Article

Cyber-Physical LPG debutanizer distillation columns: machine learning-based soft sensors for product quality monitoring.

Jože M. Rožanec [1,2,3]0000-0002-3665-639X, Elena Trajkova [1,4]0000-0001-5342-1085, Jinzhi Lu* [5]0000-0001-5044-2921, Nikolaos Sarantinoudis [6]0000-0002-4263-9123, Georgios Arampatzis [6]0000-0001-8307-2891, Pavlos Eirinakis [7]0000-0002-5262-7265, Ioannis Mourtos [8]0000-0002-9243-7327, Melike K. Onat [9], Deren Ataç Yilmaz [9], Aljaž Košmerlj [1], Klemen Kenda [1,2,3]0000-0002-4918-0650, Blaž Fortuna [1,2]0000-0002-8585-9388, and Dunja Mladenić [1]0000-0003-4480-082X

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Qlector d.o.o., Rovšnikova 7, 1000 Ljubljana, Slovenia
[3] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[4] University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia
[5] EPFL SCI-STI-DK, Station 9, CH-1015 Lausanne, Switzerland
[6] Technical University of Crete, School of Production Engineering and Management, Akrotiri Campus, Chania 731 00, Greece
[7] University of Piraeus, Department of Industrial Management and Technology, Karaoli and Dimitriou 80, Pireas 185 34, Greece
[8] Athens University of Economics and Business, Department of Management Science and Technology, Patission 76, 104 34 Athens 2, Greece
[9] Tüpras, Izmit, Turkey
* Correspondence: jinzhi.lu@epfl.ch (J.L)

Featured Application: The outcomes of this work can be applied to forecast C5 content in debutanizer columns based on data obtained by a few pressure and temperature sensors. In addition, the proposed visualization can be used as a models' global explanation, highlighting opportunities regarding feature selection, the most important features guiding the forecasts, and threshold values within which the forecasting model can operate.

Abstract: Refineries execute a series of interlinked processes, where the product of one unit serves as the input to another process. Potential failures within these processes affect the quality of the end products, operational efficiency, and revenue of the entire refinery. In this context, implementation of a real-time cognitive module, referring to predictive machine learning models, enables to provide equipment state monitoring services and to generate decision-making for equipment operations. In this paper, we propose two machine learning models: 1) to forecast the amount of pentane (C5) content in the final product mixture; 2) to identify if C5 content exceeds the specification thresholds for the final product quality. We validate our approach by using a use case from a real-world refinery. In addition, we develop a visualization to assess which features are considered most important during feature selection, and later by the machine learning models. Finally, we provide insights on the sensor values in the dataset, which help to identify the operational conditions for using such machine learning models.

Keywords: Artificial Intelligence; Machine Learning; Explainable Artificial Intelligence; Soft Sensors; Industry 4.0; Smart Manufacturing; Cyber-Physical System; Crude Oil Distillation; Debutanization; LPG Purification

1. Introduction

Petroleum refineries receive crude oil of different provenances with their specific characteristics. The inlet crude oil feedstock is transformed into final products through multiple processes. Each process provides products whose qualities are prescribed by different standards, such as the Liquefied Petroleum Gas (LPG), a mixture of hydrocarbon gases used in heating appliances and vehicles. The final mixture product usually contains 48% propane, 50% butane, and up to 2% pentane (hereafter, also referred to as C5) which is developed based on local regulations and composition requirements, the intended use, and even seasonal limitations (e.g., a higher proportion of propane is used in winter due to its evaporation point).

In order to achieve the desired quality, the LPG obtained from crude oil distillation must undergo several processes to remove impurities. One of these purification processes is debutanization, which removes C5. To ensure the final mixture meets the specification standards, samples are taken in various stages of the refinement and purification process and undergo lab analysis. Results are passed on to production engineers so that they can adjust process settings if required. However, lab analysis may take up to several hours to be completed and is not conducted every day, causing the identification of an already existing off-specs situation to be delayed. This, in turn, makes recovery harder, since the sooner an off-spec situation is identified and resolved, the better it is for the recovery efforts in terms of both, time and cost. Hence, there is a need for early (or ideally real-time) identification of situations where C5 content exceeds specifications. This brings a strong motivation to create a model-based approach for a cognition module supporting the operation which alerts of an C5 off-spec situation, and enables real-time decision-makings based on such alerts.

Currently, several approaches to estimate debutanization process outcomes exist. Among them, Aspen HYSYS [1] uses mathematical models to simulate a debutanizer unit and predict process outputs. Such simulation models frequently use elaborate math and complicated equations to achieve enough generalization to be applied across different units. Data-driven models overcome limitations regarding equation solving complexity by utilizing past data to learn and produce possible solutions. While the ability to reuse them across units strongly depends on the model design, once trained, such models can provide forecasts with almost no latency. If forecasts are good enough, the models can get frequently insights regarding C5 content in the LPG, for providing ground for earlier off-spec product identification and timely decision making.

Real-time prediction of C5 content during the debutanization processes provide new insights that guide decision-makings for process monitoring and control. To create machine learning models capable of such forecasts, we utilize historical sensor data regarding operational temperature and pressure, as well as laboratory results obtained from the samples analysis. Such data and analysis results enable to support machine learning model training and evaluation by identifying correlations between sensed conditions and measured outcomes for two purposes: (i) with real-time sensor data, such models can provide real-time C5 content estimates; (ii) with new real-time sensor data and lab analysis data update, the machine learning model performance is expected to be promoted in time, if retrained with the new data available.

In this paper, we develop machine learning models for a real-world use case, based on sensor data provided by a Tüpras[2] refinery. By examining the actual process in the use case, we found that different debutanizer columns have different features because of their different designs. Moreover, only a few sensors are located in the debutanizer column. Most sensor data corresponded to the pipping system that connected the debutanizer column with the condensation unit and the units that follow. We used several debutanizer unit diagrams to understand where the sensors are located and which sensors are close to the distillation column exit. Temperature and pressure conditions are identified by the ones near the column exit, and hence the first ones placed in the pipes close to the related exit but before the condensation unit. We assume such data provides good insight on how operating conditions relate to extracted samples and

[1] https://www.aspentech.com/en/products/engineering/aspen-hysys

[2] https://www.tupras.com.tr

measured composition. Furthermore, we observed that there are some cases where both, temperature and pressure sensors, exist for any given point in the debutanizer column, but at least one of them exists. Considering these limitations, machine learning models are developed to predict C5 content based on the inputs of two sensors (one pressure sensor and one temperature sensor). Finally, we develop two machine learning models that provide predictions based on the data from these two sensors for independent estimate: (i) one that predicts the expected amount of C5 in the LPG; and (ii) one that forecasts whether C5 content is off-spec (higher than 2%).

The contribution of this paper is the utilization of operational temperature and pressure sensor data to develop:

1. a machine learning model to predict C5 content in LPG stream;
2. a machine learning model to predict if C5 content exceeds specification levels

Machine learning models built utilizing data from a few sensors can be more easily applied to a broad range of debutanizer columns since they impose fewer restrictions on the number of input data sources required to provide forecasts. Thus, we consider that a major strength of our approach is the fact that it relies only on data of two sensors, one measuring pressure and the second one measuring temperature in the debutanizer column - both placed at separate locations within the column.

Along with the development of the aforementioned models, we also provide a prototype dashboard, which provides global explanations to understand which features were considered most relevant during the feature selection, and which features were considered relevant by the forecasting model. In addition, we provide insights on the sensor values' distribution in the training set, to understand the models' operational limitations.

To evaluate our models, we have utilized three metrics: two for measuring regression features and one for measuring the classification features. We assess the regression models' performance with the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The MAE is not sensitive to outliers and can thus provide a reasonable estimate of the models' performance for normal C5 levels. The RMSE penalizes large errors and thus better indicates if out-of-spec measurements were predicted adequately. The classification models' performance is measured with the Area Under the Receiver Operating Characteristic Curve (AUC ROC [1]). AUC ROC is invariant to *a priori* class probabilities, referring to a relevant property when measuring models' discrimination power in an imbalanced dataset. After evaluating the models, results show that our approach is applied to effectively provide real-time C5 content predictions in the LPG debutanization process of our given use case.

The rest of this paper is structured as follows: Section 2 presents related work. Section 3 describes a Tüpras refinery use-case,and Section 4 introduces the features created for the C5 content forecasting model, as well as the way to develop and evaluate these models. Section 5 presents the experiments we performed and the obtained results. Finally, Section 6 offers our conclusions and provides an outline for future work.

2. Related work

2.1. Distillation process-related models

Debutanizer columns are an important part of several processing units in oil refineries. Therefore, the objective of the online composition of debutanizer outlet streams is to maximize the production of LPG while meeting the corresponding quality standards. Currently, the quality of the debutanizer output is measured via laboratory analysis. Hence, changes in the quality are identified only upon the analysis of the sample, which may take several hours. Therefore, in order to maintain the quality of the product within the predefined specifications, it is of imperative importance to predict the top and bottom outputs of the debutanization process precisely [2].

To realize this objective, [3] identifies three major approaches to develop the required models: (i) first-principle (a.k.a. fundamental) models, which consider mass, energy, and momentum principles and equations to provide a forecast; (ii) machine learning models, which are created by training an algorithm on input-output data of the process; and (iii) hybrid models, which combine both the fundamental and the empirical models.

First-principle models involve sets of non-linear differential equations (usually in the order of 10^2 or 10^3 non-linear differential equations) and a comparable number of algebraic equations [4,5]. The equations usually take into account the global balance of matter, partial balances of matter, pressure, temperature, flow, reflux policies, and the relationship between component concentrations at different levels of the distillation column [6,7]. While additional information regarding the structure of the distillation column can further enhance such models (e.g., the number of trays in a column or the column hydraulics [8]) with the increasing computational complexity of such models.

To alleviate the computational needs, simplified distillation column models have been proposed [9,10], at the expense of an increased error whose applicability often restricted to a single column [11]. These models are usually implemented in Advanced Process Control systems (APC), such as a Multivariable Model Predictive Control (MPC), for managing relevant process variables and their dynamics. The equations mentioned above govern the control logic between variables. Algorithms that perform matrix computations are used to solve such system dynamic models with multiple variables simultaneously. In addition to their computational complexity, the usefulness of such models is constrained to the model assumptions, e.g., sensor colocation points[12].

Data-driven models provide an alternative modeling approach for developing the forecast models [13]. In particular, machine learning models are developed based on the prior knowledge of the physical processes for creating good features of model outputs. The models are trained with the collected data from the actual operations of the unit: 1) The raw data is transformed into a dataset for developing models which perform features that reflect different dynamic features for the raw data variables; 2) Through the developed models, observed outputs are generated through the the feature vectors. Through such model features, the machine learning models can accurately learn non-linear features from the data, even when some noises exist in the data [14].

Hybrid models arise from the combination of the first-principle and data-driven models [15]. Such models are used to retain the theoretical knowledge of the process, which is mirrored in equations. In contrast, the data-driven models can augment such knowledge using data, and can be used to model parts of the process that are hard to formulate and would otherwise require overly complex first principle models [3,16]. Hybrid models have been implemented widely in various chemical processes such as batch distillation [17], reactive distillation [18], and polymerization process [19,20]. However, only a handful of models have been implemented in continuous distillation columns.

