not zero. On the other hand, time series which was obtained after the filtering process presented in the right hand side figure, has a noticeably lower amount of irregular values, due to the fact that these were periods when production was in place. Similar thing could be observed from the histograms in Figure 5.

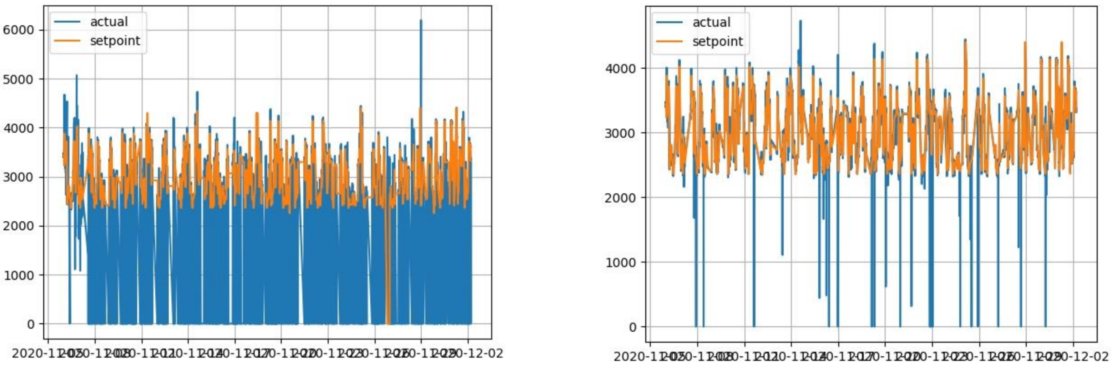


Figure 4 - Comparison between the original product piece weight measured characteristics and the one after filtering data from production nominal regime

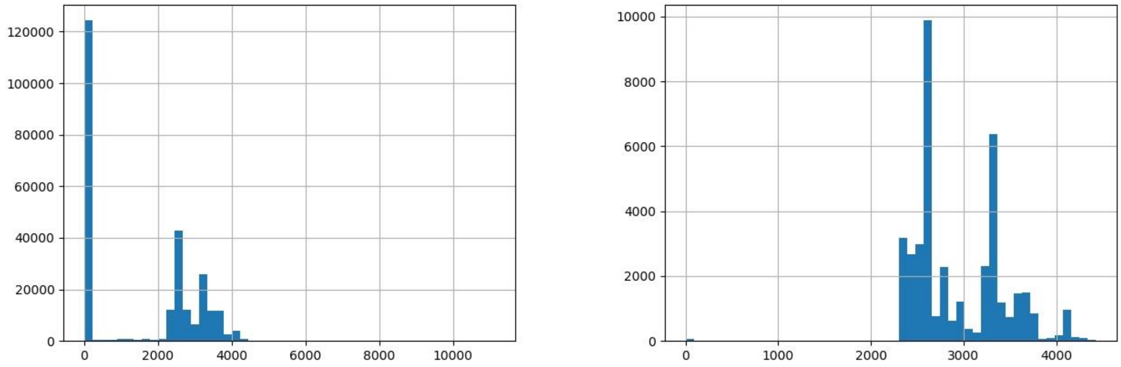


Figure 5 - Histograms of measured product weight before and after filtering

3.1.1.2 Forecasting models and the corresponding results comparison

With the goal of developing as precise a model as possible, various approaches have been tested, and here they are elaborated, and the corresponding results given. In order not to overcomplicate the content of this report, in this section results are presented for one selected product characteristic – *product weight*.

Random forest

The first model developed for estimating product characteristics was random forest (RF). Random forest is a data-driven approach designed as an ensemble model with regression tree as a weak learner, meaning that the final estimation is the mean value of individual estimations of the single tree. In this way, more granularity of the final estimation can be achieved and estimation variance is decreased.

In order to be trained, a set of hyper parameters has to be set up before RF training. The parameters that were optimized for the purpose of SDDM development are the following: *max depth of the regression tree (max_depth), number of trees (num_tree), max number of leafs per node (max_leaf)*. The RF model showed high performance, as suggested in Table 5, as its root mean square error (**RMSE**) was **216** and mean absolute error (**MAE**) was **98** for a weight characteristic on testing data. Taking into consideration that mean output value on the whole testing set is **3092**, it could be concluded that this model is highly precise with relative square error (**RSE**) of **7%** and relative absolute error (**RAE**) of **3%**. These performances have been evaluated on testing set which contained data for product characteristics in a wide range, from 2000 to 4200, showing that model is tracking the output adequately for different conditions.

Neural networks

Neural networks are widely used and performable models frequently used for solving various complex problems. Hence, within in this task, they were developed, trained and compared with other ML models for achieving the most precise predictions. As it is well known, the most important step in utilization of any neural network is designing its architecture. Various different layers have been presented, with different purposes:

- Dense layer
- Convolutional layer
- Long Short-term Memory (LSTM) layer
- Recurrent layer
- Bidirectional layer
- Dropout layer

Hence, one of the beneficial steps during the development was optimization neural network model architecture. Combinations and parameters of different layers have been tested and the final architecture is given in Table 4, whilst the optimization criterion was performance on the validation data set.

Finally, the performances on testing data are as follows **RMSE = 274**, **RSE = 9%** also given in Table 5. In comparison with RF, estimations are in total less precise, but are more fluctuating, which could be beneficial from an optimization perspective. Namely, RF, tends to estimate similar samples with completely the same value. Hence, this could disable the optimization engine to carry out slight adaptations.

Table 4 - Final neural network architecture used for SDDM

Layer	num. of ker.	kernel size	activation func.	ratio
LSTM	32		tanh	
LSTM	64		tanh	
Conv1D	15	3	relu	
Dropout				0.5
Conv1D	30	3	relu	
Flatten				
Dense	100		relu	
Dense	100		relu	
Dense	1		linear	

Support vector regression (SVR)

The third model that was considered within this report was support vector regression. Unlike all of the previous models which are optimizing model parameters to minimize total estimation error, SVR allows a certain error as long as it is within the acceptable range. Various model parameters could be optimized with the goal of improving estimation performances. The first one, that directs other numerical parameters that could be tuned, is kernel function. Different kernel functions are present in literature, such as polynomial, sigmoid and radial basis kernel function. All of these have been tested, and the best performance was achieved by radial basis kernel function. Apart from the kernel function, two additional hyper parameters were considered for optimization – regularization parameter C and parameter epsilon which specifies the epsilon tube in which no penalty is associated in the loss function. The optimal model resulted and performances as follows: **RMSE = 177, RSE = 5.7%, MAE = 76, RAE = 2.5%**. As expected, by the nature of SVR, its MAE and RAE is smaller than that of RF, but its MSE and RMSE are higher.

K Nearest Neighbors (kNN)

kNN approach is a common ML algorithm which is calculating the output as the weighted sum of output values of k the most similar training examples. The weights could be uniform or could be dependent on the distance between the new and trained examples. Hence, this approach is efficient when training examples are forming a number of groups, so that the output value is determined only by those which are similar.

This is exactly the case with Continental data. Namely, in high dimensional space of process parameters and product characteristics, data are not scattered, rather forming different groups depending on the nominal regime. Hence, it is expected that this particular approach could offer highly precise estimation. This is corroborated by numerical performance evaluation on the testing data set where **MAE = 72**, **RMSE = 189**, **RAE = 2.3%**, **RSE = 6.1%**, also summarized in the Table 5, from where it could be observed that the lowest MAE and RAE are achieved by the kNN model. Similarly, to what was underlined for neural networks, the output of kNN model is continuous when tested, in contrast with RF, where tendency of discrete levels for the output could be observed.

Table 5 - Performance comparison between the different forecasting models for product weight

	Testing data			
Model / performance	**RMSE**	**RSE**	**MAE**	**RAE**
Random forest	216	7%	98	3.2%
Neural networks	274	9%	123	4%
Support vector regression	177	5.7%	76	2.5%
kNN	189	6.1%	72	2.3%

3.1.1.3 XAI supporting approaches and the corresponding results

Apart from the forecasting models which were developed with the main goal of supporting optimization, XAI approaches were envisioned primarily to be given to the end user in order to improve transparency of the model and to increase acceptance of the artificial intelligence in the production process. Within this task, different methodologies have been tested and a summary of the obtained results and conclusions will be elaborated in this report.

Local Interpretable Model-Agnostic Explanations (LIME)

The first approach that has been tested together with all of the previously presented models was LIME. As already extensively elaborated in Section 2.1, LIME is a general XAI approach which is utilized for unwrapping black box models. Since it is not specifically developed for a particular ML approach, it was combined with all of the previously presented models.

Having said that, it could be concluded that the advancement brought by LIME to different ML models is an improvement of the explainability and transparency of the output. Namely, not only that the precision itself would be available, but also LIME offers analysis of the influence of each input on the obtained output. Hence, the most probable causes could be obtained by utilization of this methodology. The example of the output that is provided by LIME is given in Figure 6. Namely, in the figure, inputs have been sorted in accordance with their influence on the output for a specific prediction. Additionally, similar information could be observed, also from Figure 7, where influence on the output depending on the input ordinal number is given. The most influential ones are around the 47th input with a positive influence of 0.36 meaning that with the increase of 1 the particular normalized input, normalized output is expected to increase by 0.36. These inputs are related to speed of the fifth extruder, resulting in the conclusion that it highly influences the product width in this particular case. Furthermore, significant negative influence could be noticed by 72nd input (temperature in extruder 2) where it is expected that when normalized value of the input is increased by 1 that normalized value will be decreased by 0.11. This particular type of the output information will be utilized in two ways:

1. LIME will be utilized as the root cause analyzer, and the most influential input features will be presented to the end user
2. Apart from the presentation to the end user, the same extracted information will be used by the optimizer. Namely, process parameters which have been selected to be highly influential when product characteristics degradation is estimated are the ones that are likely to need update. Hence, they will be suggested for improvement to the optimization engine.

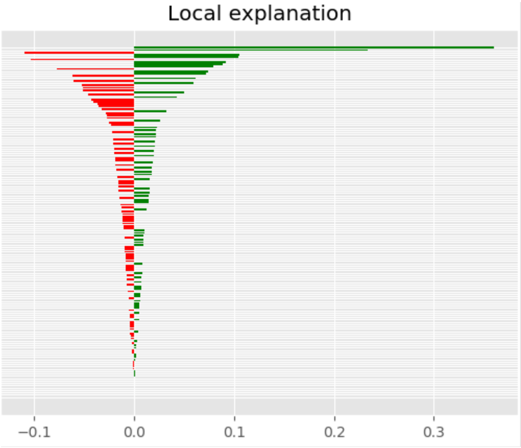


Figure 6 - Output of LIME approach sorted by the most influential inputs on the output

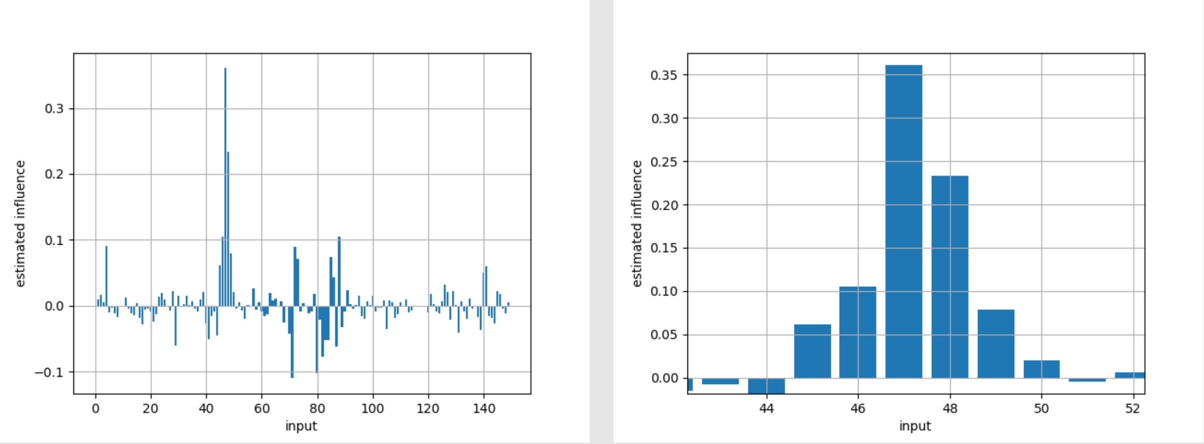


Figure 7 - Output of LIME approach per ordinal number of input

DeepLIFT

In comparison with the previously presented approaches which could be utilized for unwrapping various different ML approaches, DeepLIFT, model specifically developed for neural networks has, also, been tested. DeepLIFT is a technique for attributing the predictions of a deep learning model to the input features (also called "input attributions") that contribute the most to the prediction. It uses a reference input, which is typically an all-zero tensor or a tensor with the mean feature values, to compute the input attributions. It uses a reference input as a starting point. Model prediction on the reference input is estimated. Moreover, prediction on the actual input is also calculated. Finally, the difference between these two predications is attributed to the input features. This difference is called the "residual" and it can be positive or negative, depending on whether the feature increased or decreased the prediction compared to the reference input.