In the literature, there are some attempts to model continuous distillation processes in refineries. Such attempts not only include debutanizer columns [12,21], but also various other units such as Crude Distillation Units (CDU) [22,23] and Fluid Catalytic Crackers (FCC) [24,25]. Among the models developed for debutanizer columns, we find the artificial neural networks (ANN) [23,26,27], partial least square regression [28,29], support vector regression (SVR) [23,28], principal component regression [30], supervised latent factor analysis [31,32], probabilistic regression [33], and state-dependent autoregressive model with exogenous variables [34].

To evaluate C5 and C4 product concentrations in the debutanizer column, [3] created a dynamic neural model that acts as a soft sensor based on the data provided. In a similar manner, [35] developed an ANN model to predict LPG composition at the top and bottom of a distillation column, comparing its performance to a partial least squares model. A comparison between different models was also performed by [27], which developed multiple linear regression, principal components regression, and neural networks models for a debutanizer column. They concluded that the performance of such models was superior to least square regression models and support vector regression models reported in the literature. Finally, [36] aimed to identify the governing equations regarding a distillation column using a white-box machine learning approach.

Cyber-physical systems describe systems that integrate the physical processes into the digital world, where monitoring and analytics can be performed [37,38]. A standard abstraction model considers three significant layers: physical, cybernetic, and an interface between both [39]. The concept of cyber-physical systems has been successfully implemented in petrochemical plants [40].

This paper highlights the importance of artificial intelligence applications compared to traditional analytic methods based on mathematical models. It proposes a cyber-physical integration using machine-learning models to provide real-time LPG C5 content estimates based on streamed sensor data. In our use case, sensor data regarding pressure and temperature was available only from a few sensors at the top of the debutanizer column; hence, such models could not be replicated. Nevertheless, we have acknowledged the algorithms described in the related work and implemented models based on them and our set of features.

2.2. Explainable Artificial Intelligence

The machine learning models are growing in complexity and sophistication providing accurate forecasts based on historic data. At the same time, there is an increasing need to understand the logic behind such models, to comply with regulatory requirements, and provide ground for responsible decision-making [41,42]. Insights on the process followed by such models when applying operations on the input to provide a forecast enable to decide whether such forecasts can be trusted or not [43,44]. To respond to such challenges, research on techniques, approaches and visualizations is done in a sub-field of artificial intelligence, known as Explainable Artificial Intelligence (XAI).

Multiple taxonomies were proposed to categorize the different XAI approaches. Arrieta et al. distinguish between transparent models and post-hoc explainability techniques, dividing the last category into model-agnostic and model-specific approaches [45]. Transparent models are also known as inherently-interpretable or *white-box models*, while the models that do not fall into this category, are considered opaque or *black-box models* [46]. A more elaborate taxonomy was proposed by Das et al. [47], who considered dividing XAI techniques based on three criteria: scope (considering global or local explanations), methodology (if the technique focuses on the input data or model parameters), and usage (if is model-agnostic or model-specific). Regarding the scope, local explanations provide insights regarding a particular forecast, while global explanations attempt to describe the overall model's behavior [48].

When providing global explanations for models trained on tabular data, a frequent model specific approach is to consider the features' weight in the model to determine the features' relevance ranking. Model agnostic alternatives have been devised by several authors using surrogate models [49–51]. While much research has been done on explaining models' behavior, less research was invested towards crafting comprehensive explanations with insights regarding the data and the model creation process. Part of this void was addressed by MELODY (MachinE Learning MODel SummarY) [52], and SUBPLEX [53], which connect local explanations to data analytics either summarizing insights regarding the whole dataset or a relevant subpopulation. INFUSE [54], on the other side, focused on providing explanations regarding the feature selection process, and the influence of different feature selection strategies on it. While INFUSE takes into account a process of cross-validation, it does not bind it to the resulting model and any model related explanations.

Visual interpretations are considered particularly effective to explain the models' forecasting rationale [55]. While much work was invested towards developing XAI techniques, some researchers consider not enough research was invested on making such explanations end-user-centered [56,57]. Visual explanations comprehend insights regarding the dataset and feature contributions at a local and global level. Scatterplots are frequently used to visualize data distribution, using some dimensionality reduction technique to map the high-dimensional dataset into two dimensions [58,59]. Color-coded instances are frequently used in classification tasks, and interactive interfaces provided, to enable the user focus on specific instances and conduct further research [48]. To represent features' contributions, horizontal bar plots [53,60,61], breakdown plots [62,63], heatmaps [64,65], Partial Dependence Plots [66], or Accumulated Local Effects Plots [67] are used.

In this research, we complemented our model development with a dashboard, that provides insights into the most informative features within the dataset, when considering feature selection, while also informing their relevance from the models' point of view. In addition, we inform the value ranges of each sensors' readings found in the dataset. Such values must be taken into account, since the model is able to issue good predictions within the observed ranges, and not outside them.

3. Problem Statement

3.1. Tüpras refinery

The use case corresponds to a Tüpras refinery located in Izmit, which began oil production in 1961 and currently has a design capacity to process 11.3 million tons of crude oil per year. The crude oil is supplied from four countries: Iran, Iraq, Russia, and Saudi Arabia. Each of these crude oil feedstocks have different characteristics such as density, sulfur content, and impurities. The refinery complies with Euro 5 standards [68] and produces mostly diesel, gasoline, and LPG. The entire refining unit consists of atmospheric and vacuum distillation units, hydrocrackers, fluid catalytic crackers, continuous catalyst regeneration reformers, diesel and kerosene desulphurization units, merox, asphalt units, and sulfur recovery units. In Fig. 1, we provide a diagram showing the relation between processing units of this refinery, highlighting the LPG and gas flows. In this research, we focus on the LPG debutanizer units, which is implemented for the atmospheric distillation process. While feedstock changes regularly, experts pointed out that they do not display much difference in light hydrocarbons content regardless of the crude oil provenance. The atmospheric distillation process ameliorates this difference before the LPG enters the debutanizer unit. Given that the concentration of pentanes does not depend on the crude oil provenance, we consider it as a specific function of the debutanization process which is a distillation process with two control variables: pressure and temperature.

3.2. Debutanization process

The debutanization process is a fractional distillation process that aims to recover the light gases ($C1 - C4$) and the Liquefied Petroleum Gas (LPG) from the overhead distillate coming from the distillation unit [27]. This distillation process aims to separate liquid components by heating a liquid to vapor, condensing the vapor back to liquid to purify or separate it. To that end, three components are required: (1) a distillation column (used to separate a liquid mixture into its fractions based on the differences in volatilities); (2) a reboiler (used to provide the necessary vaporization of the distillation process); (3) a condenser (used to cool and condense the overhead vapor).

The $CH4$ (methane) component exists in the feedstock with the other alkane compounds, with the $C4H10$ (butane) fraction only gaining its freedom when the vapourised gas condenses inside the array of valve trays that line the interior of the debutanizer column. Based on thermal unit conversion technology, the butane is efficiently siphoned from the raw feed. To achieve this, the boiling point of butane is used as a reference point to determine temperature and pressure conditions. Pure butane condenses in the debutanizer column when the architecture of the column locks in the mandated variables, so few impurities can form. Similarly, propane, ethane, and methane are liberated and refined as valuable fuel sources in the other alkane processing columns.

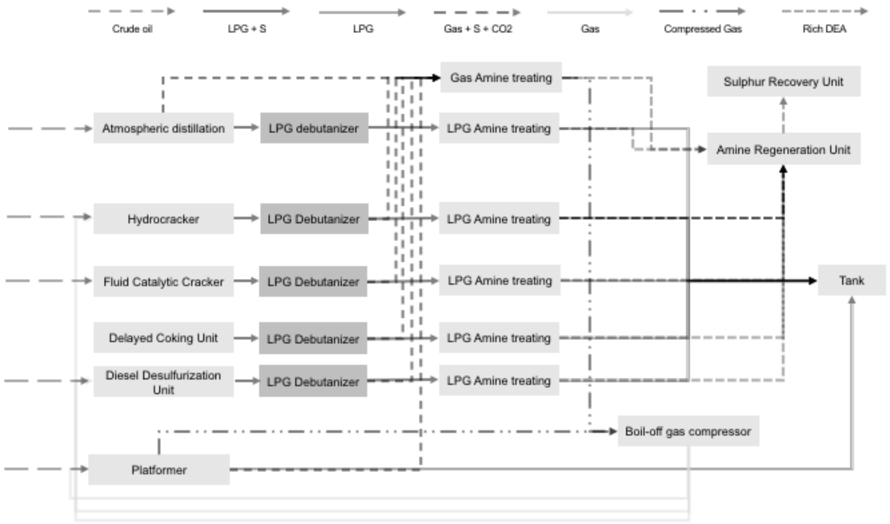

Figure 1. High-level schematic diagram of some of units found in crude oil refineries. In this research we focus on one of the LPG debutanizer units.

The distillation process can be manipulated with three control variables: the feed flow rate, reflux flow rate, and reboiler flow rate. Feed flow rate controls the feed of the distillation column, reflux flow rate controls the overhead temperature, and reboiler flow rate controls the bottom temperature. However, modeling such a column is a complex process as it involves various non-linearities and includes multiple variables with interactions between them [69].

In the Tüpras refinery, there is an abundance of sensors to monitor the entire debutanization processes. Data from these sensors include measurements of input and output flows, temperatures, and pressures across the whole refinery. These are used in feedback loops to maintain the process stable and control the system dynamics close to the set-point values that the process engineers have selected for seamless plant operation. While rich sensor data exists, we only obtained the data from the temperature and the pressure sensors on the top of the debutanizer columns. Although a limited number of sensors was provided, our proposed approach presents excellent results as shown in Section 5.

3.3. *Relevant physical and chemical principles and laws*

In our use case, we have sensor data for the temperature and pressure measurement. To formalize meaningful features enabling the models to predict C5 content, we have considered the following laws and equations from physics:

- **Raoult's law** states that the total pressure of a component equals the vapor pressure of its pure components multiplied by their mole fraction (see Eq. 1);
- **Antoine's equations** provide a relationship between the vapor pressure of a pure component and three empirically measured constants at a given temperature (see Eq. 2);
- **Combined Gas Law** states that the ratio of the product of pressure and volume and the absolute temperature of a gas equal a constant (see Eq. 3);
- **Clausius-Clapeyron relation** describes pressure at a given temperature T_2 if the enthalpy of vaporization and vapor pressure are known at some other temperature T_1 (see Eq. 4)

For the case study, we obtained data from sensors P1 and T2 of the debutanizer unit (in the Fig. 1); while we missed sensor readings from T1 and P2.