Similarly, to LIME, DeepLIFT was also designed to estimate the influence of the particular input on the output. Hence, visualization of DeepLIFT output is given in Figure 8. Namely, on the x axis the ordinal number of the input is given, whilst on the y axis its influence on the output for a specific set of input parameters. What is crucial to point out here is that groups of neighboring inputs have similar influence. This is a consequence of the input format. Namely, it has already been pointed out that for each selected input feature, a short window of its measurements has been utilized as the input as a time series. Therefore, these neighboring inputs are different samples in time of the same process measurements, leading to the conclusion that the results provided by DeepLIFT approach are valid.

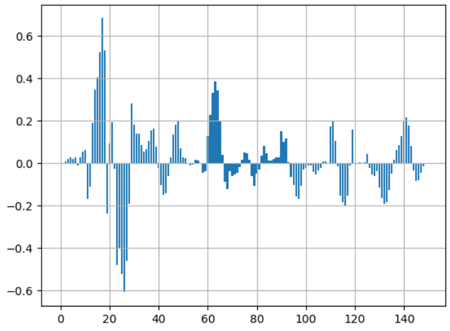


Figure 8 - Estimate influence of the neural network input to its output by DeepLIFT approach

Apart from the validity and precision of the estimation provided by previously presented XAI approaches, a crucial metric to be compared for them is execution time. Namely, validation of root cause analysis is not a straight forward task, since there are no ground truth labels that guarantee the biggest influence of the process parameter on the product characteristics. Nevertheless, by analyzing provided inputs and experience transferred by the domain experts, it was concluded that root cause analysis is satisfactory.

3.1.2 Post-hoc explainable analysis module (IBE)

The Post-hoc Explainable Anomaly Analysis (PEAA) is a module designed for the early anomalies detection and root cause identification of the historical data offered by Continental to improve the quality analysis in CONTI10.

PEAA consists of three submodules to facilitate analysis for operators or managers:

- Anomaly Detection (AD)
- Data Quality Report (DQR)
- Feature Selection using Explainable Artificial Intelligence techniques (FS_XAI)

The aim is to provide information on possible anomalies in the historical data referring to a day, a week, or a month of Continental production, as well as to provide operators and managers with a service that allows them to perform an analysis of specific signals to determine possible incidents in the production process.

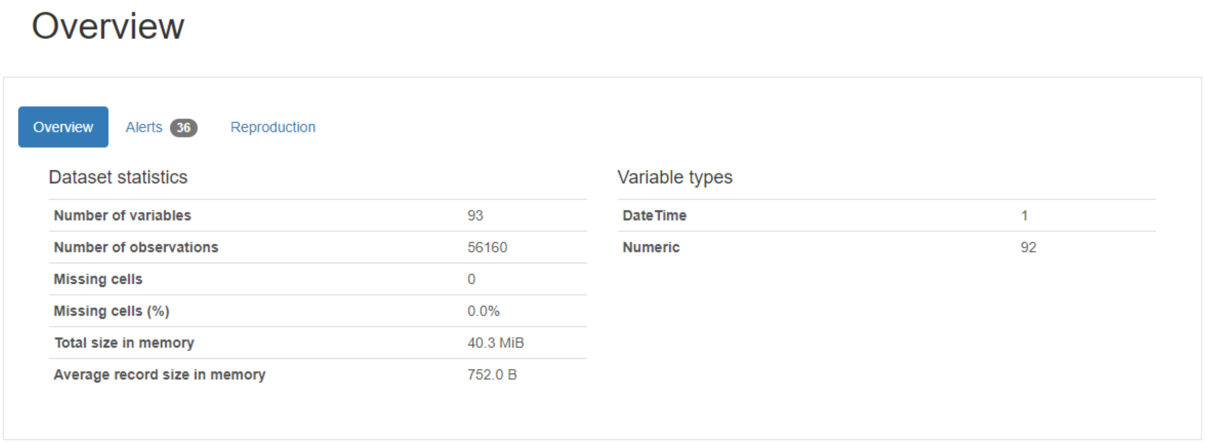
The AD submodule aims to detect possible anomalies in the analyzed historical data to avoid including these values in future analyses. Specific anomaly detection techniques are used, such as clustering, distances, isolation forest, etc. Although this functionality was initially designed to be integrated with the FS_XAI service, it has finally been decided to keep it as an independent service and leave it to the quality manager whether to use the result of anomaly detection in other processes or not. Hence, more details regarding AD will be included in the coming project deliverables as D2.3.

3.1.2.1 Data Quality Report

The DQR submodule performs a statistical analysis of the signals provided by Continental. This functionality aims to analyze the data's quality for its use in prediction models.

This submodule has been designed so that operators and managers can have a global vision of the main variables they wish to analyze and that it is quick and easy to understand. As a result, they can see the main statistics of the selected variables, their distribution, if they have many missing values or contain atypical elements. Visually, they can understand this information later used for in-depth data analysis. An example of the interface is shown in Figure 10.

We can analyze the entire dataset or selected variables. The problems will show as Alerts that suggests to managers which signals may affect or not the production and can exclude them from the future analysis of the cause of degradation.



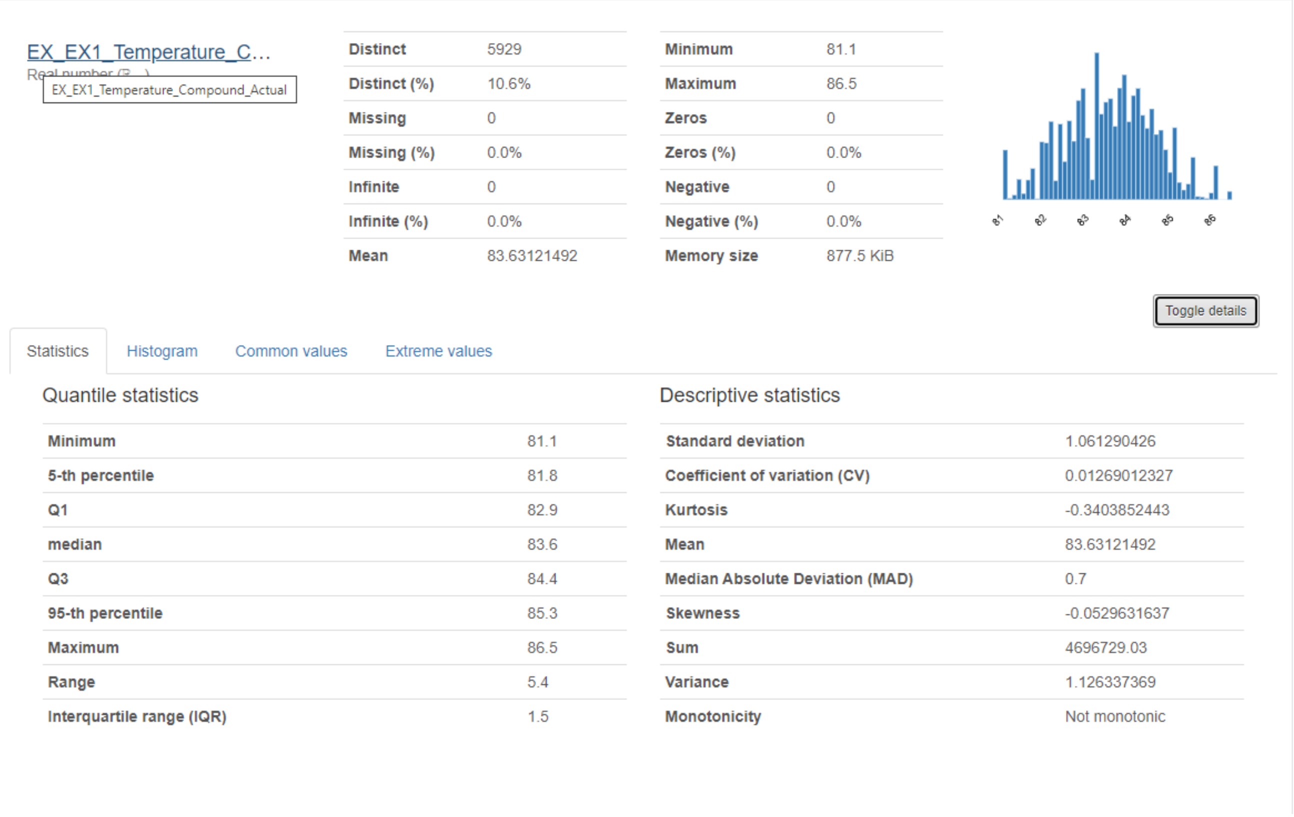
Dataset statistics

Number of variables	93
Number of observations	56160
Missing cells	0
Missing cells (%)	0.0%
Total size in memory	40.3 MiB
Average record size in memory	752.0 B

Variable types

DateTime	1
Numeric	92

Figure 9 - Overview Data Quality Report



EX_EX1_Temperature_C…

Real number (ℝ …)

EX_EX1_Temperature_Compound_Actual

Distinct	5929	Minimum	81.1
Distinct (%)	10.6%	Maximum	86.5
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	83.63121492	Memory size	877.5 KiB

Quantile statistics

Minimum	81.1
5-th percentile	81.8
Q1	82.9
median	83.6
Q3	84.4
95-th percentile	85.3
Maximum	86.5
Range	5.4
Interquartile range (IQR)	1.5

Descriptive statistics

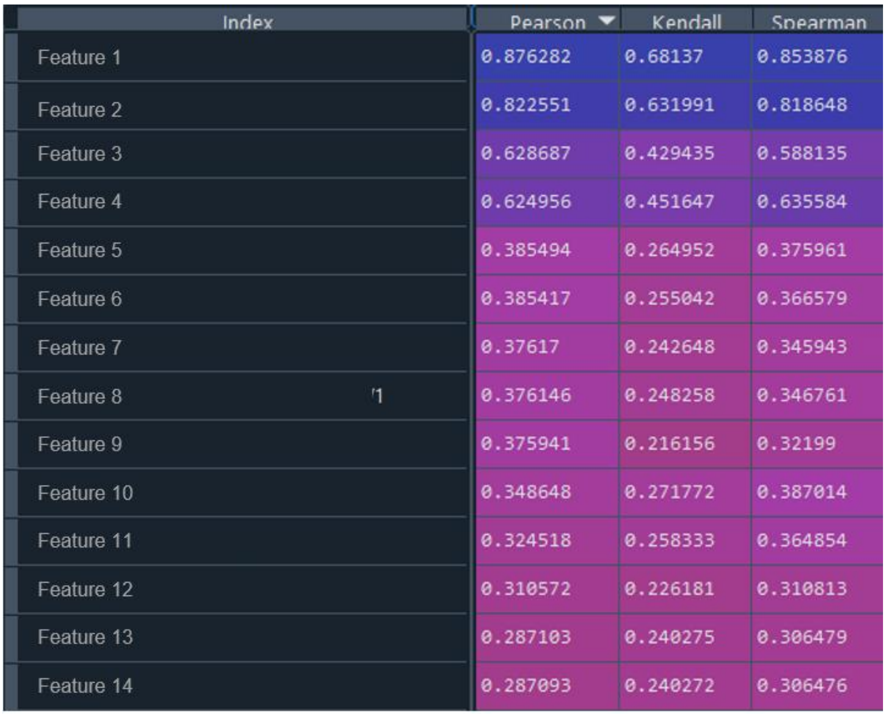
Standard deviation	1.061290426
Coefficient of variation (CV)	0.01269012327
Kurtosis	-0.3403852443
Mean	83.63121492
Median Absolute Deviation (MAD)	0.7
Skewness	-0.0529631637
Sum	4696729.03
Variance	1.126337369
Monotonicity	Not monotonic

Figure 10 - Example of statistics information of one variable

3.1.2.2 Feature Selection using Explainable Artificial Intelligence techniques

Once the information regarding the irrelevant process variables has been obtained through DQR service, a feature selection solution based on XAI techniques could be exploited. Different variables were selected based on linear and non-linear statistics. This solution allows the quality manager to analyze the mutual dependence between different production process variables to determine which signals significantly affect the target variable selected. The idea is that the manager can choose the signal he/she want to study and a date filter to indicate the period wish to run the analysis. In particular, the extruder temperature was chosen to explain the developed models. It is the parameter of the essential extruder in the production process for analysis for one week.

Different correlation tests (Pearson, Spearman, and Kendall) were initially used to check the relations, linear correlation, and monotony between the variables. The results for the different correlation tests, given in Figure 11, are consistent since all ordered variables' influences are in the same manner concerning the chosen objective. Since authors are not allowed to present specific measurements, due to security policy, only numerical outputs are present in the figure and anonymized feature tags.