Not having both temperature and pressure at a given point of the debutanizer column prevents us from using the Ideal Gas Law equation to compute the gas molar

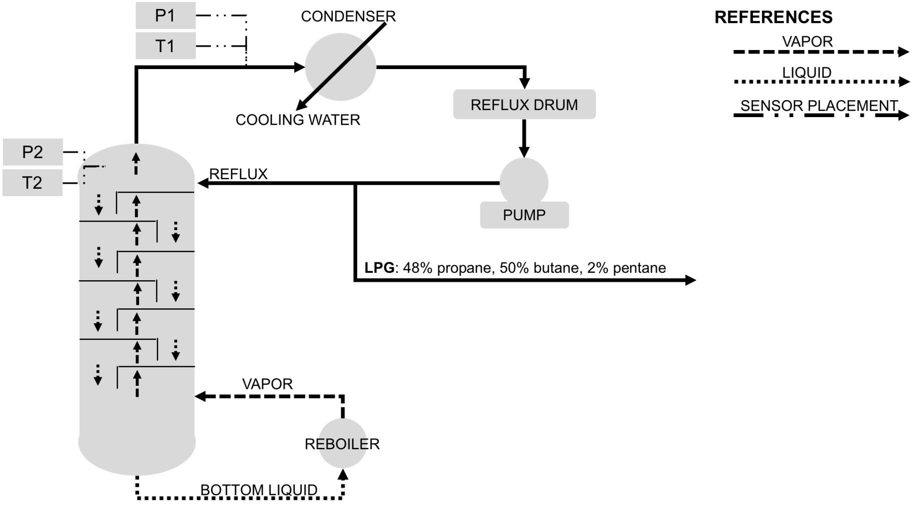

Figure 2. Schematic diagram of an LPG debutanizer column. In the diagram we reference two locations on which the sensors are placed. In this research, we developed models that take into account only sensors P1 and T2.

$$P = P_1 \cdot x_1 + ... + P_n \cdot x_n \tag{1}$$

Equation 1: Raoult's law. P refers to pressure, x refers to mole fraction, and the n indicates different mixture components.

$$log_{10}P = \frac{A\check{}B}{C+T} \tag{2}$$

Equation 2: Antoine's equation. P refers to pressure, T refers to temperature. A, B, C are empirical, component specific constants.

weight (see Eq. 5) of the mixture. The gas molar weight could provide further insights on the mixture composition using the gas molar weight equation (see Eq. 6). The gas molar weight equation expresses that the gas molar weight equals to the sum of the molar weights of the pure components multiplied by their mole fraction.

LPG specifications require LPG to have a mixture of propane and butane, with no more than 2% of the volume of five carbon components (C5) and no more than 5% of the volume of two and five carbon components (C2 and C5). Though many possible components have two and five carbons, we decided to approximate them as a single pure component. Considering the laws, equations, and restrictions described above, we can derive a set of equations, which provide meaningful cues on the expected mixture composition, and thus drive better forecasts. E.g., from the Eq. 1 and approximating the LPG composition to the four elements described above, we obtain that C5 proportion can be expressed as Eq. 7. While pressure is known from the sensor readings (P1), we do not know the exact proportion of propane and butane. We also miss sensor data regarding the temperature at the same point where the pressure is sensed (T1). Considering that the relationship between temperature and pressure is linear, and given a snapshot of sensor data, we approximate T1 based on P1 and T2. Such an approximation allows us to compute saturation pressures for pure LPG components based on Antoine's equations (see Eq. 2). Considering various scenarios of possible LPG composition, we compute features (see Section 4.3) signaling expected pressure for given conditions and how it compares to the pressure sensed in the debutanizer unit. When considering the constants for Antoine's equations, we approximated two carbon hydrocarbon elements with methyl-disulfide ($C_2H_6S_2$), and five carbon hydrocarbon elements with pentane

$$k = \frac{P \cdot V}{T} \tag{3}$$

Equation 3: Combined Gas Law equation. P refers to pressure, V refers to volume, and T refers to temperature. k is a constant.

$$ln\left(\frac{P_1}{P_2}\right) = -\frac{L}{R} \cdot \left(\frac{1}{T_2} - \frac{1}{T_1}\right) \tag{4}$$

Equation 4: Clausius-Clapeyron relation. P refers to pressure, T refers to temperature, L is the specific latent heat of the substance, and R is the specific gas constant.

(C_5H_{12})). When doing so, we considered the vaporization temperature and sulfur content (sulfur is removed in later stages).

4. Methodology

4.1. Data preparation

In order to realize the proposed machine learning models, we used data provided by Tüpras. The data included temperature and pressure sensor data, and 263 laboratory measurements (167 measurements from the debutanizer *Unit A*, and 96 from the debutanizer *Unit B*), all sampled simultaneously at irregular day intervals. We consider that the irregular sampled data should not affect the machine learning model training since temperature and pressure sensor inputs are used to estimate LPG C5 content. As described in Section 3, experts informed us that light components, such as C5, do not vary much between feedstocks. The debutanizer follows a previous distillation phase, where LPG is separated from the rest of crude oil derivatives. Therefore, debutanizer's operational pressure and temperature are used to influence the observed LPG C5 concentration.

When creating the dataset for training machine learning models, we sampled sensor data at an each minute, for computing the average of raw sensor measurements between two minutes. We chose to impute values using forward filling for missing values, considering that missing sensor readings are most likely to have a value similar to the last one observed. Since we had no information regarding set-point configurations on past operations, we ran a change level detection algorithm on sensor reading time series. The algorithm identified changes in sensor data which refers to that values above a certain threshold. We empirically tried different threshold values and obtained the best results, corroborated with plots manual inspection, by setting it to 4%. In Fig. 3A, we provide an example of three time series of sensor data, enclosing some level changes within dashed squares. In Fig. 3B, we show two plots that illustrate how the change level detector works.

Experts instructed us that the timestamps from laboratory samples did not match sensor data timestamps. In order to match them, timestamps from laboratory samples had to be transposed four to five hours earlier. Since accurate data regarding time transposition was missing, we decided to consider sensor values measured in fifteen minutes slots for a time range of an hour and a half (see Fig. 4). Since operational conditions change when a new set-point is given, we computed the median sensor value since the last change level detected and the upper bound time considered to match the laboratory reading.

4.2. Data analysis

To perform our research, we focused on the data provided from the two LPG debutanizer units. We got a total of 263 laboratory analysis results: 167 for *Unit A* and 96 for *Unit B*. Sensor reading values were attached to them through the procedure described in the previous subsection. We observed that only *Unit A* had pentane concentrations

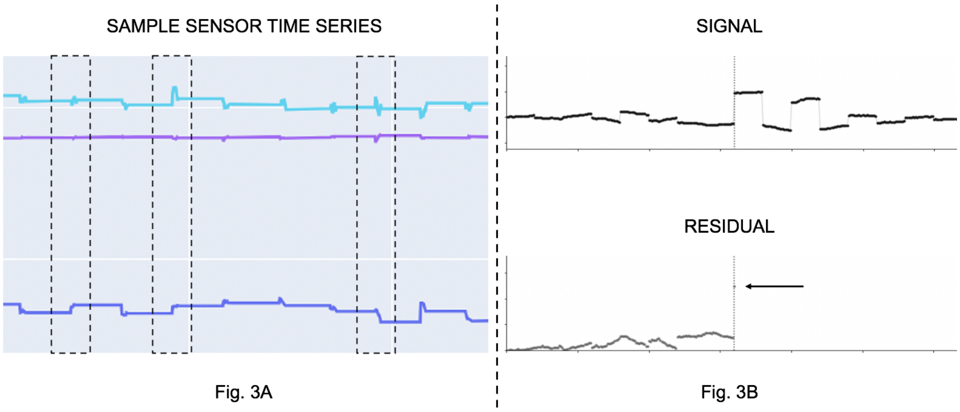

Figure 3. Fig. 3A shows a plot with three sample sensors timeseries. The dashed squares enclose some of the change levels observed in those time series. Fig. 3B shows two plots, related to the change level detector: on the top we observe the signal, and on the bottom the residual. If the residual exceeds certain threshold, a new interval is created.

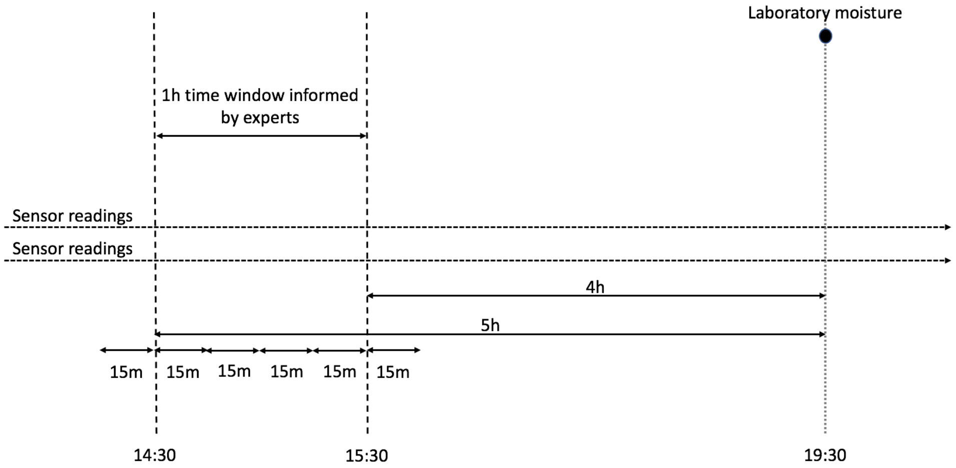

Figure 4. Timestamp conciliation between sensor and laboratory sample timestamps, based on insights provided by experts. Since a time range is provided, we decided to sample sensor values in the given interval every fifteen minutes, adding a fifteen minutes tolerance at the interval edges. Times provided in this example do not correspond to real timestamps in data.

$$P \cdot V = n \cdot R \cdot T \tag{5}$$

Equation 5: Ideal Gas Law. P stands for pressure, V stands for volume, n represents the amount of substance, R is the ideal gas constant, and T corresponds to the temperature.

$$M = M_1 \cdot x_1 + ... + M_n \cdot x_n \tag{6}$$

Equation 6: Molar weight equation. M stands for molar weight, x represents mole fractions, while the subindexes indicate different mixture components.

that exceeded the allowed out-of-specification threshold, reaching a total of fourteen off-specification events. We provide the summarized statistics of the sensor readings and target values in Table 1.

4.3. Feature creation

When creating features for our models, Raoult's law and the Gas molar weight equation in Section 3.3 assume that all the components and proportions of a given gas are known to compute the final pressure and molar weight. While specifications indicate

doi:10.20944/preprints202110.0364.v1

	Unit A							Unit B						
	mean	stdev	min	25%	50%	75%	max	mean	stdev	min	25%	50%	75%	max
P1 (kg/cm^2)	7,42	0,29	6,54	7,28	7,43	7,60	8,32	4,98	3,84	0,00	0,00	7,61	8,02	8,78
T2 (°C)	62,66	19,82	0,00	66,17	67,41	69,13	89,70	35,99	30,84	0,00	0,00	58,20	61,38	80,36
C5	0,63	1,15	0,00	0,02	0,17	0,70	6,52	0,04	0,11	0,00	0,00	0,00	0,03	0,74

Table 1. Description statistics for sensor and laboratory analysis data obtained for debutanizer *Unit A* and *Unit B*.