Index	Pearson	Kendall	Spearman
Feature 1	0.876282	0.68137	0.853876
Feature 2	0.822551	0.631991	0.818648
Feature 3	0.628687	0.429435	0.588135
Feature 4	0.624956	0.451647	0.635584
Feature 5	0.385494	0.264952	0.375961
Feature 6	0.385417	0.255042	0.366579
Feature 7	0.37617	0.242648	0.345943
Feature 8	0.376146	0.248258	0.346761
Feature 9	0.375941	0.216156	0.32199
Feature 10	0.348648	0.271772	0.387014
Feature 11	0.324518	0.258333	0.364854
Feature 12	0.310572	0.226181	0.310813
Feature 13	0.287103	0.240275	0.306479
Feature 14	0.287093	0.240272	0.306476

Figure 11 - Comparison of different correlations test: Pearson, Spearman and Kendall

Additionally, two other feature selection criteria were tested: mutual information[2] and f regression[3], and results are given in Figure 12 and Figure 13. They are recommended as a feature selection criterion to identify potentially predictive features for a downstream classifier, irrespective of the sign of the association with the target variable. Three datasets that contain the most relevant variables selected by each method Correlation, Mutual Information, and F Regression were created. The aim is to use a predictive model that explains the relationship between the most important variables chosen by the feature selection techniques and the target variable being analyzed.

[2] *Kreer, J. G. (1957). "A question of terminology". IRE Transactions on Information Theory.* ***3*** *(3): 208. doi:10.1109/TIT.1957.1057418*

[3] Snedecor, George W.; Cochran, William G. (1989). Statistical Methods (8th ed.). Ames, Iowa: Blackwell Publishing Professional. ISBN 0-8138-1561-4.

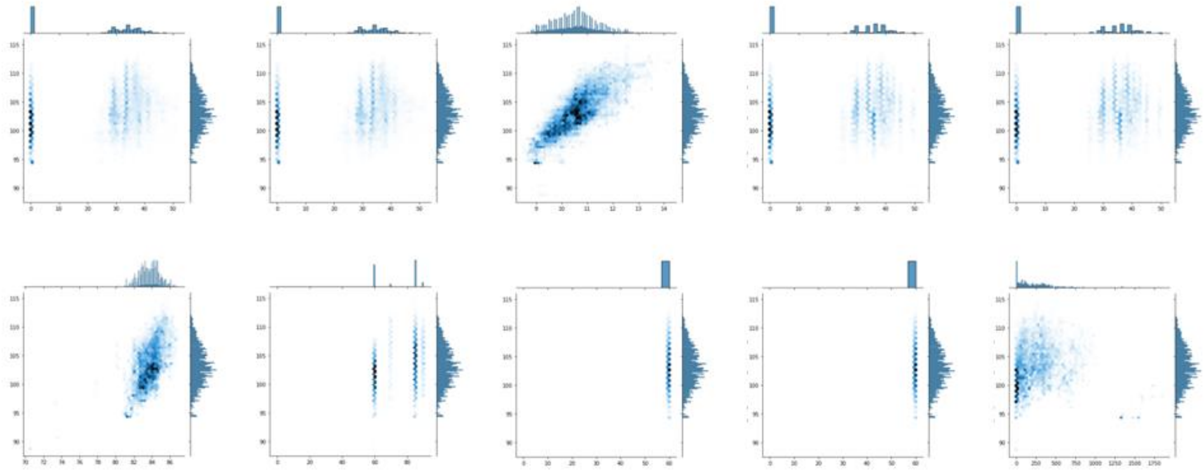


Figure 12 - Correlations of the variables selected by the Mutual Information method with temperature in extruder 2.

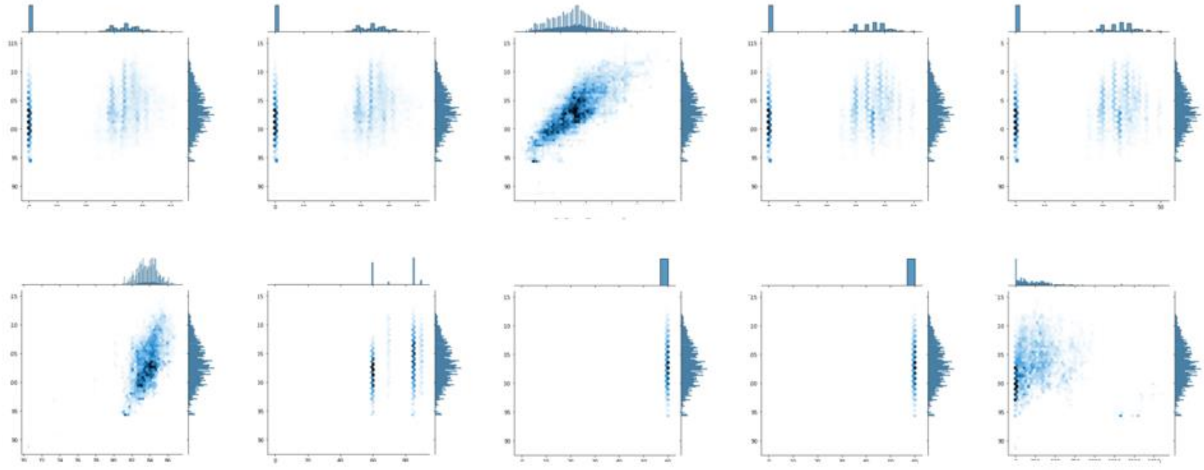


Figure 13 - Correlations of the variables selected by the f regression method with temperature in extruder 2.

A function was created to compare the output of different models on the data set to verify which predictive model should be used. The following models were tested: LASSO regression [50], Elastic Net Regression [51], Gradient Boosting Regression [52], XGBoost [53], LightGBM [54], Random Forest [55], and Support Vector Rregression [56]. The results obtained in the test set are shown in the following table.

Table 6 - Performance comparison between the different predictive models for selected datasets by feature selection techniques.

	Correlation		F Regression		Mutual Information	
Model / performance	**R2**	**Runtime (s)**	**R2**	**Runtime (s)**	**R2**	**Runtime (s)**
Random forest	0.99164	1.115200	0.99186	1.763711	0.96498	1.142860
XGBoost	0.95654	14.472292	0.96623	17.802266	0.91092	8.582994

Gradient Boosting	0.88623	3.935113	0.88232	4.072183	0.70733	2.495856
Lasso	0.82742	0.013999	0.74634	0.196000	0.32677	0.029031
Elastic Net	0.82742	0.021000	0.74627	0.110000	0.32677	0.032961
LightGBM	0.78756	1.037243	0.89108	1.083001	0.68342	0.696000

Based on these results, the Random Forest model was chosen, with the dataset generated by the F regression method to describe the relationship with the selected target variable. As the Random Forest algorithm is a black box model, explainability techniques have been used to interpret the model results. Below, in Figure 14, important of various features is given.

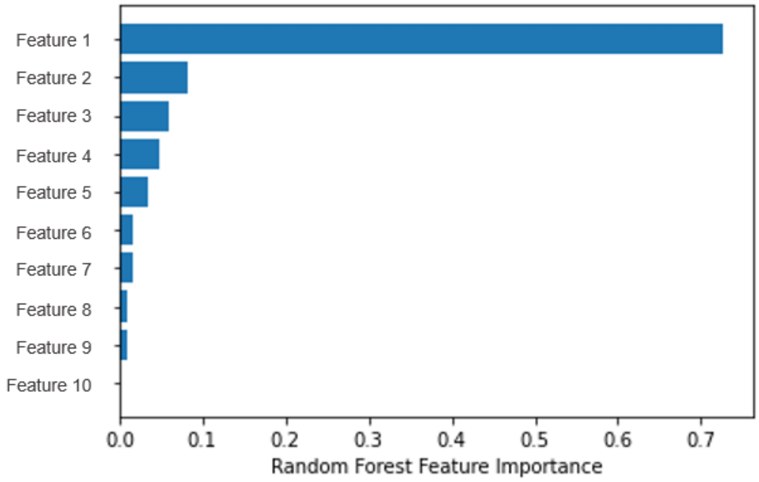


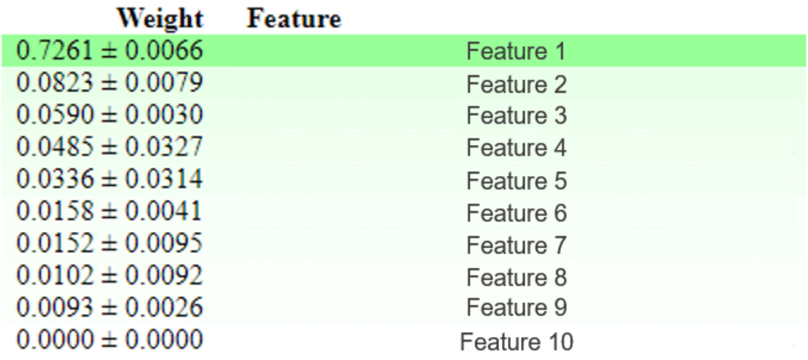
Figure 14 - Random Forest feature importance with selected variables by F regression

ELI5

By using specific Python package ELI5[4] for inspecting different machine learning models and explaining their predictions, developers are able to analyze the influence of each variable on the model output and understand its overall performance and the estimator's performance for a particular sample could be seen and compared to what combination of features and values leads to a particular prediction, which is why ELI5 was exploit in this task, as well. ELI5 is popularly used to debug algorithms such as sklearn

[4] https://eli5.readthedocs.io/en/latest/overview.html

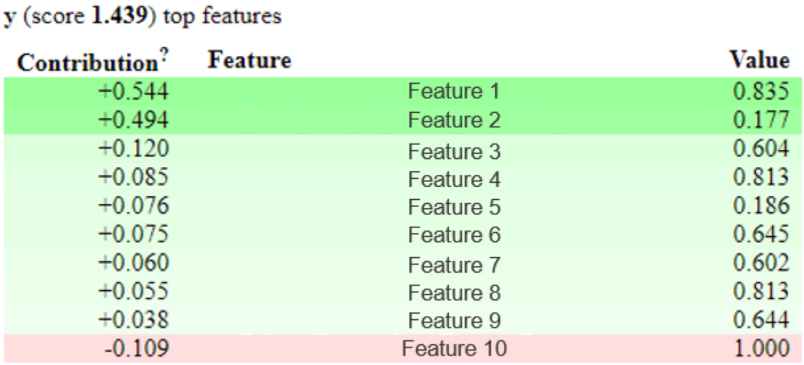
regressors and classifiers, XGBoost, CatBoost, Keras, etc. For the specific case selected, in Figure 15, the contribution of different variables to estimating the value of the Extruder 2 Temperature is given.



Weight	Feature
0.7261 ± 0.0066	Feature 1
0.0823 ± 0.0079	Feature 2
0.0590 ± 0.0030	Feature 3
0.0485 ± 0.0327	Feature 4
0.0336 ± 0.0314	Feature 5
0.0158 ± 0.0041	Feature 6
0.0152 ± 0.0095	Feature 7
0.0102 ± 0.0092	Feature 8
0.0093 ± 0.0026	Feature 9
0.0000 ± 0.0000	Feature 10

Figure 15 - Weights of different signals for the prediction of the variable temperature in extruder 2 in the Random Forest model

As can be observed from the output above, ELI5 shows the contribution of each feature in predicting the outcome. It could be also compared which combination of features and values leads to a particular prediction. The target value is: 0,1778700365665143.



y (score **1.439**) top features

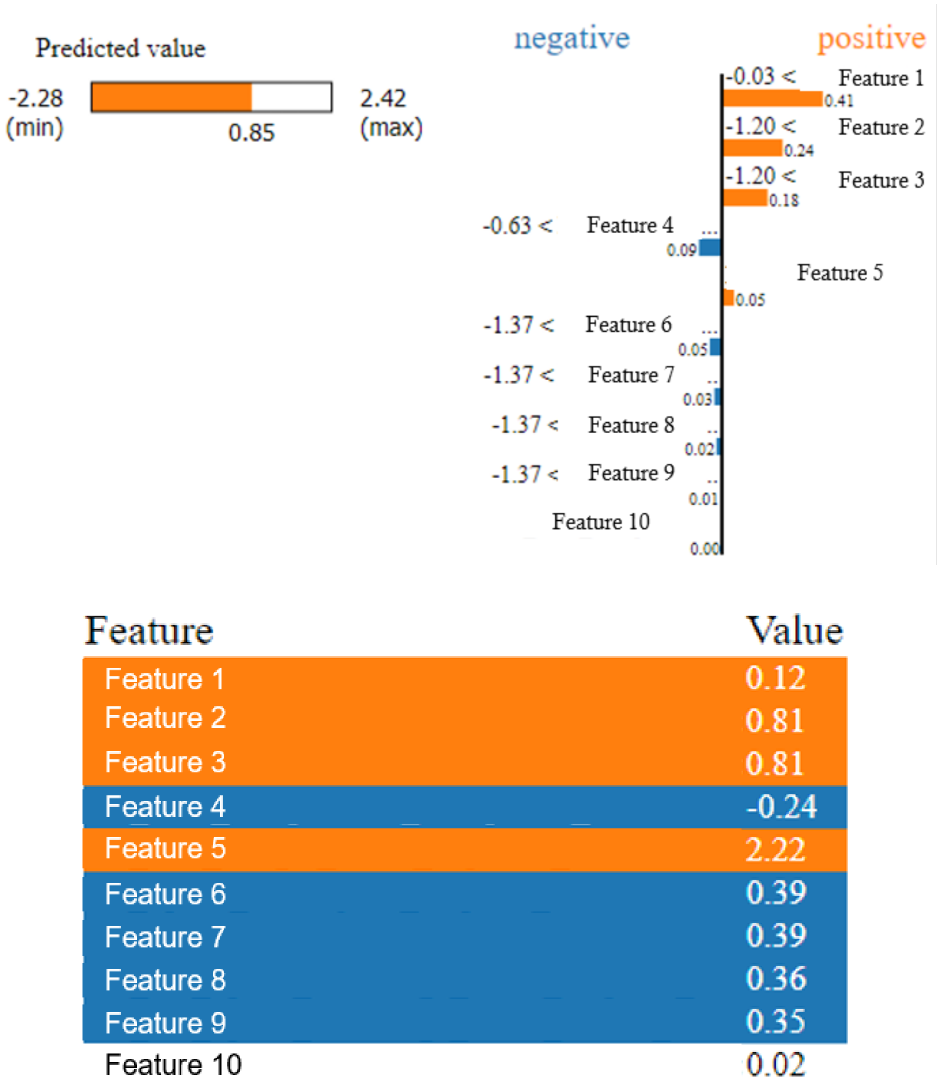
Contribution?	Feature	Value
+0.544	Feature 1	0.835
+0.494	Feature 2	0.177
+0.120	Feature 3	0.604
+0.085	Feature 4	0.813
+0.076	Feature 5	0.186
+0.075	Feature 6	0.645
+0.060	Feature 7	0.602
+0.055	Feature 8	0.813
+0.038	Feature 9	0.644
-0.109	Feature 10	1.000