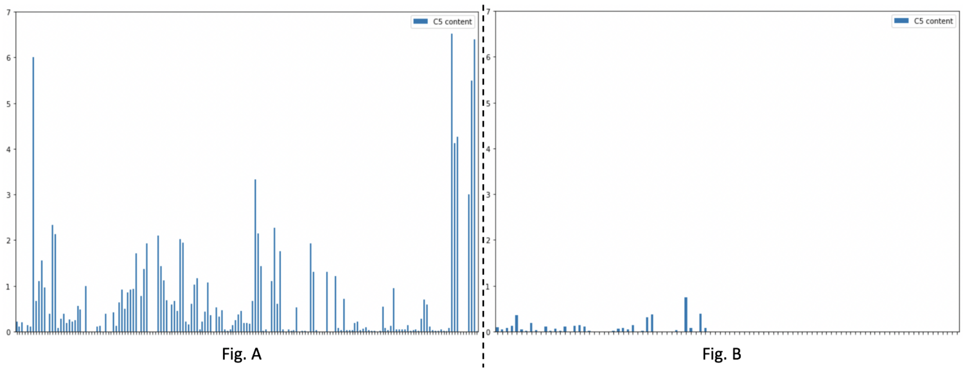

Figure 5. Measured C5 content in laboratory samples over time. Please notice, that the samples are taken at irregular intervals. On Fig 5A we present measurements from *Unit A*, while on Fig 5B we present measurements from *Unit B*.

$$x_{C5} = \frac{P - (P_B \cdot x_B + P_P \cdot x_P + P_{C2} \cdot (1 - x_B - x_P))}{P_{C5} - P_{C2}} \tag{7}$$

Equation 7: Estimated C5 content. We obtain P from sensor data, $\mathrm{P_i}$ can be computed based on a given temperature, $\mathrm{x_B}$ and $\mathrm{x_P}$ can be approximated to LPG specification, or other useful values.

that no more than 2% of the LPG volume is compound by C5 hydrocarbons and that the sum of C2+C5 hydrocarbons must not exceed 5% of the LPG volume, a wide range of possible mixture proportions is observed in reality. In some scenarios, the C5 proportion exceeds the specifications, which is detrimental to propane and butane content. The same is observed for C2 content. In our model, we decided to consider five hypothetical LPG compositions as described in Table 2. Our hypothesis is that such simplifications could be useful towards understanding the real LPG composition given temperature and pressure sensor readings. To compute specific pressures given Antoine's equations, and given the wide variety of C2 and C5 components, we approximated them with a single type of chemical compound: methyl-disulfide ($C_2H_6S_2$), and pentane (C_5H_{12}). The constants for Antoine's equations were obtained from the National Institute of Standards and Technology[3], and the University of Maryland[4], which cite the following scientific literature sources: [70–77].

For each of these scenarios, we estimated the T1 values using the Clausius-Clapeyron relation based on the enthalpy of vaporization we computed for a snapshot of data provided in debutanizer unit diagrams (see Fig. 2). By analyzing temperature and pressure for three segments of measurements, we identified that high or low C5 content is likely associated to certain pressure thresholds. We thus created dummy variables considering those thresholds.

In Table 3 we describe some of the features we developed for our machine learning models. We grouped then in *Feature Groups*, based on their common characteristics. While features from *Features Group 1* correspond to raw sensor readings, the rest of the features was developed based on physical principles and equations presented in Section

[3] https://webbook.nist.gov/

[4] https://user.eng.umd.edu/ nsw/chbe250/antoine.dat

LPG sample mixture	C2H6S2	C3H8	C4H10	C5H12
1	0.000	0.485	0.505	0.010
2	0.000	0.480	0.500	0.020
3	0.030	0.465	0.485	0.020
4	0.000	0.465	0.485	0.050
5	0.000	0.455	0.475	0.070

Table 2. Description of sample mixtures considered to compute expected pressure for certain temperature, given the mixture composition and constants from Antoine's equations.

Features Group (FG)	FG ID	Feature	Description	Type
Sensor reading values	1	P1	Pressure measurement from sensor P1	Real number
		T2	Temperature measurement from sensor T2	Real number
Expected mixture vapor saturation pressure for temperature T2	2	spt002	Mixture #1	Real number
		spt0	Mixture #2	Real number
		spt1	Mixture #3	Real number
		spt2	Mixture #4	Real number
		spt3	Mixture #5	Real number
		spt4	Mixture #6	Real number
Pressure P1 in range	3	p<7.06	Pressure below 7.06 kg/cm2	Boolean
		p<7.14	Pressure below 7.14 kg/cm2	Boolean
		p>7.63	Pressure above 7.63 kg/cm2	Boolean
Expected T1 temperature for mixture	4	T1-spt1	Mixture #3	Real number
		T1-spt2	Mixture #4	Real number
		T1-spt3	Mixture #5	Real number
		T1-spt4	Mixture #6	Real number
Relative pressure, comparing pressure P1 and expected mixture pressure for temperature T2.	5	spr002	spt002/P1	Real number
		spr0	spt1/P1	Real number
		spr1	spt2/P1	Real number
		spr2	spt3/P1	Real number
		spr3	spt4/P1	Real number
		spr4	spt5/P1	Real number
Ratio between estimated T1 temperature for mixture, and the P1 pressure.	6	T1/P1-spt1-T2	Mixture #3	Real number
		T1/P1-spt2-T2	Mixture #4	Real number
		T1/P1-spt3-T2	Mixture #5	Real number
		T1/P1-spt4-T2	Mixture #6	Real number
Categorical feature indicating whether the relationship between estimated T1 temperature and P1 pressure is above or below the value measured from normal operating conditions, from values obtained in diagrams provided.	7	T1/P1-spt1.vref	Mixture #3	Boolean
		T1/P1-spt2.vref	Mixture #4	Boolean
		T1/P1-spt3.vref	Mixture #5	Boolean
		T1/P1-spt4.vref	Mixture #6	Boolean

Table 3. Some of the features we created for the machine learning models. *spr* abbreviates *saturation pressure ratio*, while *spt* abbreviates *saturation pressure total*.

3.3. Features corresponding to *Features Group 2* indicate the expected vapor pressure at P2 for the sensed temperature at T2, considering the mixtures from Table 2. *Features Group 3* groups three categorical features defined in relation to P1, where thresholds were defined based on average P1 pressure values and standard deviations of each group and their relation to measured LPG C5 content. The features in *Features Group 4* are analogous to the features from the *Features Group 2*, computing the expected T1 temperature based on pressure P1, for LPG mixtures specified in Table 2. These features are used to compute the *Features Group 5* when contrasted with sensed pressure at P1. The *Features Group 6* computes the ratio between the estimated T1 temperature, and the pressure at P1. Finally, *Features Group 7* indicates whether the ratio between the estimated temperature T1 and pressure P1 is greater than the value measured from the diagrams obtained under normal operating conditions.

We created a total of 198 features. To avoid overfitting the machine learning models, we selected only K features, obtaining K from $\sqrt{N}$, where N is the number of instances in the training subset, as suggested by [78]. Feature selection was performed by computing their mutual information [79], and selecting the *top K* most informative ones. We describe the correlation between the selected features and target C5 content values we aim to forecast in Fig. 6, 7, and 8.

4.4. Machine learning model Development

4.4.1. Regression Machine learning models

Forecasting LPG C5 content solely from temperature and pressure data is a challenging task. In Fig. 5, C5 content could reach very disparate values: from values close to zero to a range of valid values when considering the LPG specifications, and some peaks corresponding to samples with C5 concentrations well above the specification

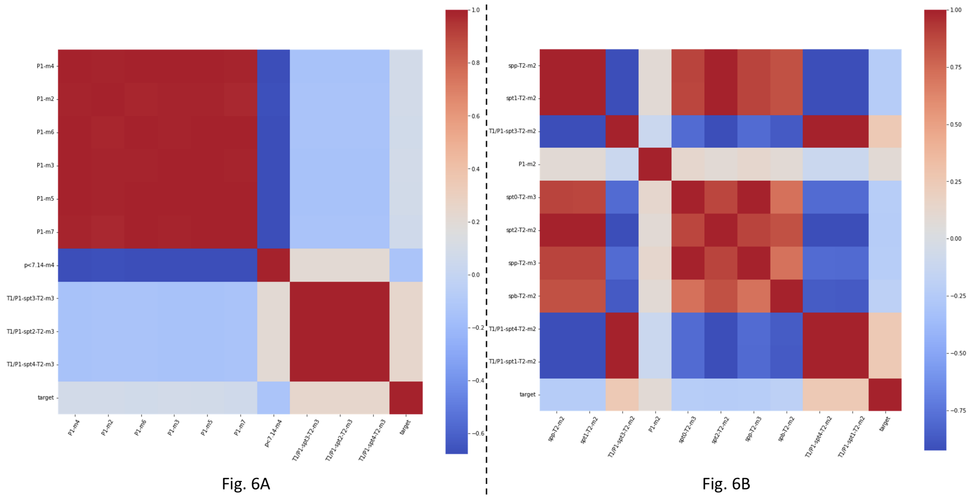

Figure 6. Feature correlation for ten selected features in each case, when forecasting the amount of C5 present in distilled LPG at the end of the distillation process in the debutanizer columns, for Unit A. On Fig. 6A we present feature correlations for Experiment 1, while on Fig. 6B we present feature correlations for Experiment 2.

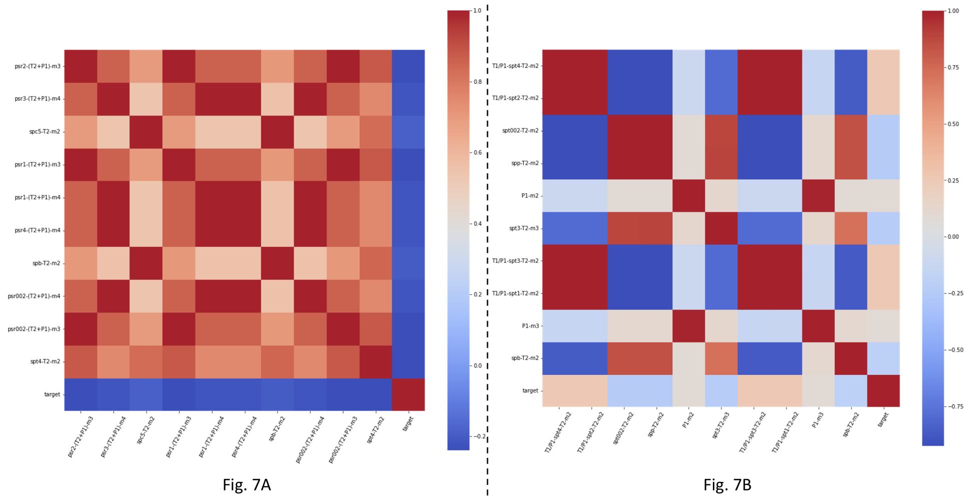

Figure 7. Feature correlation for ten selected features in each case, when forecasting the amount of C5 present in distilled LPG at the end of the distillation process in the debutanizer columns, for Unit B. On Fig. 7A we present feature correlations for Experiment 1, while on Fig. 7B we present feature correlations for Experiment 2.

ranges. In our research, we developed and compared six models. These models include two baseline models and four models that aim to provide enhanced forecasts, and which we describe below:

- **Baseline 1 (C5 median)**: our prediction is the median of C5 values observed in the data set for model training;
- **Baseline 2 (LiR)**: linear regression to predict C5 content based on raw temperature and pressure sensor measurements (P1 and T2 from Fig. 1, described in Feature Group ID #1 at Table 3);
- **Model 1 (LiR)**: linear regression considering raw sensor measurements of P1 and T2 sensors at fifteen minute intervals (see Fig. 1, and all features described in Table 3), for the time range as presented in Fig. 4;
- **Model 2 (SVR)**: Support Vector Regressor [80], which takes into account most relevant features assessed over all created features;
- **Model 3 (MLPR)**: Multi-layer Perceptron regressor [81], which takes into account most relevant features assessed over all created features;

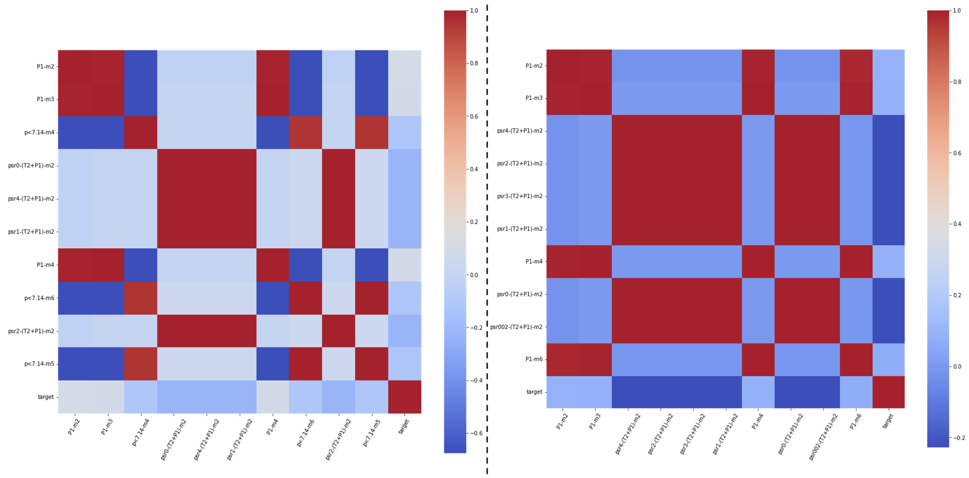

Fig. 8A

Fig. 8B

Figure 8. Feature correlation for ten selected features in each case, when forecasting if the amount of C5 present in distilled LPG at the end of the distillation process in the debutanizer columns remains within the required specification thresholds, for *Unit A*. On Fig. 8A we present feature correlations for Experiment 1, while on Fig. 8B we present feature correlations when considering Experiment 2.