Figure 16 - Contribution and real value of each signal for one prediction of the variable temperature in extruder 2

LIME

The second XAI technique tested was LIME. As explained in the State-of-the-art section, LIME generates fake data using our input data, trains a simple ML model that performs similarly to our complex black-box model, and uses this model's weights to describe features' importance.

In this particular case, the local explanation for one prediction is obtained. It could be noticed that the visualization has a progress bar, bar chart, and table. The progress bar shows the range in which the value varies and the actual prediction. The bar chart shows features that contributed positively and negatively to prediction. The table shows actual feature values.



Feature	Value
Feature 1	0.12
Feature 2	0.81
Feature 3	0.81
Feature 4	-0.24
Feature 5	2.22
Feature 6	0.39
Feature 7	0.39
Feature 8	0.36
Feature 9	0.35
Feature 10	0.02

Figure 17 - Output of LIME explanation for one prediction.

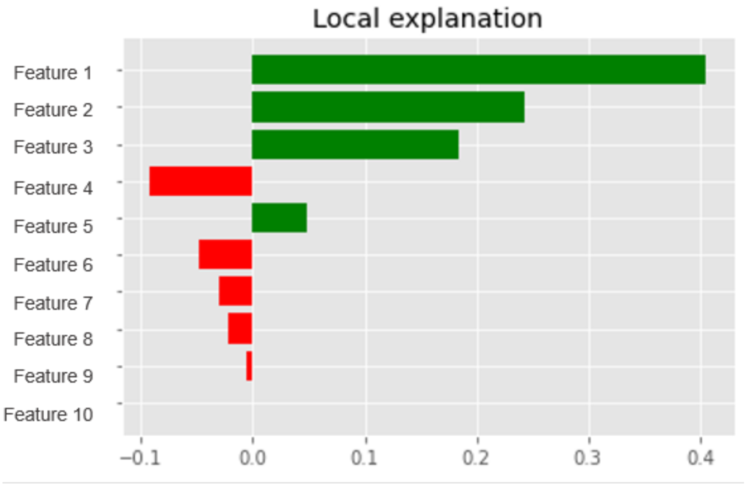


Figure 18 - Output of LIME approach sorted by the most influential inputs on the output

With this analysis the most influential signals on the target variable are obtained. For example, in the case of extruder 2 temperature, the increase of one unit of the feature 1, the normalized output is expected to increase by 0,41. Meanwhile, when the feature 4 increases in one unit of the normalized input value, the expected output decrease by 0,09.

SHAP

SHAP is the third and last XAI technique to compare the results of explainable methods on this RF model. It is another way of interpreting the predictions of the ML models through the Shapley values. The key idea of SHAP is to calculate the Shapley values for each sample feature to be interpreted. Each Shapley value represents the impact that the feature with which it is associated generates in the prediction.

SHAP allows users to obtain the relationship of all the variables with the model and their impact. The graph below shows the SHAP value of each feature in the training set. The importance of the variables is listed from highest to lowest, the first being the most relevant for the model.

In Figure 19, we can see the feature 1 as the most relevant for the forecast model. The concentration of SHAP values for this variable in red or blue indicates how the different values of the variable affect the prediction. The parameters in red are the ones that make the prediction have a higher value, while the parameters in blue make have a lower value. This relationship can be better appreciated later when we analyze one prediction individually.

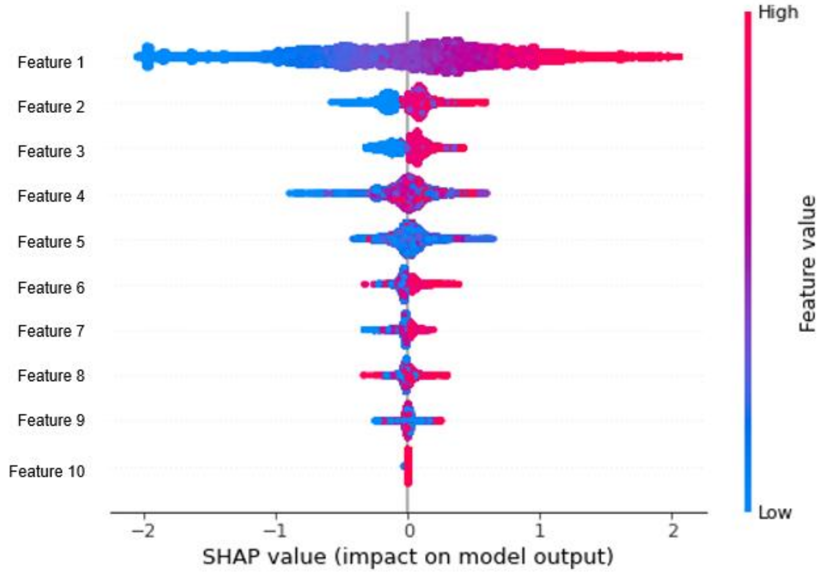


Figure 19 - Output of the SHAP impact on the Random Forest model

It also allows to see the relationship of a single input parameter concerning the model's predictions. These charts also show the variable with more interaction with the parameter analyzed. For example, in the first graph, it could be seen that feature 1, has a linear negative relationship with the target (SHAP values), and the variable with which it interacts the most is feature 3.

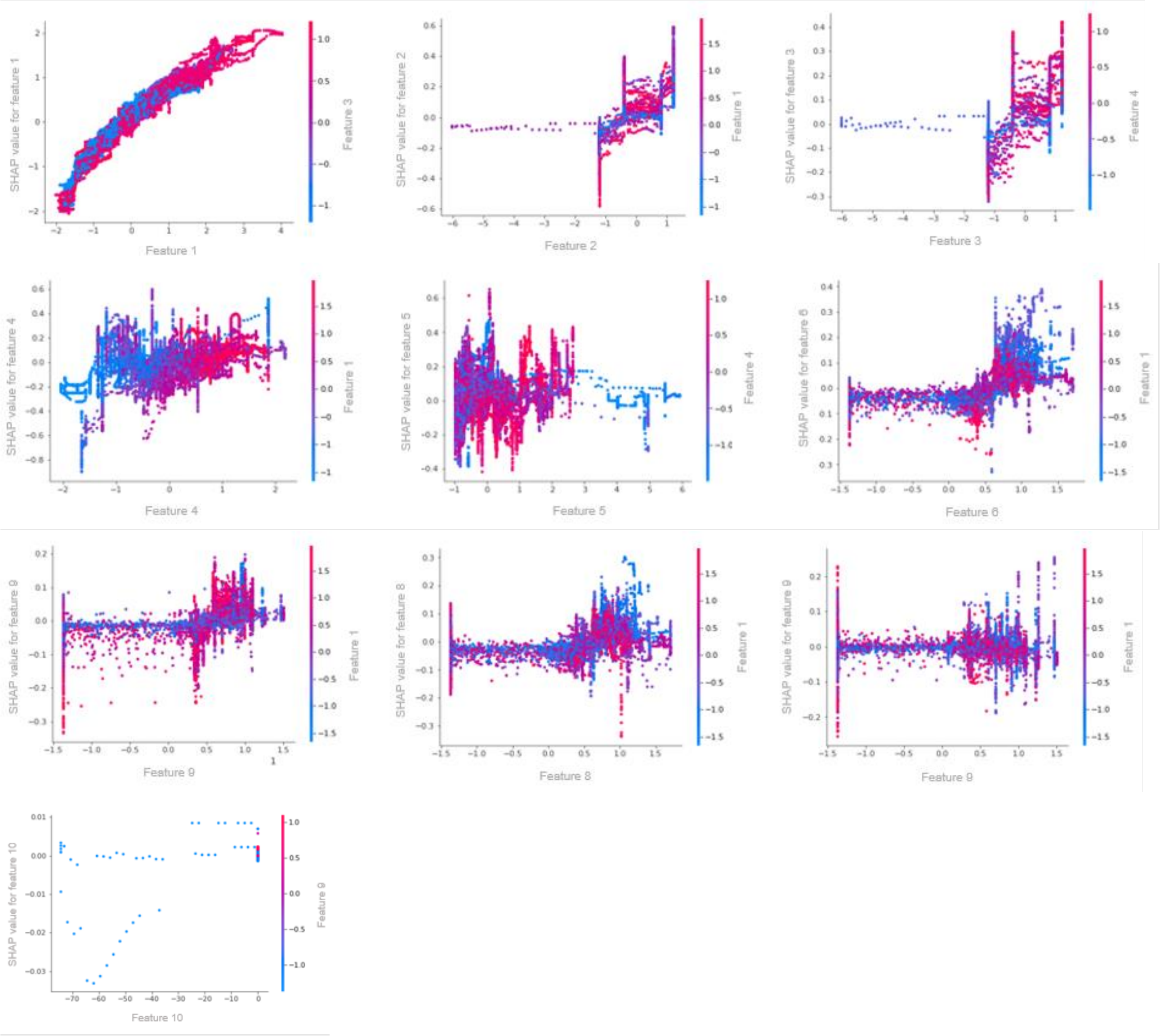


Figure 20 - Impact of each variable with the target value.

Another of SHAP's advantages is that it allows for explaining the value of the model's parameters for a particular prediction. In Figure 20, we see the *base value = –0.10913911,* which is the average prediction in the training dataset. It is the base value with which the model works. At the same time, *f(x) = -0.77* is the model's prediction for that specific dataset input, with the characteristics that were delivered to it. The parameters in red are the ones that make the forecast have a higher value, while the parameters in blue make the prediction have a lower value. The values in blue have a greater weight in this case than those in red, which is why the value f(x) is less than the base value of the model.

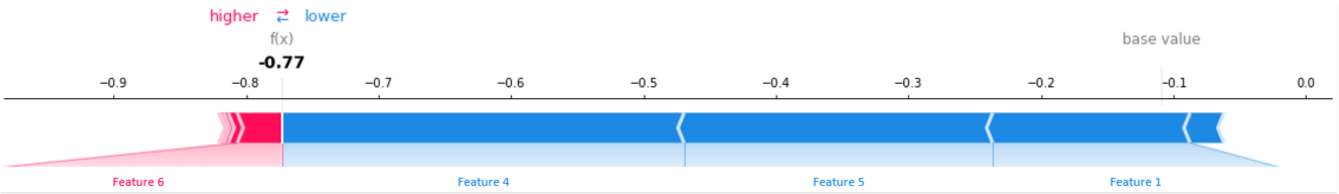


Figure 21 - SHAP explanation for one prediction.

After having implemented the three techniques, we consider that the most complete one to show the results would be SHAP, with the drawback that its execution time when we look for the relationships in the global model can be pretty high. However, predictions of a particular point take the same time as the other two methods. Its visualizations are far above the rest because they provide much more information that allows easier visual understanding. A summary Table 7 was created to understand the differences between the three explainable methods analyzed.

Table 7 - Comparison between the different XAI techniques.

XAI technique / Features	ELI5	LIME	SHAP
Shows the relationships of the variables with the model in general.	X		X
Shows the relationships of the variables with a specific prediction.	X	X	X
Execution Time	Low	Low	High (to explain relationships with the model in general) Low for a prediction.
Plots	No. It is only possible to display a list with the weights, contributions and features values of the variables.	Bar Plots	Good graphics and visualizations.

Although in the future, the idea is to offer the quality manager in the dashboard the possibility of choosing how he/she wishes to analyze the results of the predictive model, for the first deployment version, the solution has been designed choosing to represent the feature selection f regression with a Random Forest model to explain the target variable and a SHAP model that allows interpreting the results of the black box model.

3.2 Semantic technologies (TEK)

In AI-PROFICIENT, based on a semantic approach, FIDES[5] Ontology has been developed. FIDES aims to represent, structure, and set formal relations among the ML-based models and the forecasts/suggestions that conform a ML system, disposing all the necessary information to answer all the pertinent questions in terms of the systems' accountability. For developing FIDES ontology, LOT (Linked Open Terms) methodology has been followed. This methodology follows two main steps for development: a first step of requirements specifications and a second implementation step. And two additional steps for publication and maintenance aspects.

FIDES envisions the representation of two main knowable topics: the forecast/suggestion made by the ML-based model, and the procedure followed by such a ML-based model for making the forecast/suggestion. To formalise the information requirements for each knowable topic, the Competency Questions (CQ) were used as proposed by LOT. For CQs definition, a team of 4 AI experts and 2 ontologists met several times.

For the procedure followed to construct the ML-based model for making forecasts, the team established that the relevant information could be divided into, on the one hand, the information addressing the data

[5] Fides was the Roman goddess of trust.

used to train the ML-based model, and on the other, the information concerning the details of the procedure implemented by the ML-based model. The training data can be characterized by its features, including the amount of data used, the dependent and independent variables considered, as well as statistical characteristics such as the variance, mean or median of the data.

The complete implementation of FIDES is available in https://w3id.org/fides. For this, the new classes and properties have been first properly documented, the corresponding checking ensuring that no syntactic or semantic error has been performed, and using WIDOCO tool, the necessary documentation have been generated.

To validate the ontology, CONTI2 and CONTI5 use cases predictive model development has been used to populate the ontology and validate that all the necessary information can be properly represented according to FIDES to get a correct answer for the defined CQs, at least for the model development procedure. The validation for the prediction/suggestion part will be further performed, when the model is running and providing forecasts.

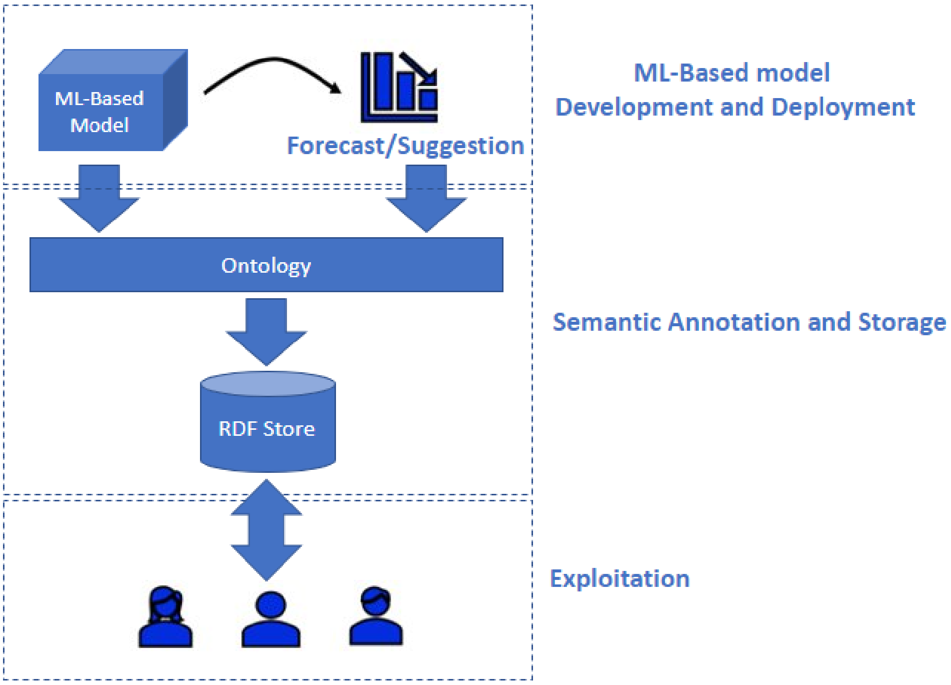

ML-Based Model
Forecast/Suggestion
ML-Based model Development and Deployment
Ontology
Semantic Annotation and Storage
RDF Store
Exploitation

Figure 22 - AI-Models Accountability Semantic Approach

All in all, FIDES ontology is intended to be used in the context of T4.4, as the core element of a semantic framework for AI-models accountability. The semantic approach consists of three phases as shown in Figure 22.

In the first phase, the predictive model that will solve the problem at hand is developed. Depending on the type of the problem addressed, and the quality and the amount of the data available, some algorithms may provide better results than others. Furthermore, the adequate fine-tuning of the hyperparameters of the algorithms may have a direct effect on the performance of the final model. Therefore, at this stage, the data scientists in charge of developing the model will need to make the opportune choices. Once the model is generated, it is deployed into production in order to generate the aimed forecasts. FIDES is model-agnostic with a view to be valid for the wide variety of existing ML algorithms, and it works with predictive models developed in R programming language and deployed in Rserve[6], a server that allows to execute R implementations.

In the second phase, the relevant information related to the developed predictive model and the generated forecasts are retrieved from R and Rserve respectively, and annotated with the adequate ontological terms. Then, the resulting RDF triples are automatically stored in an Openlink Virtuoso[7] repository. Both the retrieval and semantic annotation of the data, and the storage of the RDF triples is fully automated with a service based on Apache Jena[8], so there is no need for human intervention. The main goal of this service is to minimise potential performance issues and errors derived from manual practices.

In the third phase, the end-users are provided with a user interface that facilitates the retrieval of the information that helps to make ML systems accountable. This user interface implements a set of API methods developed in Python FastAPI, which in turn execute a set of predefined parameterizable SPARQL queries, thus abstracting end-users from the underlying query language.

3.2.1 Predictive Model Development and Deployment phase

The predictive models have been developed by a team of data scientists in the R programming language. CONTI2 have been implemented using a Random Forest Classifier, and in case of CONTI2 has been used a custom prediction implementation. The developed predictive models have been exported in the form of .rds files and put into production in an Rserve version 3.2.5 deployed in a Docker[9] container.

3.2.2 Semantic Annotation and Storage phase

The predictive models have been generated and deployed in Rserve, their corresponding RDF triples have been automatically generated and stored in the Virtuoso Open-Source Edition version 7.2.5.1 repository. Likewise, each time a forecast has been generated, its corresponding RDF triples have been automatically generated and stored in the same repository.

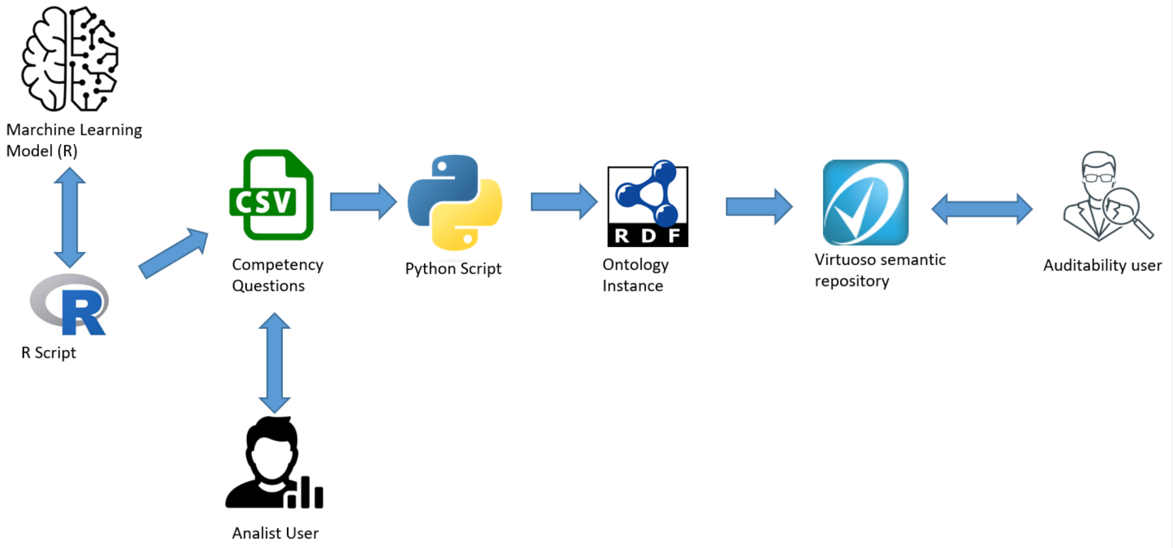


Figure 23 - Auditability system pipeline

[6] https://www.rforge.net/Rserve/
[7] https://virtuoso.openlinksw.com/
[8] http://jena.apache.org/
[9] https://www.docker.com/

Data Exploitation phase

Once the forecasts and details of the procedure used by predictive models to generate such forecasts are semantically annotated and stored in the RDF Store, FIDES make use of a user interface to let end-users interact with this information. FIDES implements and API REST which exposed the different questions grouped in 9 sets:

- Development system information
- Model authoring
- Model implementation
- Model metadata
- Information related to dataset
- Deployment information
- Model execution meta data
- Performance metrics
- Performance metrics values

For instance, the SPARQL query to get development system information for CONTI2 is show in next table.

Table 8. SPARQL sentence for Development system information in CONTI2 Random Forest model

DevSystemInformation

```
DEFINE input:inference 'urn:ai-proficient-auditability'

prefix mls: <http://www.w3.org/ns/mls#>

prefix fides: <https://w3id.org/fides#>

prefix dcterms: <http://purl.org/dc/terms/>

prefix :  <http://www.aiproficient-auditability#>

SELECT ?Software ?Version ?OperatingSystem from <urn:ai-proficient-auditability> WHERE {

        ?r a mls:Run.

        ?r mls:hasOutput :CONTI2RFClassifier20220620.

        ?r mls:executes ?i.

        ?s mls:hasPart ?i.

        ?s rdfs:label ?Software.

        ?s fides:hasOperatingSystem ?OperatingSystem.

        ?s dcterms:hasVersion ?Version.

}
```

JSON Result

```
{
    "head": {
        "link": [],
        "vars": [
            "Software",
            "Version",
            "OperatingSystem"
        ]
    },
    "results": {
        "distinct": false,
        "ordered": true,
        "bindings": [
            {
                "Software": {
                    "type": "literal",
                    "value": "python 3.6.13 in windows7"
                },
                "Version": {
                    "type": "literal",
                    "value": "3.6.13"
                },
                "OperatingSystem": {
                    "type": "literal",
                    "value": "windows 10"
                }
            }
        ]
    }
}
```

3.2.3 Evaluation

The usability of FIDES has been measured with the SUS (System Usability Scale). It consists in a questionnaire with ten questions, where participants are asked to score them with one of five responses that range from Strongly Agree (5 points) to Strongly disagree (1 point). It allows to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites and applications, and it has become an industry standard. The SUS questionnaire has been used after participants have interacted with FIDES at least once and before any discussion took place. Furthermore, as suggested by the methodology itself, participants have been asked to record an immediate response to each question, rather than thinking about items for a long time. The average score obtained was 80.8 out of 100, so it can be concluded that the overall usability of FIDES is very good. The most remarkable outcomes of this questionnaire are that all participants think that they would use FIDES frequently as it is easy to use and quick to learn.

4 Integration with the AI-PROFICENT platform and service deployment

In this section, details regarding XAI service integration and deployment will be given, as the one final step before the exploitation of the XAI services presented in the previous section.

4.1 Surrogate explainable data-driven model (IMP)

Various approaches for SDDM model have been tested and compared in section 3.1.1, with the goal of achieving a high quality of the performance on the shop floor. Finally, the last step that was crucial was SDDM integration with the rest of the AI-PROFICIENT platform. Three crucial components from the integration point of view are presented – ***generative holistic optimizer (GHO), data storage*** and ***human machine interface (HMI),*** as given in Figure 24 from D5.5. As already presented, SDDM was developed with the main goal of serving GHO. Since SDDM is used for fitness function evaluation, it was necessary to enable easy and fast communication between these two models, so that GHO could provide outputs in as short a time as possible. Therefore, integration between the two was achieved through API. Furthermore, SDDM inputs are obtained from AI-PROFICIENT cloud platform, and, therefore corresponding integration has been carried out. Presentation of SDDM outputs on HMI is achieved through a service output relational data base. Finally, service, developed in Python, was successfully deployed with all the corresponding dependencies on AI-PROFICIENT server using Docker environment.

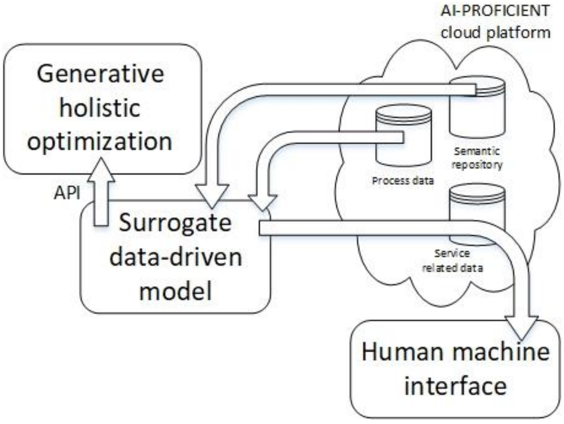

AI-PROFICIENT cloud platform
Generative holistic optimization
API
Semantic repository
Process data
Surrogate data-driven model
Service related data
Human machine interface

Figure 24 - Integration of SEDDM with AI-PROFICIENT platform

4.2 Post-hoc explainable analysis module (IBE)

For post-hoc explainable analysis service, all modules are implemented in Python using specific libraries for machine learning and explainable artificial intelligence. They are designed to connect to a MySQL database, read the preprocessed data, and then store the corresponding results of each submodule in the MySQL database itself, as shown in Figure 25.