- **Model 4 (VR)**: composite model introduced in Fig 9, and described in detail later in this section. The model takes into account most relevant features assessed over all created features.

While the *Baseline 2 (LiR)* is a linear regression model that forecasts C5 content based only on raw temperature and pressure readings obtained from two sensors, *Model 1 (LiR)* provides insights on how the forecasting quality is improved by introducing a more extensive set of features (all features presented in Table 3), considering Roult's law and Antoine's equations, given the assumptions and simplifications described in Section 4.3. *Model 2 (SVR)* and *Model 3 (MLPR)* were built based on the SVR and MLPR algorithms, which were frequently reported in the related work. We instantiated the *Model 2 (SVR)* model with a radial basis function kernel, using an epsilon value of 0,1 and non-scaled L2 regularization. We did not impose constraints on the number of iterations required by the solver. *Model 3 (MLPR)* was instantiated with a single hidden layer of a hundred neurons, using a ReLU activation [82] and the Adam solver [83]. The learning rate was set to a fixed constant (0,001), and we trained it for 300 iterations.

We designed *Model 4 (VR)* (see Fig 9) as a voting regressor (VR) [84] that takes the input from four estimators to decide on the final forecast. Two estimators are Catboost [85] models ((A) and (B)), each of them optimized with a different metrics function. (A) is optimized for the Root Mean Square Error (RMSE) metric, which tends to give more weight to points further away from the mean, and thus focuses on better adjusting off-specification values. On the other side, (B) is optimized for the Mean Absolute Error (MAE), which is not sensitive to outliers and ends up providing better estimates on the usual C5 levels. For both models, we use the expectile loss [86], which places unequal weights on disturbances. The expectile level (α) represents the center of mass of a probability distribution. The probabilities to the right are measured with α, while the probabilities to the left are measured with 1 - α[87]. Providing an asymmetric penalization of errors for the scored instances emphasizes instances whose output was not properly learned and yielded a greater forecast error. Both estimators are fed to the voting regressor. The voting regressor then issues a final forecast computing the mean predicted regression targets of the estimators in the ensemble.

It is important to highlight that though C5 content data is available from laboratory analysis, we avoid using features based on past C5 measurements to ensure that the final model can load the sensor data and provide real-time C5 content estimates.

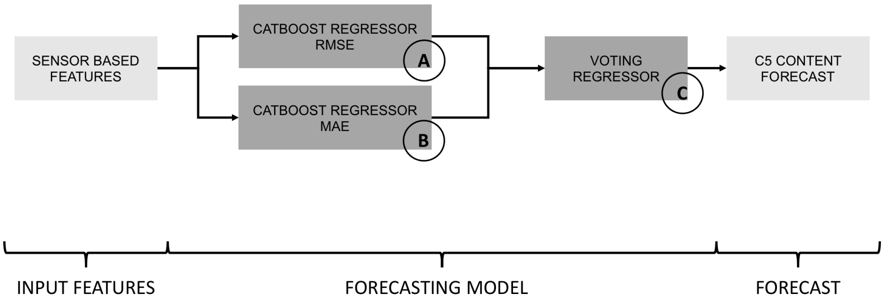

Figure 9. To estimate C5 content we created a voting regressor, that considers only sensor data as input. Regressors (A) and (B) correspond to Catboost models with different optimization objectives: (A) optimizes against RMSE, penalizing large errors, while (B) optimizes against MAE to achieve best median performance. Outputs from models (A), and (B) are weighted by the voting regressor (C), to create the final forecast.

4.4.2. Classification machine learning models

Forecasting if C5 content is off-specification is a challenging task given the strongly imbalanced data. Only 14 out of 167 measurements in *Unit A* corresponded to such events in our particular use case, while no such events were registered in *Unit B*. In our research, we compared six models; two baseline models and four models that aim to provide better predictions:

- **Baseline 1 (zero forecast)**: we predict no off-spec occurrence takes place;
- **Baseline 2 (LgR)**: logistic regression to predict C5 content based on raw temperature and pressure sensor measurements (P1 and T2 from Fig. 1, described in Feature Group ID #1 at Table 3);
- **Model 1 (LgR)**: logistic regression considering raw sensor measurements of P1 and T2 sensors at fifteen minute intervals (see Fig. 1), and all features described in Table 3), for the time range as presented in Fig. 4;
- **Model 2 (SVC)**: Support Vector Classifier [80], which takes into account most relevant features assessed over all created features;
- **Model 3 (MLPC)**: Multi-layer Perceptron Classifier [81], which takes into account most relevant features assessed over all created features;
- **Model 4 (Catboost)**: a CatBoost classifier with a Focal loss [88], which provides an asymmetric penalization to train instances, focusing more on those that are missclassified. The model takes into account most relevant features assessed over all created features.

The *Baseline 2 (LgR)* and *Model 1 (LgR)* were initialized with the same parameters, using a limited-memory Broyden–Fletcher–Goldfarb–Shanno solver algorithm [89–92], along with a L2 regularization. In both cases, a class balancing strategy was used to weights classes inversely proportional to class frequencies. *Model 2 (SVC)* was initialized with a radial basis function kernel and epsilon value of 0,1 and L2 regularization. We did not constrain the number of solvers' iterations. We initialized the *Model 3 (MLPC)* with a single hidden layer of a hundred neurons, with a ReLU activation and Adam solver. We used a constant learning rate (0,001) and trained the model over 300 iterations. Finally, the Catboost model was initialized with a Focal loss, growing asymmetric trees, a depth of six nodes, and a maximum number of sixty-four leaves. We trained the model over a thousand iterations, with a learning rate of 0,0299 evaluating against the AUC ROC metric. In all cases, we standarized features by removing the mean and scaling them to unit variance.

When building the classification models, we avoided using features based on past C5 measurements to ensure the models consume only data that can be provided in real-time, and thus issue real-time forecasts.

5. Experiments and Results

To evaluate the models presented in Section 4.4, we ran a repeated ten-fold cross-validation [93], executing fifty cross-validation runs. We conducted four experiments: two for regression models and two for classification models. Either for regression and classification, the experiments consisted of training the model only with historical data of the debutanizer unit we aim to predict for (Experiment 1), and to enrich the model validation with the data available from another debutanizer unit (Experiment 2). We present the corresponding cross-validation setting in Fig. 10. We ensure results from both experiments are comparable by preserving the same cross-validation test sets among both experiments. We also assessed if the differences in results obtained for the different models were statistically significant. To that end, we executed the Wilcoxon signed-rank test [94] and tested for significance at a 95% level.

5.1. Regression models

When implementing the experiments for the regression models presented in Section 4.4.1, we measured MAE and RMSE metrics. We present the results in Table 4 and Table 5. From Experiment 1, we observed the best overall performance was achieved with *Model 4 (VR)*, which demonstrates the best performances for all the scenarios except for one (RMSE for *Unit B*), where it achieved the second-best prediction. This performance was nearly matched by *Baseline 1 (C5 median)*, which achieved the best performance in three cases: *Unit B*, and MAE for *Unit A*. We consider the *Model 2 (SVR)* was the third-best model among the evaluated ones, achieving the second-best prediction in all cases, except for MAE at Unit A, where it matched the best performance displayed by *Model 4 (VR)* and *Baseline 1 (C5 median)*. Moreover, we found the *Model 1 (LiR)* demonstrated a significantly worse performance, which we attribute to the features' selection. We ground this conclusion on the fact that a better result was obtained by *Baseline (LiR)*, and while some improvement was observed when augmenting the data in Experiment 2, it did not match the performance of the rest of the models. The best overall performance for Experiment 2 was achieved by *Model 4 (VR)*, which achieved the best performance at *Unit A*, and second-best for *Unit B*. We consider the overall second-best performance was achieved by *Baseline 1 (C5 median)*, which had the best performance in *Unit B*, and second-best considering MAE at *Unit A*.

When comparing results from both experiments, we observed *Model 4 (VR)* displayed the best performance, surpassing the *Baseline 1 (C5 median)*. *Model 4 (VR)* is used to predict the C5 peaks providing good forecasts for low C5 levels, which reflects on the close results when compared to *Baseline 1 (C5 median)*. While the SVR algorithm is frequently reported in the literature, it achieved a third-best performance in Experiment 1 and degraded in Experiment 2. Using data from both units to train the models improved the performance of *Baseline 2 (LiR)* and *Model 4 (VR)* in all cases. It also improved the performance of *Model 1 (LiR)* for *Unit B*, and degraded the performance of *Model 3 (MLPR)* in all cases, except when measuring RMSE for *Unit B*.

Finally, we assessed which features were considered most informative by the feature selection criteria for both experiments. We found that from Experiment 1, the most relevant features were the pressure sensor readings, features from *Feature Group 6* (ratio between expected T1 temperature and P1 pressure, for the given LPG mixtures - see Table 3), and categorical features indicating whether the pressure sensor readings are below 7,14 kg/cm^2, or above 7,63 kg/cm^2. However, the set of relevant features in Experiment 2 changed. Among the most important ones, we found those from *Feature Group 2* (expected mixture saturation vapor pressure for considering temperature T2 - see Table 3), and those from *Feature Group 6*. Pressure measures were still considered relevant in Experiment 2, but their importance faded in the presence of the ones mentioned above.