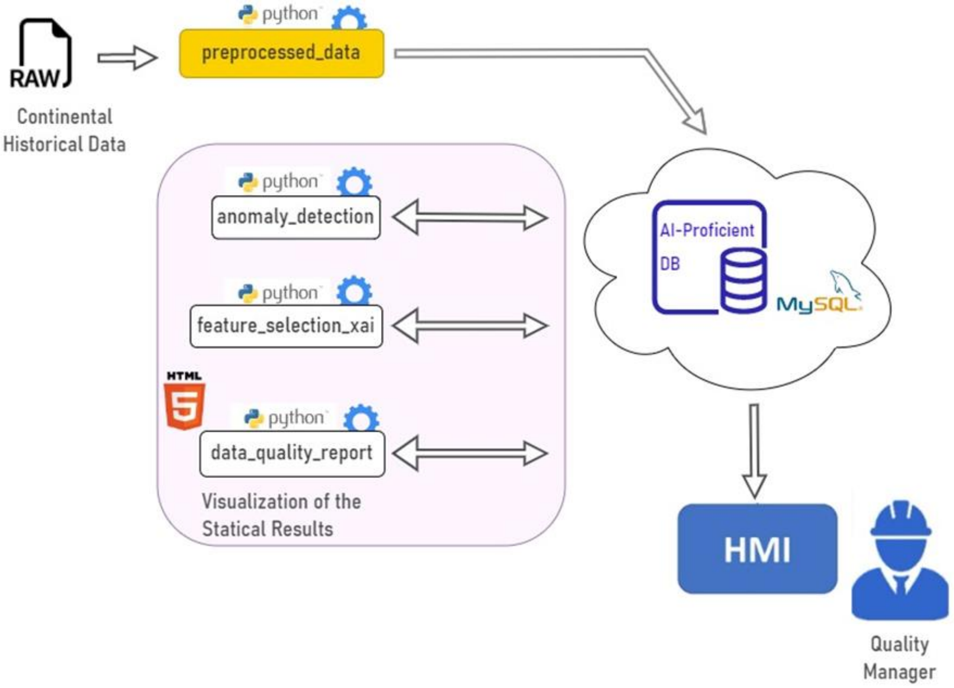
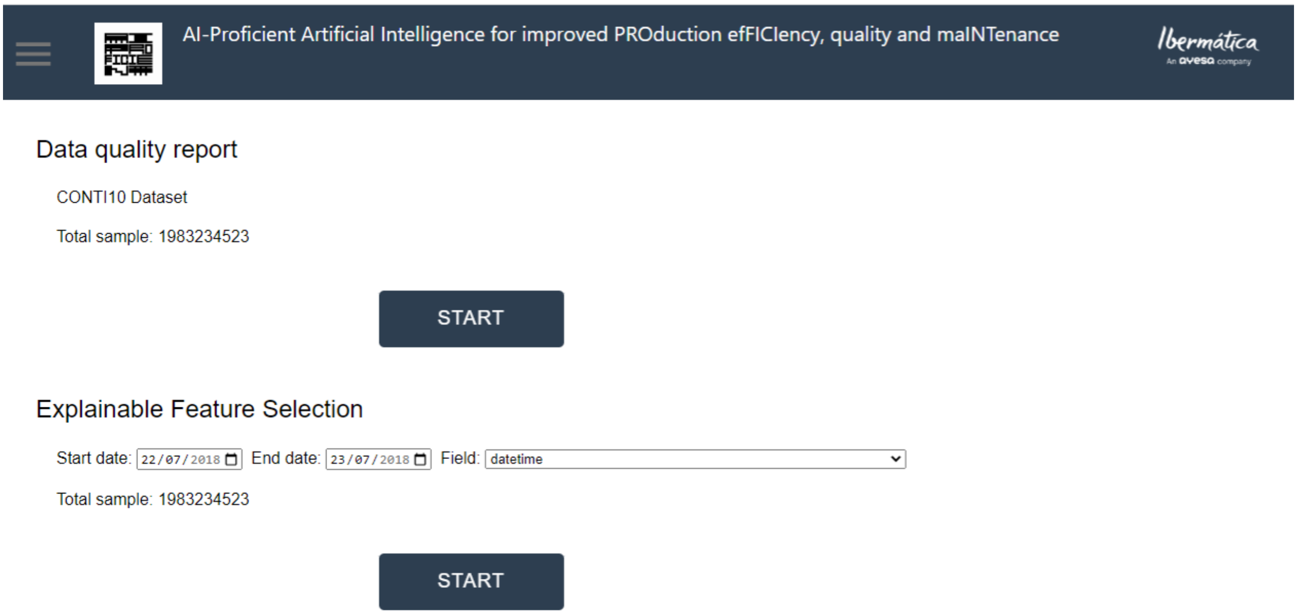


Figure 25 - Integration of PEAA with AI-PROFCIENT platform

It is intended to design an interface that allows the user to choose the variable and the period to be analyzed. It is currently in a pilot version under development, and is given in Figure 26.



AI-Proficient Artificial Intelligence for improved PROduction efFICIency, quality and maINTenance

Data quality report

CONTI10 Dataset

Total sample: 1983234523

Explainable Feature Selection

Start date: 22/07/2018 End date: 23/07/2018 Field: datetime

Total sample: 1983234523

Figure 26 - PEAA service user interface

4.3 Auditability system (TEK)

FIDES is a tool that uses ontologies for representing, structuring, and setting formal relations among the predictive models and the forecasts that conform a ML system, and provides end-users with the necessary means to exploit this knowledge and answer the pertinent questions.

The different questions have been grouped in 7 different sections. Each section or question set, is exposed through an API Rest, so that functionality can be called from AI-PROFICIENT platform, providing information about development, data, deployment and execution of machine learning module.

Moreover, an user interface is under development, providing the information directly to the user. A pilot version is shown in Figure 28.

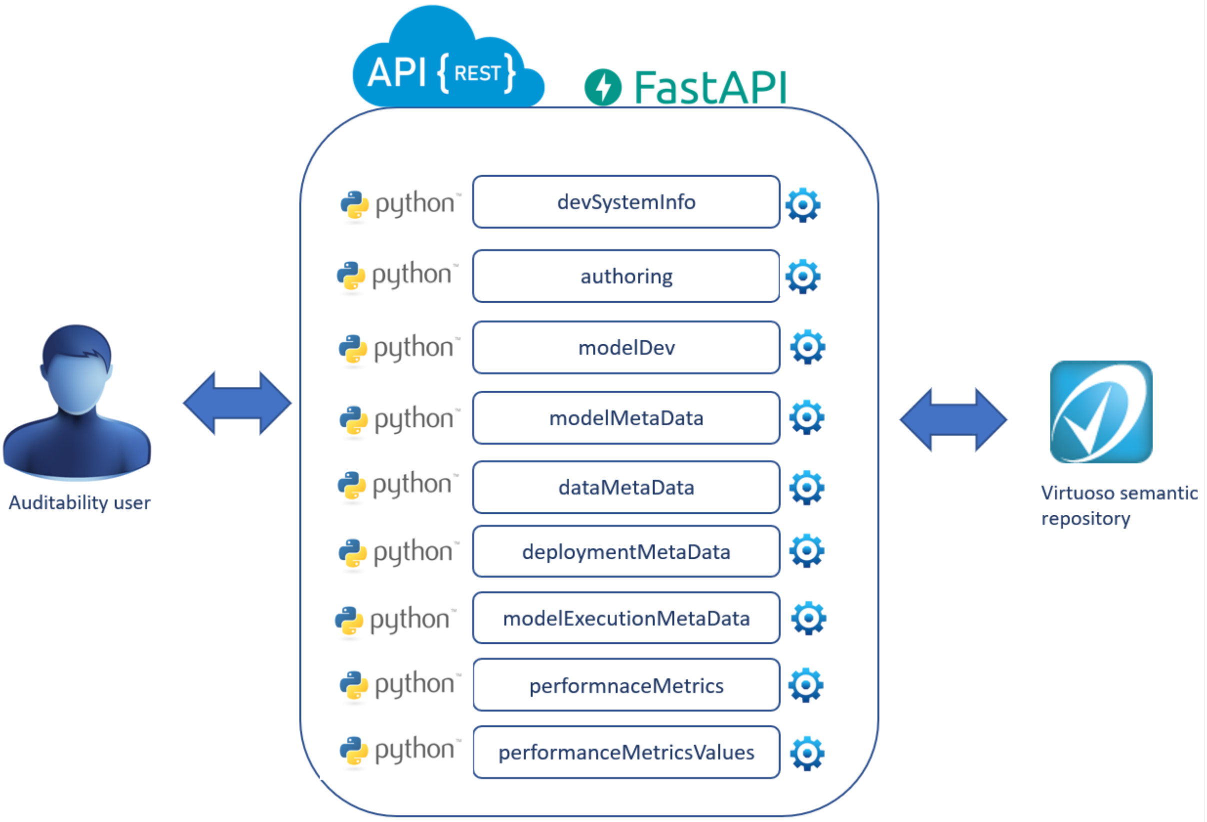
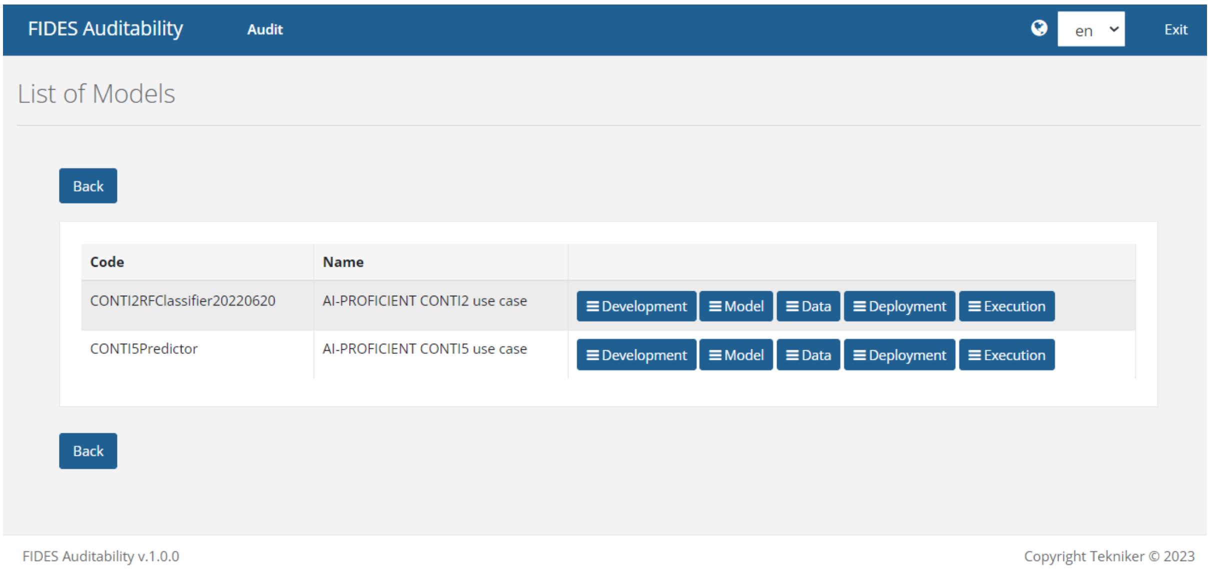


Figure 27 - API Rest services of FIDES



List of Models

Back

Code	Name	
CONTI2RFClassifier20220620	AI-PROFICIENT CONTI2 use case	☰Development ☰Model ☰Data ☰Deployment ☰Execution
CONTI5Predictor	AI-PROFICIENT CONTI5 use case	☰Development ☰Model ☰Data ☰Deployment ☰Execution

Back

FIDES Auditability v.1.0.0

Copyright Tekniker © 2023

Figure 28 – User interface for FIDES

5 Ethical considerations (UL)

Task 4.4. was to develop explainable and transparent AI related to services for the various Use Cases. Conti 2, Conti 10, and Ineos 3 were eventually selected by the partners as being suitable for the goals of the deliverable and analysis of the ethical aspects. For Conti 5, models will not been used by plan employees, so they have not been analyzed from ethical aspect.

The definitions of the terms remain problematic. Explainability and transparency are evoked in the task, but numerous researchers affirm that there is no firm consensus on the definitions of explainability and transparency. Sokol and Flach in [57] show that nearly two dozen terms have been used in the research literature, often interchangeably, around the notion of the human understanding of algorithms.

Interpretability is often spoken of as an alternative to explainability for example. Doshi-Velez and Kim in [58] implicitly understand explainability as a subsidiary aspect of interpretability. For them a machine learning model is interpretable if it can 'explain itself' or present its process in an understandable way to a human. Authors in [59] agree that interpretability encompasses explainability, with the former being passive and the latter a more active mode of explanation, e.g., deliberate post-hoc explanation. More recently [60], drawing upon psychological research defines *explainable* as: *a detailed presentation of the mechanistic process from input to output regardless of context*, in distinction from *interpretable*: *what does the output mean to the user relative to a specific context, in terms of what the algorithm was supposed to do*.

Transparency, meanwhile, is sometimes taken to be the amenability of a model to being understood by a human. For Lipton [61] this means the model is amenable to being worked though by a human in order to understand its mechanism: the human can break down the mechanism conceptually, understand its parts, its process, and conceptually simulate that process. For authors in [62], similarly, the degree of exposure of the system's inner workings is key to transparency. But again, consensus is hard to find. [59] for example, dwelling on the notion of passivity, equate transparency with interpretability, which on some readings, puts in doubt the separation of transparency from explainability.

Regardless of the difficulty of defining explainability and transparency precisely, the ethics team has chosen to focus on human centeredness as a point around which these efforts were conducted. The aim, as in other deliverables, has been *to encourage the interaction of the relevant developer partners with the primary users in the specific Use Cases* – the operators and process managers – but also *to demonstrate by the example of this task of the project that a practical, consultative, and iterative development is possible for explainability*.

Developing the technical aspects of the task according to a human centered approach involves a number of elements which were communicated early on. All of these elements clustered around the assumption that: 'the explanation is for the primary user' (the operator and process engineer)[10]. From this, given the stated aim of the project to provide such explainability, the ethical question was asked: *what does that user practically require for explanations and their context, to be satisfying*? The answers were then used to make specific recommendations and to follow up on their implementation. At a theoretical ethical level this approach operationalizes part of the spirit of discourse ethics as developed by Habermas, i.e., develop and tend to the conditions for ethical communication among a community. It facilitates community between developers and users and also among the developers themselves. It is thus general in the focus of its effects – the user context – but also amenable to very specific recommendations.

Available research indicates that there is a tradeoff between accuracy and transparency in AI model uses. But since explainability on most readings depends upon transparency, then the level of accuracy desired or decided upon will limit what can be explained. Thus, we recommended generally to decide

[10] We did not put emphasis on the data scientists as users, judging that their development of explanatory and transparency methods would have to satisfy themselves in any case since they would use such methods to make corrections to the models.

upon levels of accuracy first and then use up the room for maneuver for maximal transparency relative to that accuracy.

Further, the background of the user has been shown to change how they see the explanation of the model [63], with the user potentially making unwarranted ascriptions of intention to the AI and placing unwarranted trust in its 'intelligence.' Consequently, we recommended to gain some knowledge of the background of the user and of their background knowledge in order to decide between choices of post-hoc explanation.

Coherence of explanations is also important, as noted by authors in [64], with contradictions between instances of explanation engaged through an iterative process with the user to resolve them. This leads to the obvious recommendation of modifying the explanation design to account for the user's perception of intent in the AI and clarifying that the AI process is not intelligent. But if a full iterative process proves to be difficult timewise, e.g., the operators and process engineers haven't got time for it, then a presentation showing that the AI processes are not intelligent was recommended as a fallback.

The following specific recommendations were given relative to the above. The lead and industrial partners have carried out several recommendations and are currently carrying out others in the form of question and explanation sessions with operators and process engineers. Both groups are to be asked which post-hoc explainability models they understand best and development tailored accordingly. The ethics team has worked with the lead and industrial partners for Task 4.4 to formulate and tailor the question sessions. It is foreseen that the implementation of the Task 4.4 ethics component will continue to evolve beyond the submission of this deliverable, through the final year of the project. The questionnaire that was developed as a part of work within Task 4.4 is given in appendix.

Table 9 - Ethical issues for Explainable and transparent AI decision making

Recommendations	**Responsible**
ETHICS 4.4-1) Recommend that you discuss directly and regularly (once a month) with the process engineers most closely involved with each of the UCs under consideration as you develop the transparency models. Recommend that the process engineers discuss similarly with the operators who will be most closely involved	All task partners
ETHICS 4.4-2) Recommend that you formally clarify who will be the user(s) of the explanations for each UC (e.g., data scientists, process engineers, operators), on an individual level if possible	All task partners
ETHICS 4.4-3) Recommend that in discussion with the process engineers you clarify what level of accuracy of the AI is acceptable for each UC and then you decide which options for explainability methods remain open based upon that	All task partners
ETHICS 4.4-4) Recommend that you carry out a preliminary short survey, e.g., 10 questions, of user background knowledge (process engineers and operators) regarding AI, to be used in adjusting for potential user assumptions during XAI development	IMP; Conti; Ineos
ETHICS 4.4-5) Recommend that for all explainability methods destined for process engineers and operators, you review the models directly with the process engineers and/or operators as soon as possible after the prototype stage (or with a mock input and result), asking them directly: "do you understand this method generally?" and then adjust for their concerns	IMP; TEK; IBER

ETHICS 4.4-6) Recommend that, after you advance beyond the prototype stage you pursue an iterative development of explainability methods if project timeline allows ETHICS 4.4-7) Recommend that if project timeline does not allow for the iterative development (recommendation 4.4-6), you have several sessions with the process engineers and operators where you present clear mechanistic explanations which characterize the AI processes as un-intelligent tools	All task partners IMP; TEK; IBER

6 Conclusion (IMP)

In recent years, there has been a growing interest in XAI (Explainable Artificial Intelligence) methodologies, which aim to provide insights into the predictions and decisions made by deep learning models. These methodologies are important for several reasons. First, they can help improve the transparency and accountability of deep learning models, by allowing users to understand how AI models are making their predictions and decisions. This is especially important for applications where the model's outputs have significant consequences, which is the case with AI-PROFICIENT pilot sites. Second, XAI can facilitate the debugging and improvement of deep learning models, by identifying the input features that are most important for the predictions and highlighting potential sources of bias or error and providing an ontology-based framework which easier the development process. Third, XAI can enhance the usability and interpretability of deep learning models, by providing a human-understandable explanation of the model's output, which can be especially valuable for domain experts or non-technical users.

In this report a state-of-the-art analysis of the available XAI techniques was given and in accordance with the corresponding analysis different XAI models were developed and presented in this report. Surrogate explainable data driven model was designed to facilitate generative optimization and thus is envisioned as explainable forecasting model to be tested in three use cases – CONTI2, CONTI10 and INEOS3. The models have been developed using on site plant data and corresponding promising results were presented. The combinations of different ML approaches such as random forests, neural networks, support vector machines, kNN, etc. was merged and tested with different XAI approaches (LIME, DeepLIFT). The models showed high prediction precision and potential for integration with the optimization engine, as well as ability to be used as the root cause identification module in order to determine which process parameters are influencing the product quality degradation. Furthermore, post-hoc explainable analysis module was designed in order to be able to analyze historical product quality degradation and its causes. Hence, in the report a framework for historical data analysis is presented, together with the XAI service for detecting likely cause of previous degradation. For those purposes, additional XAI methods were tested, such as SHAP and ELI5. Finally, in order to encourage further advancement of the X(AI), AI-PROFICIENT offered auditability system which is intended to easy the development process of ML solutions to the developers and data scientist based on the semantic technologies.

All in all, this deliverable presented three XAI services for improving product characteristics applied in two different industry domains – manufacturing and chemical industry followed by the additional service for the developers in order to increase penetration of XAI technologies in the coming years.

7 Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 957391.

8 References

[1] S. S. ÓhÉigeartaigh, J. Whittlestone, Y. Liu, Y. Zeng, and Z. Liu, “Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance,” *Philos. Technol.*, vol. 33, no. 4, pp. 571–593, Dec. 2020, doi: 10.1007/s13347-020-00402-x.

[2] M. Cubric, “Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study,” *Technology in Society*, vol. 62, p. 101257, Aug. 2020, doi: 10.1016/j.techsoc.2020.101257.

[3] G. Baryannis, S. Validi, S. Dani, and G. Antoniou, “Supply chain risk management and artificial intelligence: state of the art and future research directions,” *International Journal of Production Research*, vol. 57, no. 7, pp. 2179–2202, Apr. 2019, doi: 10.1080/00207543.2018.1530476.

[4] E. Broadbent, R. Stafford, and B. MacDonald, “Acceptance of Healthcare Robots for the Older Population: Review and Future Directions,” *Int J of Soc Robotics*, vol. 1, no. 4, p. 319, Oct. 2009, doi: 10.1007/s12369-009-0030-6.

[5] L. Laranjo *et al.*, “Conversational agents in healthcare: a systematic review,” *J Am Med Inform Assoc*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018, doi: 10.1093/jamia/ocy072.

[6] J. T. Ingibergsson, U. P. Schultz, and M. Kuhrmann, “On the Use of Safety Certification Practices in Autonomous Field Robot Software Development: A Systematic Mapping Study,” in *Proceedings of the 16th International Conference on Product-Focused Software Process Improvement - Volume 9459*, Berlin, Heidelberg, Dec. 2015, pp. 335–352. doi: 10.1007/978-3-319-26844-6_25.

[7] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: evaluating claims and practices,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 469–481. doi: 10.1145/3351095.3372828.

[8] J. Sánchez-Monedero, L. Dencik, and L. Edwards, “What does it mean to ‘solve’ the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 458–468. doi: 10.1145/3351095.3372849.

[9] S. Chuang and C. M. Graham, “Embracing the sobering reality of technological influences on jobs, employment and human resource development,” *European Journal of Training and Development*, vol. 42, no. 7/8, pp. 400–416, 636715296000000000, doi: 10.1108/ejtd-03-2018-0030.

[10] M. Madsen and S. Gregor, “Measuring Human-Computer Trust,” in *11th Australasian conference on information systems (ACIS)*, 2000, vol. 53, pp. 6–8.

[11] J.-Y. Jian, A. Bisantz, and C. Drury, “Foundations for an empirically determined scale of trust in automated systems,” *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53-71., 2000, doi: 10.1207/S15327566IJCE0401_04.

[12] B. Cahour and J.-F. Forzy, “Does projection into use improve trust and exploration? An example with a cruise control system,” *Safety Science*, vol. 47, no. 9, pp. 1260–1270, Nov. 2009, doi: 10.1016/j.ssci.2009.03.015.

[13] D. Gunning, “Explainable artificial intelligence (xai).” Defense Advanced Research Projects Agency (DARPA), 2017.

[14] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, “Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI.” arXiv, Feb. 05, 2019. doi: 10.48550/arXiv.1902.01876.

[15] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, “Evaluating XAI: A comparison of rule-based and example-based explanations,” *Artificial Intelligence*, vol. 291, p. 103404, Feb. 2021, doi: 10.1016/j.artint.2020.103404.

[16] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, “Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation,” in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, Maastricht, Aug. 2018, pp. 1–8. doi: 10.1109/CIG.2018.8490433.

[17] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” arXiv, Aug. 09, 2016. doi: 10.48550/arXiv.1602.04938.

[19] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions.” arXiv, Nov. 24, 2017. doi: 10.48550/arXiv.1705.07874.

[20] A. Datta, S. Sen, and Y. Zick, “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems,” in *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016, pp. 598–617. doi: 10.1109/SP.2016.42.

[21] A. Henelius, K. Puolamäki, and A. Ukkonen, “Interpreting Classifiers through Attribute Interactions in Datasets.” arXiv, Jul. 24, 2017. doi: 10.48550/arXiv.1707.07576.

[22] J. R. Zilke, E. Loza Mencía, and F. Janssen, “DeepRED – Rule Extraction from Deep Neural Networks,” in *Discovery Science*, Cham, 2016, pp. 457–473. doi: 10.1007/978-3-319-46307-0_29.

[23] L. Fu, “Rule generation from neural networks,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 8, pp. 1114–1124, Aug. 1994, doi: 10.1109/21.299696.

[24] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[25] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep Taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[26] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-Wise Relevance Propagation: An Overview,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Cham: Springer International Publishing, 2019, pp. 193–209. doi: 10.1007/978-3-030-28954-6_10.

[27] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences.” arXiv, Apr. 11, 2017. doi: 10.48550/arXiv.1605.01713.

[28] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences.” arXiv, Oct. 12, 2019. doi: 10.48550/arXiv.1704.02685.

[29] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: removing noise by adding noise.” arXiv, Jun. 12, 2017. doi: 10.48550/arXiv.1706.03825.

[30] M. Sundararajan, A. Taly, and Q. Yan, “Gradients of Counterfactuals.” arXiv, Nov. 15, 2016. doi: 10.48550/arXiv.1611.02639.

[31] J. Fox, “The uncertain relationship between transparency and accountability,” *Development in Practice*, vol. 17, no. 4–5, pp. 663–671, Aug. 2007, doi: 10.1080/09614520701469955.