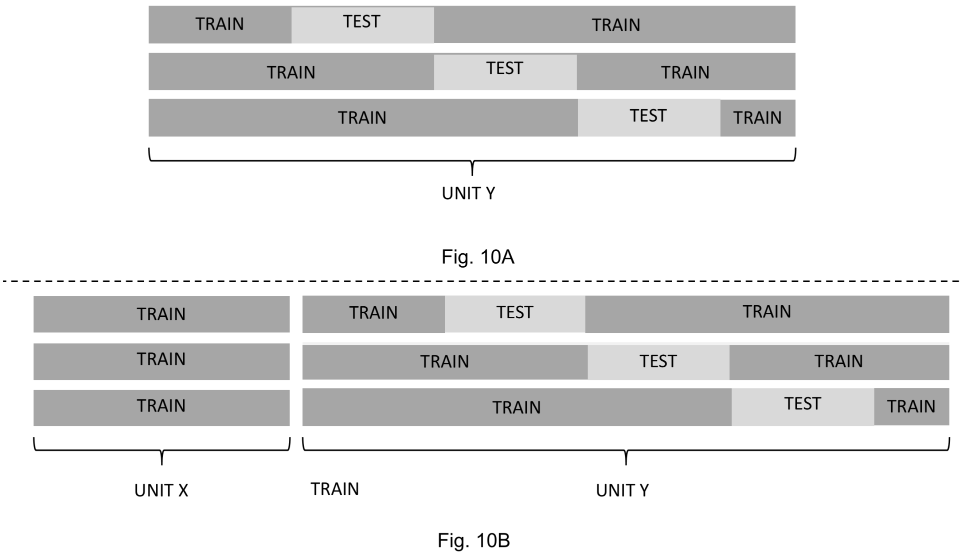

Fig. 10A

Fig. 10B

Figure 10. We pose two experiments: Experiment 1 trains models only with data of the debutanizer unit we aim to predict for (Fig. 10A), while Experiment 2 enriches the training set with data from other debutanizer unit (Fig. 10B).

Model	**Unit A**		**Unit B**	
	RMSE$_{\textbf{mean}}$	**MAE**$_{\textbf{mean}}$	**RMSE**$_{\textbf{mean}}$	**MAE**$_{\textbf{mean}}$
Baseline 1 (C5 median)	**1,1179	__*0,6028__	__0,1174__	__*0,0853__
Baseline 2 (LR)	**1,1794	0,7601	1,4150	0,7248
Model 1 (LR)	1632,9693	560,4650	96826,5563	46730,6566
Model 2 (SVR)	_*1,0754_	__*0,6087__	_*0,1240_	_0,0991_
Model 3 (MLPR)	_*1,0728_	_0,7122_	0,2115	0,1424
Model 4	__1,0352__	__*0,6127__	_*0,1201_	__*0,0871__

Table 4. Experiment 1 results. Mean RMSE and MAE values we obtained for different models with a ten-fold cross-validation, repeated fifty times. Best results are bolded, second-best are reported in italics. The results within the same column, which have no statistically significant difference between them when tested with a Wilcoxon paired rank test at a 95% confidence level, are marked with * and **.

5.2. Classification models

When implementing the experiments for the classification models presented in Section 4.4.2, we measured the AUC ROC metric. From the Table 6, we found that the best classification performance was obtained by *Model 4 (Catboost)* in both experiments, with an AUC ROC of at least 0,7359, and surpassing the second-best model by 0,065 points in the worst case. However, best results were achieved in Experiment 2.

The second-best model in Experiment 1 was the *Model 2 (MLPC)*, and the *Model 1 (LgR)* for Experiment 2. When comparing the models' performance across experiments, we observed an increased performance in Experiment 2 for the *Baseline 2 (LR)* and *Model 2 (SVC)* models. On the other hand, a decreased discrimination power was measured for *Model 1 (LgR)*, and *Model 3 (MLPC)*. *Model 2 (SVC)* performed worse than a zero forecast in both experiments, but this difference was not statistically significant in Experiment 2. We hypothesize that the performance decrease in Experiment 2 for certain models can be related to the stronger class imbalance (8% event occurrence in Experiment 1 is reduced to 5% event occurrence in Experiment 2). Such imbalance influences the learning of the algorithms, and most likely affects the discrimination power of the trained models. We found the differences between results within the experiments were statistically significant, except between the *Baseline 1 (zero forecast)* and *Model 2 (SVC)* in Experiment 2.

Model	**Unit A**		**Unit B**	
	RMSE$_{mean}$	**MAE$_{mean}$**	**RMSE$_{mean}$**	**MAE$_{mean}$**
Baseline 1 (C5 median)	1,1760↓	*0,6141*↓	**0,1152**↑	**0,0818**↑
Baseline 2 (LR)	*1,0603*↑	0,7009↑	**0,2126↑	0,1978↑
Model 1 (LR)	1198266158,9503↓	411001902,3054↓	**0,2753↑	0,1900↑
Model 2 (SVR)	1,1098↓	*0,6193*↓	**0,1287*↓	0,1021↓
Model 3 (MLPR)	1,0771↓	0,7234↓	0,2044↑	0,1581↓
Model 4	**0,9655**↑	**0,5743**↑	**0,1270*↓	*0,0852*↑

Table 5. Experiment 2 results. Mean RMSE and MAE values we obtained for different models with a ten-fold cross-validation, repeated fifty times. Best results are bolded, second-best are reported in italics. The results within the same column, which have no statistically significant difference between them when tested with a Wilcoxon paired rank test at a 95% confidence level, are marked with * and **. The arrows indicate whether the mean result improved (↑), or degraded (↓) when compared to Experiment 1.

Model	**Experiment 1**	**Experiment 2**
	AUC ROC$_{mean}$	**AUC ROC$_{mean}$**
Baseline 1 (zero forecast)	0,5000	*0,5000
Baseline 2 (LR)	0,5656	↑0,5675
Model 1 (LR)	0,6567	↓*0,6059*
Model 2 (SVC)	0,4491	*↑0,4897
Model 3 (MLPC)	*0,6709*	↓0,5381
Model 4 (Catboost)	**0,7359**	↑**0,7670**

Table 6. Out-of-specification detection results for *Unit A*. Mean ROC AUC values we obtained for different models with a ten-fold cross-validation, repeated fifty times. Best results are bolded, second-best are reported in italics. The results within the same column, which have no statistically significant difference between them when tested with a Wilcoxon paired rank test at a 95% confidence level, are marked with *. The arrows indicate whether the mean result improved (↑), or degraded (↓) when compared to Experiment 1. *Unit B* is not reported, since the dataset did not include out-of-spec measurements for *Unit B*.

Finally, we analyzed which features were considered most informative under the mutual information criteria for each experiment. For Experiment 1 the most informative features were the readings from the pressure sensor, categorical features indicating whether the sensed pressure is below 7,14 kg/cm^2, or below 7,06 kg/cm^2, and features from *Features Group 4* (see Table 3). This changed for Experiment 2, where the most important features were related to readings from the pressure sensor and the *Features Group 5*.

5.3. *Explaining Artificial Intelligence models*

While models' accuracy is of great importance, insights on models' rationale are required to assess the main factors driving the forecast are reasonable, and thus the forecast can be trusted. While much research in the literature was devoted to global explanations, we found a few authors taking into account the data or models' training process. Furthermore, we found no authors combined insights regarding feature selection, and how relevant the selected features are to the model across a repeated cross-validation. We therefore propose a novel visualization that summarizes the aforementioned insights (see Fig. 11).

While much research work in the literature has been devoted to global explanations, and some related work focuses on the characteristics of the dataset, little research has been done on integrating insights regarding the dataset, the experimental setting, and the resulting model. We therefore propose a novel visualization which combines the three aforementioned parts. Fig. 11A provides a brief description regarding the experimental setting. In this particular case, it states that the corresponding plots result from data obtained when training a forecasting model in a repeated 10-fold cross-validation setting, repeating the cross-validation 50 times. Fig. 11B shows a horizontal stacked bar plot,

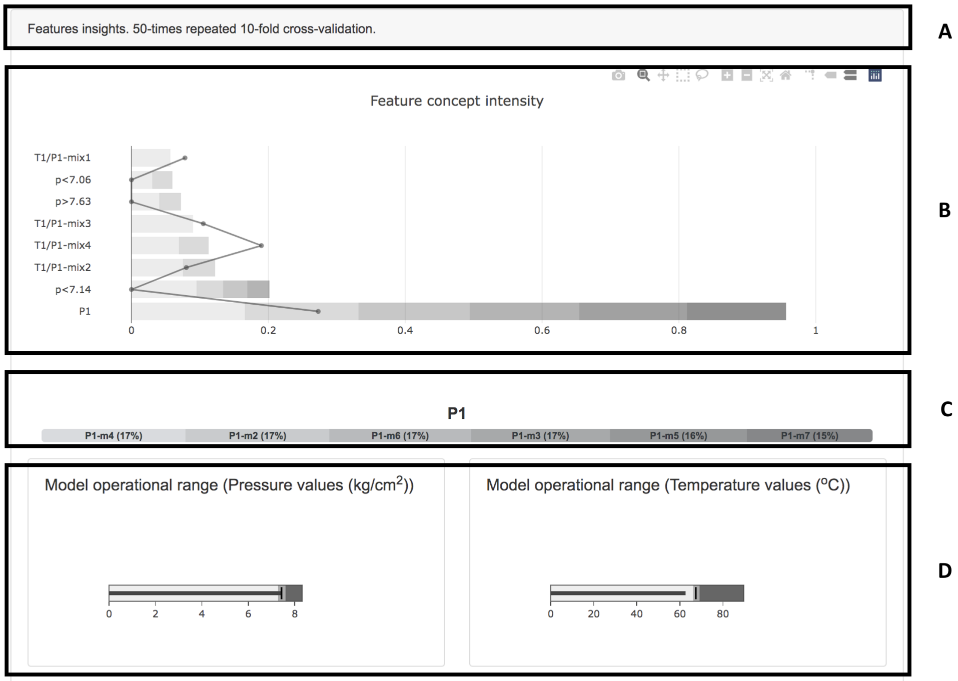
Features insights. 50-times repeated 10-fold cross-validation.
A
Feature concept intensity
T1/P1-mix1
p<7.06
p>7.63
T1/P1-mix3
T1/P1-mix4
T1/P1-mix2
p<7.14
P1
0
0.2
0.4
0.6
0.8
1
B
P1
P1-m4 (17%)
P1-m2 (17%)
P1-m6 (17%)
P1-m3 (17%)
P1-m5 (16%)
P1-m7 (15%)
C
Model operational range (Pressure values (kg/cm^2))
0
2
4
6
8
Model operational range (Temperature values (°C))
0
20
40
60
80
D

Figure 11. The visualization summarizes relevant information regarding the dataset, and forecasting model: (A) describes the cross-validation setting, (B) informs most relevant feature concepts when considering feature selection and models' features relevance, (C) details the weight of particular features within a feature concept, and (D) provides insights regarding values distribution for sensor data.

where the intensity of *feature concepts* is presented to the users. We consider *feature concepts* as semantic abstractions that group certain features based on features' metadata. In our particular case, such grouping was performed for features computed with the same formula, but using sensor data at different points in time (see Fig. 4). The shades of gray within the horizontal stacked bars represent how frequently was each feature of the *feature concepts* abstraction chosen when performing feature selection, within the 50 times 10-fold cross-validation. The line chart overlayed to the horizontal stacked bar plot informs how relevant were those feature concepts to the forecasting machine learning models, on average. Such overlay provides useful information to the machine learning engineer, who can remove features found not informative to the model, to give room to better ones. In this particular case, we found three such cases ($p < 7.06$, $p > 7.63$, and $p < 7.14$), referring to features with boolean values, assessing whether the pressure values obtained from the sensor were above or below certain threshold value. By clicking a particular *feature concept* in the horizontal bar stacked plot, the section highlighted in Fig. 11C is updated. Fig. 11C enriches the aforementioned view, detailing each feature's relevance within the specific *feature concept*. Finally, the bullet charts in Fig. 11D provide insights into the values distribution for both sensors (P1 and T2). We explain the bullet chart with greater detail in Fig. 12. The bullet chart has three segments, corresponding to quartiles Q1, Q2+Q3, and Q4. A vertical bar in the Q2+Q3 marks the median value, while a dark horizontal bar within the bullet chart, shows the mean value observed in the readings.

6. Conclusions

In this paper, two machine learning models are developed to forecast the concentration of pentanes (C5) in the LPG debutanization process. The first one is a regression model that provides pentane concentration estimates. The second one is a classifier that predicts whether the pentane concentration levels exceed allowed thresholds. Both models were designed to provide real-time forecasts based on sensor data. The advantages

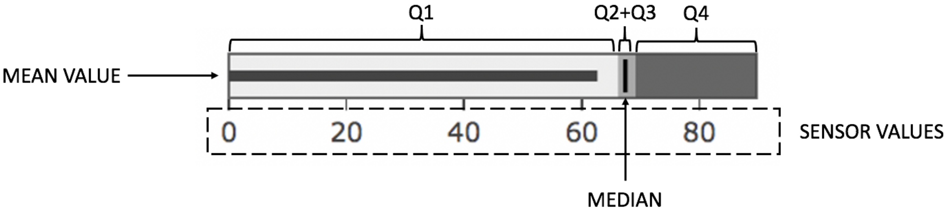

Figure 12. The bullet chart summarizes the sensor values distribution: three segments, Q1, Q2+Q3, and Q4 mark values related to the quartiles; a dark vertical bar within the Q2+Q3 segment represents the median value, while the dark horizontal bar within the bullet chart informs the mean readings' value.

of the models are that only two sensors are required (temperature and pressure sensors, located at two distinct points at the top of the debutanizer column). Both models were compared against several baseline models, and machine learning models developed based on algorithms cited in the literature.

Our experiments show that the best results for the pentanes concentration estimation was obtained with a voting regressor, trained with historical data of the debutanizer unit. The model surpasses the performance achieved by a baseline predicting the pentane concentration as a median of past values and a linear regressor predicting the concentration from raw sensor values. When predicting the off-specification detection, best results were achieved with a CatBoost classifier trained with a focal loss over the data of both debutanizer units considered in this research. The model achieved an AUC ROC of 0,7670. In both cases, the addition of data from another debutanizer unit boosted the learning and consequent performance of most of the models.

In addition to the aforementioned models, we developed a prototype dashboard, that allows to visualize relevant information regarding feature selection, features relevance to the model, and sensor reading values within which the model was trained. Such a dashboard is useful to assess strengths, limitations and improvement opportunities regarding the developed models.

We envision several directions for future research. Firstly, we would like to extend these experiments to a broader range of debutanizer units. Secondly, we would like to compare the current approach to more complex settings, where a broader range of sensors is available. Finally, we consider that this approach can be applied in other industries using distillation processes and where soft-sensors predicting specific substance concentrations are helpful.

Author Contributions: Conceptualization, J.M.R.; methodology, J.M.R.; software, J.M.R., E.T., A.K., and N.S.; validation, J.M.R. and A.K.; formal analysis, J.M.R. and A.K.; investigation, J.M.R., N.S.; resources, M.K.O., D.A., A.K and G.A., D.M., B.F.; data curation, A.K., N.S., J.M.R. and E.T.; writing—original draft preparation, J.M.R.; writing—review and editing, J.M.R., A.K., G.A., J.L., P.E, I.M., B.F. and D.M.; visualization, J.M.R.; supervision, K.K., A.K., G.A., B.F. and D.M.; project administration, G.A., B.F. and D.M.; funding acquisition, P.E., I.M., G.A., B.F. and D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovenian Research Agency and European Union's Horizon 2020 program project FACTLOG under grant agreement number H2020-869951.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
APC	Advanced Process Control
AUC ROC	Area Under the Receiver Operating Characteristic Curve
C1	Molecules with a single carbon atom
C2	Molecules with two carbon atoms
C4	Molecules with four carbon atoms
C5	Pentanes
CDU	Crude Distillation Unit
FCC	Fluid Catalytic Cracker
FG	Features Group
LgR	Logistic Regression
LiR	Linear Regression
LPG	Liquified Petroleum Gas
MAE	Mean Absolute Error
MLPC	Multi-layer Perceptron Classifier
MLPR	Multi-layer Perceptron regressor
MPC	Multivariable Model Predictive Control
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
SVC	Support Vector Classifier
SVR	Support Vector Regressor
VR	Voting Regressor

References

1. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **1997**, *30*, 1145 – 1159. doi:https://doi.org/10.1016/S0031-3203(96)00142-2.
2. Ferrer-Nadal, S.; Yélamos-Ruiz, I.; Graells, M.; Puigjaner, L. On-line fault diagnosis support for real time evolution applied to multi-component distillation. In *European Symposium on Computer-Aided Process Engineering-15, 38th European Symposium of the Working Party on Computer Aided Process Engineering*; Puigjaner, L.; Espuña, A., Eds.; Elsevier, 2005; Vol. 20, *Computer Aided Chemical Engineering*, pp. 961–966. doi:https://doi.org/10.1016/S1570-7946(05)80002-1.
3. Abdullah, Z.; Aziz, N.; Ahmad, Z. Nonlinear Modelling Application in Distillation Column. *Chemical Product and Process Modeling* **2007**, *2*. doi:doi:10.2202/1934-2659.1082.
4. Michelsen, F.A.; Foss, B.A. A comprehensive mechanistic model of a continuous Kamyr digester. *Applied Mathematical Modelling* **1996**, *20*, 523–533. doi:https://doi.org/10.1016/0307-904X(95)00171-F.
5. Franzoi, R.E.; Menezes, B.C.; Kelly, J.D.; Gut, J.A.; Grossmann, I.E. Cutpoint Temperature Surrogate Modeling for Distillation Yields and Properties. *Industrial & Engineering Chemistry Research* **2020**, *59*, 18616–18628.
6. Friedman, Y.; Neto, E.; Porfirio, C. First-principles distillation inference models for product quality prediction. *Hydrocarbon Processing* **2002**, *81*, 53–60.
7. Abdullah, Z.; Aziz, N.; Ahmad, Z. Nonlinear modelling application in distillation column. *Chemical Product and Process Modeling* **2007**, *2*.
8. Garcia, A.; Loria, J.; Marin, A.; Quiroz, A. Simple multicomponent batch distillation procedure with a variable reflux policy. *Brazilian Journal of Chemical Engineering* **2014**, *31*, 531–542.
9. Ryu, J.; Maravelias, C.T. Computationally efficient optimization models for preliminary distillation column design and separation energy targeting. *Computers & Chemical Engineering* **2020**, *143*, 107072.
10. Küsel, R.R.; Wiid, A.J.; Craig, I.K. Soft sensor design for the optimisation of parallel debutaniser columns: An industrial case study. *IFAC-PapersOnLine* **2020**, *53*, 11716–11721.
11. Ibrahim, D.; Jobson, M.; Guillen-Gosalbez, G. Optimization-based design of crude oil distillation units using rigorous simulation models. *Industrial & Engineering Chemistry Research* **2017**, *56*, 6728–6740.
12. Schäfer, P.; Caspari, A.; Schweidtmann, A.M.; Vaupel, Y.; Mhamdi, A.; Mitsos, A. The Potential of Hybrid Mechanistic/Data-Driven Approaches for Reduced Dynamic Modeling: Application to Distillation Columns. *Chemie Ingenieur Technik* **2020**, *92*, 1910–1920.
13. Bachnas, A.; Tóth, R.; Ludlage, J.; Mesbah, A. A review on data-driven linear parameter-varying modeling approaches: A high-purity distillation column case study. *Journal of Process Control* **2014**, *24*, 272–285.
14. Eikens, B.; Karim, M.N.; Simon, L. Neural Networks and First Principle Models for Bioprocesses. *IFAC Proceedings Volumes* **1999**, *32*, 6974–6979. 14th IFAC World Congress 1999, Beijing, Chia, 5-9 July, doi:https://doi.org/10.1016/S1474-6670(17)57190-6.
15. McBride, K.; Sanchez Medina, E.I.; Sundmacher, K. Hybrid Semi-parametric Modeling in Separation Processes: A Review. *Chemie Ingenieur Technik* **2020**, *92*, 842–855.