[32] J. Kroll *et al.*, “Accountable Algorithms,” *University of Pennsylvania Law Review*, vol. 165, no. 3, pp. 633-705., 2016.

[33] V. Beaudouin *et al.*, “Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach.” Rochester, NY, Mar. 23, 2020. doi: 10.2139/ssrn.3559477.

[34] D. Oberle, “How ontologies benefit enterprise applications,” *Semantic Web*, vol. 5, no. 6, pp. 473–491, Jan. 2014, doi: 10.3233/SW-130114.

[35] S. Russell and P. Norvig, *Artificial intelligence: A Modern Approach*. Pearson, 2009.

[36] A. Seeliger, M. Pfaff, and H. Krcmar, *Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review*. 2019.

[37] R. Confalonieri, T. Weyde, and T. R. Besold, “TREPAN Reloaded: A Knowledge-Driven Approach to”.

[38] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, and D. L. McGuinness, “Explanation Ontology: A Model of Explanations for User-Centered AI,” in *The Semantic Web – ISWC 2020*, Cham, 2020, pp. 228–243. doi: 10.1007/978-3-030-62466-8_15.

[39] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI: an ontology-based approach to black-box sequential data classification explanations,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 629–639. doi: 10.1145/3351095.3372855.

[40] F. Lecue and J. Chen, “Knowledge-Based Explanations for Transfer Learning,” *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, pp. 180–195, 2020, doi: 10.3233/SSW200018.

[41] F. Lecue and J. Wu, “Semantic Explanations of Predictions.” arXiv, May 27, 2018. doi: 10.48550/arXiv.1805.10587.

[42] P. Hitzler, F. Bianchi, M. Ebrahimi, and M. K. Sarker, “Neural-symbolic integration and the Semantic Web,” *Semant. web*, vol. 11, no. 1, pp. 3–11, 2020, doi: 10.3233/SW-190368.

[43] F. Bianchi, G. Rossiello, L. Costabello, M. Palmonari, and P. Minervini, “Knowledge Graph Embeddings and Explainable AI,” *Studies on the Semantic Web*, vol. 47: Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges, pp. 49–72, Apr. 2020, doi: 10.3233/SSW200011.

[44] S. Moon, P. Shah, A. Kumar, and R. Subba, “OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 845–854. doi: 10.18653/v1/P19-1081.

[45] J. Huang, W. X. Zhao, H. Dou, J.-R. Wen, and E. Y. Chang, “Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks,” in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, New York, NY, USA, Jun. 2018, pp. 505–514. doi: 10.1145/3209978.3210017.

[46] F. Lecue, “On the role of knowledge graphs in explainable AI,” *Semant. web*, vol. 11, no. 1, pp. 41–51, Jan. 2020, doi: 10.3233/SW-190374.

[47] S. Chari, D. Gruen, O. Seneviratne, and D. Mcguinness, “Foundations of Explainable Knowledge-Enabled Systems,” *Studies on the Semantic Web*, vol. 47: Knowledge Graphs for eXplainable

Artificial Intelligence: Foundations, Applications and Challenges, pp. 23–48, 2020, doi: doi:10.3233/SSW200010.

[48] “H.L.E.G.o.A.I. HLEG-AI, Ethics Guidelines for Trustworthy AI,” 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed Dec. 26, 2022).

[49] I. Esnaola-Gonzalez, “Semantic Technologies Towards Accountable Artificial Intelligence: A Poultry Chain Management Use Case,” in *Artificial Intelligence XXXVII*, Cham, 2020, pp. 215–226. doi: 10.1007/978-3-030-63799-6_17.

[50] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[51] H. Zou and T. Hastie, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[52] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[53] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System.” Jun. 10, 2016. doi: 10.1145/2939672.2939785.

[54] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Jan. 16, 2023. [Online]. Available: https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[55] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[56] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” *Adv Neural Inform Process Syst*, vol. 28, pp. 779–784, Jan. 1997.

[57] K. Sokol and P. Flach, “Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence.” arXiv, Sep. 08, 2022. doi: 10.48550/arXiv.2112.14466.

[58] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning.” arXiv, Mar. 02, 2017. doi: 10.48550/arXiv.1702.08608.

[59] S. Vollert, M. Atzmueller, and A. Theissler, “Interpretable Machine Learning: A brief survey from the predictive maintenance perspective,” in *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sep. 2021, pp. 01–08. doi: 10.1109/ETFA45728.2021.9613467.

[60] D. A. Broniatowski, “Psychological Foundations of Explainability and Interpretability in Artificial Intelligence,” *NIST*, Apr. 2021, Accessed: Dec. 26, 2022. [Online]. Available: https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence

[61] Z. C. Lipton, “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.,” *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.

[62] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert, “It’s Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI,” *ACM Trans. Comput.-Hum. Interact.*, vol. 29, no. 4, p. 35:1-35:33, Mar. 2022, doi: 10.1145/3495013.

[63] U. Ehsan *et al.*, “The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations.” arXiv, Jul. 28, 2021. doi: 10.48550/arXiv.2107.13509.

[64] A. Jacovi, J. Bastings, S. Gehrmann, Y. Goldberg, and K. Filippova, “Diagnosing AI Explanation Methods with Folk Concepts of Behavior.” arXiv, Jul. 19, 2022. doi: 10.48550/arXiv.2201.11239.

Appendix

Ethical questionnaire

Operator Question Session

1. What is the back ground of the operators who will be using our system?

    a. What level of education do they have?

2. How familiar with the advanced analytical services are the operators who will be using our system? Have they had some opportunities to use systems with artificial intelligence during their working routine? If so, could they give an example of the system they used, were they satisfied or not and why?

3. What are the examples of the artificial intelligence uses or tools that operators that will be using our system are aware of?

4. Are the operators that will be using our system familiar with explainable AI?

5. Are the operators that will be using our system interested in understanding the model that will provide recommendations to them? If so, following questions should be asked:

    a. Are the operators that will be using our system interested in understanding which parameters are considered in order to provide certain suggestion to them?

    b. Are the operators that will be using our system interested in the high level understanding of the AI model?

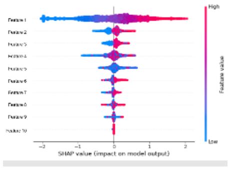
    c. Are the operators that will be using our system interested in mathematics behind the model?

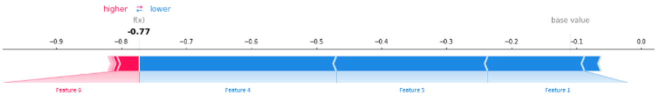
6. Figures bellow should be explained to the operates, and the following feedback on top of that should be obtained:

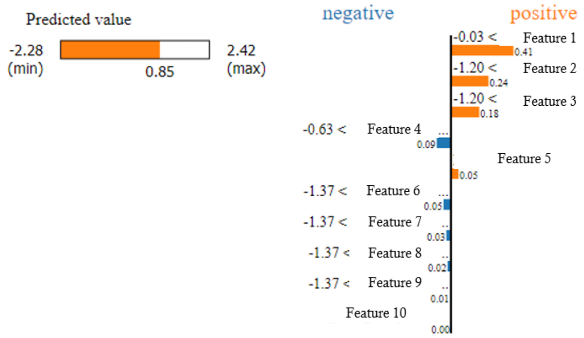
    a. From the following visualizations, what is the most comprehensive one and why?

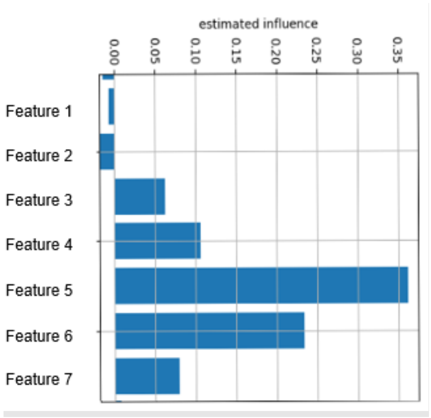
    b. Would that particular visualization be practical for your everyday work – would you be able to track the information given on it, wouldn't it be too time consuming?

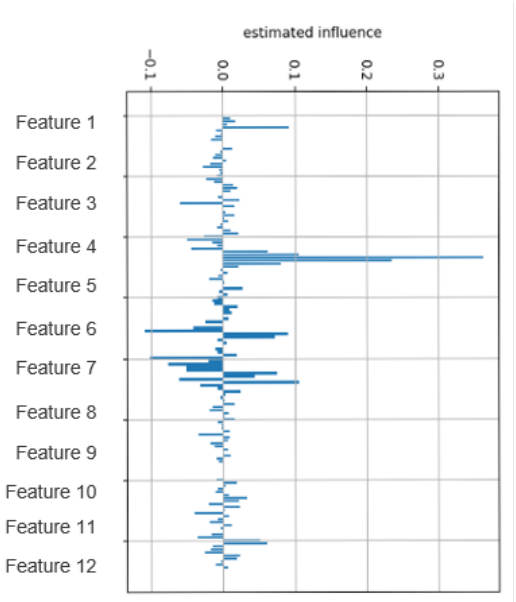
    c. Is there any other type of visualization that you find better for this purpose?











Process Managers Question Session

7. What is the back ground of the process managers who will be using our system?

    a. What level of education do they have?

8. How familiar with the advanced analytical services are the process managers who will be using our system? Have they had some opportunities to use systems with artificial intelligence during their working routine? If so, could they give an example of the system they used, were they satisfied or not and why?

9. What are the examples of the artificial intelligence uses or tools that process managers that will be using our system are aware of?

10. Are the process managers that will be using our system familiar with explainable AI?

11. Are the process managers that will be using our system interested in understanding the model that will provide recommendations to them? If so, following questions should be asked:

    a. Are the process managers that will be using our system interested in understanding which parameters are considered in order to provide certain suggestion to them?

    b. Are the process managers that will be using our system interested in the high level understanding of the AI model?

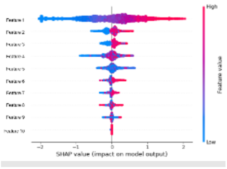
    c. Are the process managers that will be using our system interested in mathematics behind the model?

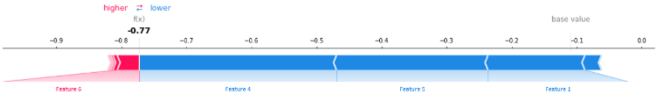
12. Figures bellow should be explained to the process managers, and the following feedback on top of that should be obtained:

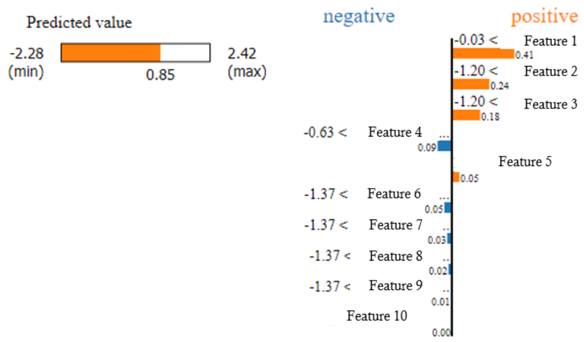
    a. From the following visualizations, what is the most comprehensive one and why?

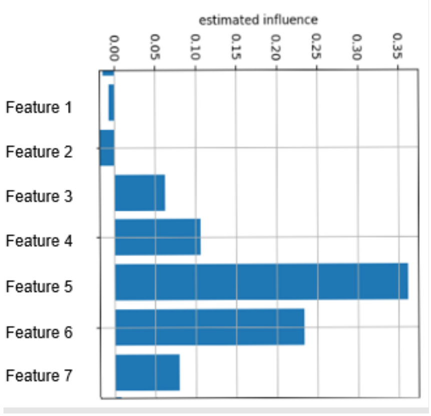
    b. Would that particular visualization be practical for your everyday work – would you be able to track the information given on it, wouldn't it be too time consuming?

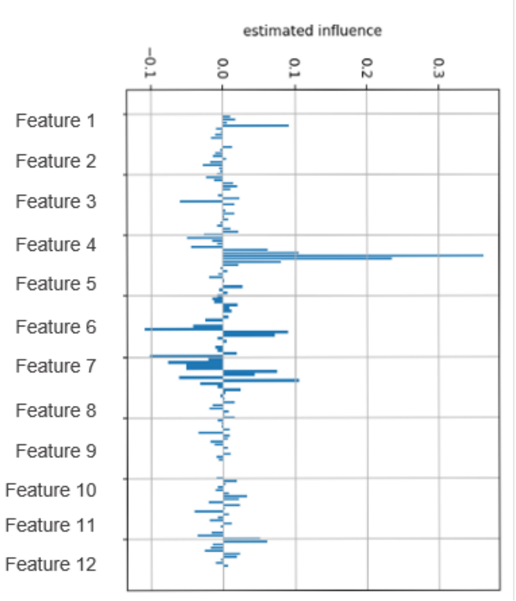
    c. Is there any other type of visualization that you find better for this purpose?











13. Are there any additional representatives from CONTI plant that will, potentially, be interested in understanding the XAI solution provided by AI-PROFICIENT project?