16. Schweidtmann, A.M.; Bongartz, D.; Huster, W.R.; Mitsos, A. Deterministic global process optimization: flash calculations via artificial neural networks. In *Computer Aided Chemical Engineering*; Elsevier, 2019; Vol. 46, pp. 937–942.
17. van Lith, P.F.; Betlem, B.H.; Roffel, B. Combining prior knowledge with data driven modeling of a batch distillation column including start-up. *Computers and Chemical Engineering* **2003**, *27*, 1021–1030.
18. Chen, L.; Hontoir, Y.; Huang, D.; Zhang, J.; Morris, A. Combining first principles with black-box techniques for reaction systems. *Control Engineering Practice* **2004**, *12*, 819–826. PC-B02-Process Control IFAC 2002, doi:https://doi.org/10.1016/j.conengprac.2003.09.006.
19. Cubillos, F.; Callejas, H.; Lima, E.; Vega, M. Adaptive control using a hybrid-neural model: application to a polymerisation reactor. *Brazilian Journal of Chemical Engineering* **2001**, *18*, 113 – 120.
20. Chang, J.S.; Lu, S.C.; Chiu, Y.L. Dynamic modeling of batch polymerization reactors via the hybrid neural-network rate-function approach. *Chemical Engineering Journal* **2007**, *130*, 19–28. doi:https://doi.org/10.1016/j.cej.2006.11.011.
21. Siddharth, K.; Pathak, A.; Pani, A. Real-time quality monitoring in debutanizer column with regression tree and ANFIS. *Journal of Industrial Engineering International* **2018**, *15*. doi:10.1007/s40092-018-0276-4.
22. Ochoa-Estopier, L.M.; Jobson, M. Optimization of Heat-Integrated Crude Oil Distillation Systems. Part I: The Distillation Model. *Industrial & Engineering Chemistry Research* **2015**, *54*, 4988–5000. doi:10.1021/ie503802j.
23. Shang, C.; Yang, F.; Huang, D.; Lyu, W. Data-driven soft sensor development based on deep learning technique. *Journal of Process Control* **2014**, *24*, 223–233.
24. Michalopoulos, J.; Papadokonstadakis, S.; Arampatzis, G.; Lygeros, A. Modelling of an Industrial Fluid Catalytic Cracking Unit Using Neural Networks. *Chemical Engineering Research and Design* **2001**, *79*, 137–142. doi:https://doi.org/10.1205/02638760151095944.
25. Bollas, G.; Papadokonstadakis, S.; Michalopoulos, J.; Arampatzis, G.; Lappas, A.; Vasalos, I.; Lygeros, A. Using hybrid neural networks in scaling up an FCC model from a pilot plant to an industrial unit. *Chemical Engineering and Processing: Process Intensification* **2003**, *42*, 697–713. Application of Neural Networks to Multiphase Reactors, doi:https://doi.org/10.1016/S0255-2701(02)00206-4.
26. Fortuna, L.; Graziani, S.; Xibilia, M. Soft sensors for product quality monitoring in debutanizer distillation columns. *Control Engineering Practice* **2005**, *13*, 499–508. doi:https://doi.org/10.1016/j.conengprac.2004.04.013.
27. Pani, A.K.; Amin, K.G.; Mohanta, H.K. Soft sensing of product quality in the debutanizer column with principal component analysis and feed-forward artificial neural network. *Alexandria Engineering Journal* **2016**, *55*, 1667–1674. doi:https://doi.org/10.1016/j.aej.2016.02.016.
28. Ge, Z.; Song, Z. A comparative study of just-in-time-learning based methods for online soft sensor modeling. *Chemometrics and Intelligent Laboratory Systems* **2010**, *104*, 306–317. doi:https://doi.org/10.1016/j.chemolab.2010.09.008.
29. Zheng, J.; Song, Z.; Ge, Z. Probabilistic learning of partial least squares regression model: Theory and industrial applications. *Chemometrics and Intelligent Laboratory Systems* **2016**, *158*, 80–90. doi:https://doi.org/10.1016/j.chemolab.2016.08.014.
30. Ge, Z. Active learning strategy for smart soft sensor development under a small number of labeled data samples. *Journal of Process Control* **2014**, *24*, 1454–1461. doi:https://doi.org/10.1016/j.jprocont.2014.06.015.
31. Ge, Z. Supervised Latent Factor Analysis for Process Data Regression Modeling and Soft Sensor Application. *IEEE Transactions on Control Systems Technology* **2016**, *24*, 1004–1011. doi:10.1109/TCST.2015.2473817.
32. Yao, L.; Ge, Z. Locally Weighted Prediction Methods for Latent Factor Analysis With Supervised and Semisupervised Process Data. *IEEE Transactions on Automation Science and Engineering* **2017**, *14*, 126–138. doi:10.1109/TASE.2016.2608914.
33. Yuan, X.; Ye, L.; Bao, L.; Ge, Z.; Song, Z. Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA. *Chemometrics and Intelligent Laboratory Systems* **2015**, *147*, 167–175. doi:https://doi.org/10.1016/j.chemolab.2015.08.014.
34. Bidar, B.; Sadeghi, J.; Shahraki, F.; Khalilipour, M.M. Data-driven soft sensor approach for online quality prediction using state dependent parameter models. *Chemometrics and Intelligent Laboratory Systems* **2017**, *162*, 130–141. doi:https://doi.org/10.1016/j.chemolab.2017.01.004.
35. Mohamed Ramli, N.; Hussain, M.; Mohamed Jan, B.; Abdullah, B. Composition Prediction of a Debutanizer Column using Equation Based Artificial Neural Network Model. *Neurocomputing* **2014**, *131*, 59–76. doi:https://doi.org/10.1016/j.neucom.2013.10.039.
36. Subramanian, R.; Moar, R.R.; Singh, S. White-box Machine learning approaches to identify governing equations for overall dynamics of manufacturing systems: A case study on distillation column. *Machine Learning with Applications* **2021**, *3*, 100014. doi:https://doi.org/10.1016/j.mlwa.2020.100014.
37. Shi, J.; Wan, J.; Yan, H.; Suo, H. A survey of cyber-physical systems. 2011 international conference on wireless communications and signal processing (WCSP). IEEE, 2011, pp. 1–6.
38. Chen, H. Applications of cyber-physical system: a literature review. *Journal of Industrial Integration and Management* **2017**, *2*, 1750012.
39. Lu, Y. Cyber physical system (CPS)-based industry 4.0: A survey. *Journal of Industrial Integration and Management* **2017**, *2*, 1750014.
40. Khodabakhsh, A.; Ari, I.; Bakir, M.; Ercan, A.O. Multivariate Sensor Data Analysis for Oil Refineries and Multi-mode Identification of System Behavior in Real-time. *IEEE Access* **2018**, *6*, 64389–64405.
41. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* **2018**, *6*, 52138–52160.
42. Keller, P.; Drake, A. Exclusivity and Paternalism in the public governance of explainable AI. *Computer Law & Security Review* **2021**, *40*, 105490.

43. El-Assady, M.; Jentner, W.; Kehlbeck, R.; Schlegel, U.; Sevastjanova, R.; Sperrle, F.; Spinner, T.; Keim, D. Towards XAI: Structuring the Processes of Explanations. ACM Workshop on Human-Centered Machine Learning, 2019.
44. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A brief survey on history, research areas, approaches and challenges. CCF international conference on natural language processing and Chinese computing. Springer, 2019, pp. 563–574.
45. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; others. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **2020**, *58*, 82–115.
46. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113.
47. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* **2020**.
48. Alicioglu, G.; Sun, B. A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics* **2021**.
49. Messalas, A.; Kanellopoulos, Y.; Makris, C. Model-agnostic interpretability with shapley values. 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2019, pp. 1–7.
50. Frye, C.; de Mijolla, D.; Begley, T.; Cowton, L.; Stanley, M.; Feige, I. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272* **2020**.
51. Polley, S.; Koparde, R.R.; Gowri, A.B.; Perera, M.; Nuernberger, A. Towards trustworthiness in the context of explainable search. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2580–2584.
52. Chan, G.Y.Y.; Bertini, E.; Nonato, L.G.; Barr, B.; Silva, C.T. Melody: Generating and Visualizing Machine Learning Model Summary to Understand Data and Classifiers Together. *arXiv preprint arXiv:2007.10614* **2020**.
53. Chan, G.Y.Y.; Yuan, J.; Overton, K.; Barr, B.; Rees, K.; Nonato, L.G.; Bertini, E.; Silva, C.T. SUBPLEX: Towards a Better Understanding of Black Box Model Explanations at the Subpopulation Level. *arXiv preprint arXiv:2007.10609* **2020**.
54. Krause, J.; Perer, A.; Bertini, E. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics* **2014**, *20*, 1614–1623.
55. Seifert, C.; Aamir, A.; Balagopalan, A.; Jain, D.; Sharma, A.; Grottel, S.; Gumhold, S. Visualizations of deep neural networks in computer vision: A survey. In *Transparent data mining for big and small data*; Springer, 2017; pp. 123–144.
56. Jin, W.; Carpendale, S.; Hamarneh, G.; Gromala, D. Bridging ai developers and end users: an end-user-centred explainable ai taxonomy and visual vocabularies. *Proceedings of the IEEE Visualization, Vancouver, BC, Canada* **2019**, pp. 20–25.
57. Hudon, A.; Demazure, T.; Karran, A.; Léger, P.M.; Sénécal, S. Explainable Artificial intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. *Fred D. Davis, René Riedl, Jan vom Brocke, Pierre-Majorique Léger, Adriane B. Randolph, Gernot Müller-Putz (Eds.)*, p. 263.
58. Joia, P.; Coimbra, D.; Cuminato, J.A.; Paulovich, F.V.; Nonato, L.G. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics* **2011**, *17*, 2563–2571.
59. Wang, J.; Gou, L.; Zhang, W.; Yang, H.; Shen, H.W. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE transactions on visualization and computer graphics* **2019**, *25*, 2168–2180.
60. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
61. Collaris, D.; van Wijk, J.J. ExplainExplore: Visual exploration of machine learning explanations. 2020 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2020, pp. 26–35.
62. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* **2016**.
63. Alvarez-Melis, D.; Jaakkola, T.S. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538* **2018**.
64. Viton, F.; Elbattah, M.; Guérin, J.L.; Dequen, G. Heatmaps for Visual Explainability of CNN-Based Predictions for Multivariate Time Series with Application to Healthcare. 2020 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2020, pp. 1–8.
65. Rožanec, J.; Trajkova, E.; Kenda, K.; Fortuna, B.; Mladenić, D. Explaining Bad Forecasts in Global Time Series Models. *Applied Sciences* **2021**, *11*. doi:10.3390/app11199243.
66. Greenwell, B.M. pdp: An R package for constructing partial dependence plots. *R J.* **2017**, *9*, 421.
67. Apley, D.W.; Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2020**, *82*, 1059–1086.
68. UNION, P. Regulation (EC) No 715/2007 of the European Parliament and of the Council. Technical report, 715/2007/EC, 2007.
69. Behnasr, M.; Jazayeri-Rad, H. Robust data-driven soft sensor based on iteratively weighted least squares support vector regression optimized by the cuckoo optimization algorithm. *Journal of Natural Gas Science and Engineering* **2015**, *22*, 35–41. doi:https://doi.org/10.1016/j.jngse.2014.11.017.
70. Aston, J.; Messerly, G. Additions and Corrections-The Heat Capacity and Entropy, Heats of Fusion and Vaporization, and the Vapor Pressure of n-Butane. *Journal of the American Chemical Society* **1941**, *63*, 3549–3549.
71. Das, T.R.; Reed Jr, C.O.; Eubank, P.T. PVT [pressure-volume-temperature] surface and thermodynamic properties of butane. *Journal of Chemical and Engineering Data* **1973**, *18*, 244–253.

72. Carruth, G.F.; Kobayashi, R. Vapor pressure of normal paraffins ethane through n-decane from their triple points to about 10 mm mercury. *Journal of Chemical and Engineering Data* **1973**, *18*, 115–126.
73. Kemp, J.; Egan, C.J. Hindered rotation of the methyl groups in propane. The heat capacity, vapor pressure, heats of fusion and vaporization of propane. Entropy and density of the gas. *Journal of the American Chemical Society* **1938**, *60*, 1521–1525.
74. Rips, S. On a Feasible Level of Filling in of Reservoires by Liquid Hydrocarbons. *Khim. Prom.(Moscow)* **1963**, *8*, 610–613.
75. Helgeson, N.; Sage, B.H. Latent heat of vaporization of propane. *Journal of Chemical and Engineering Data* **1967**, *12*, 47–49.
76. Yaws, C.; Yang, H. To estimate vapor pressure easily. *Hydrocarbon Processing;(USA)* **1989**, *68*.
77. Osborn, A.G.; Douslin, D.R. Vapor-pressure relations for 15 hydrocarbons. *Journal of Chemical and Engineering Data* **1974**, *19*, 114–117.
78. Hua, J.; Xiong, Z.; Lowey, J.; Suh, E.; Dougherty, E.R. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* **2005**, *21*, 1509–1515.
79. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Physical review E* **2004**, *69*, 066138.
80. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V.; others. Support vector regression machines. *Advances in neural information processing systems* **1997**, *9*, 155–161.
81. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **1943**, *5*, 115–133.
82. Fukushima, K. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics* **1969**, *5*, 322–333.
83. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
84. An, K.; Meng, J. Voting-averaged combination method for regressor ensemble. International Conference on Intelligent Computing. Springer, 2010, pp. 540–546.
85. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* **2018**.
86. Aigner, D.; Lovell, C.K.; Schmidt, P. Formulation and estimation of stochastic frontier production function models. *Journal of econometrics* **1977**, *6*, 21–37.
87. Ehm, W.; Gneiting, T.; Jordan, A.; Krüger, F. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* **2016**, pp. 505–562.
88. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
89. Broyden, C.G. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics* **1970**, *6*, 76–90.
90. Fletcher, R. A new approach to variable metric algorithms. *The computer journal* **1970**, *13*, 317–322.
91. Goldfarb, D. A family of variable-metric methods derived by variational means. *Mathematics of computation* **1970**, *24*, 23–26.
92. Shanno, D.F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation* **1970**, *24*, 647–656.
93. Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* **1974**, *36*, 111–133.
94. Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in statistics*; Springer, 1992; pp. 196–202.