Results in Engineering 13 (2022) 100362

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Machine learning based adaptive soft sensor for flash point inference in a refinery realtime process

Izaskun Mendia [a,*], Sergio Gil-López [a], Itziar Landa-Torres [b], Lucía Orbe [b], Erik Maqueda [a]

[a] TECNALIA, Basque Research and Technology Alliance (BRTA), 48160, Derio, Bizkaia, Spain
[b] Petronor Innovación S.L., 48550, Muskiz, Bizkaia, Spain

ARTICLE INFO

Keywords:
Flash-point temperature
Control industry process
Adaptive soft sensor
Virtual sensing
Inferential sensing
Data-driven techniques

ABSTRACT

In industrial control processes, certain characteristics are sometimes difficult to measure by a physical sensor due to technical and/or economic limitations. This fact is especially true in the petrochemical industry. Some of those quantities are especially crucial for operators and process safety. This is the case for the automotive diesel Flash Point Temperature (FT). Traditional methods for FT estimation are based on the study of the empirical inference between flammability properties and the denoted target magnitude. The necessary measures are taken indirectly by samples from the process and analyzing them in the laboratory, this process implies time (can take hours from collection to flash temperature measurement) and thus make it very difficult for real-time monitorization, which in fact results in security and economical losses. This study defines a procedure based on Machine Learning modules that demonstrate the power of real-time monitorization over real data from an important international refinery. As input, easily measured values provided in real-time, such as temperature, pressure, and hydraulic flow are used and a benchmark of different regressive algorithms for FT estimation is presented. The study highlights the importance of sequencing preprocessing techniques for the correct inference of values. The implementation of adaptive learning strategies achieves considerable economic benefits in the productization of this soft sensor. The validity of the method is tested in the reality of a refinery. In addition, real-world industrial data sets tend to be unstable and volatile, and the data is often affected by noise, outliers, irrelevant or unnecessary features, and missing data. This contribution demonstrates with the inclusion of a new concept, called an adaptive soft sensor, the importance of the dynamic adaptation of the conformed schemes based on Machine Learning through their combination with feature selection, dimensional reduction, and signal processing techniques. The economic benefits of applying this soft sensor in the refinery's production plant and presented as potential semi-annual savings.

1. Introduction

Refineries produce multiple petroleum subproducts that are used for various applications. For safety considerations and to be commercialized, these products have to meet a set of quality specifications [1]. One of the critical quality specifications for automotive diesel is Flash Point Temperature (FT) [2]. The FT is a property of the diesel that indicates the lowest temperature at which there will be enough flammable vapor to ignite when an ignition source is applied. It is determined by the number of light hydrocarbons present in the diesel and dictates the flammability of the fuel. It is commonly monitored after the diesel has been processed at the desulphuration industrial unit. In the desulfurization unit, the process of conditioning the load streams (coming from the atmospheric and vacuum distillation units and thermal and catalytic units) to commercial diesel specifications takes place. This process consists mainly of catalytic hydrogenation that removes sulfur, nitrogen, oxygen compounds, and other metallic impurities present in the feedstock streams, resulting in a more refined product that meets the customer's handling, transportation, and combustibility requirements. The process carried out in the refinery plant is complex, and it's formed of a large number of physicochemical reactions to meet the quality and safety specifications (European standard UNE-EN 590).

The common methods for FT prediction are based on the study of the empirical correlation between physical properties based on the

* Corresponding author.
E-mail addresses: izmendi@hotmail.com, izaskun.mendia@tecnalia.com (I. Mendia), sergio.gil@tecnalia.com (S. Gil-López), itziar.landa@repsol.com (I. Landa-Torres), lucia.orbe@repsol.com (L. Orbe), erik.maqueda@tecnalia.com (E. Maqueda).

https://doi.org/10.1016/j.rineng.2022.100362
Received 17 December 2021; Received in revised form 18 January 2022; Accepted 1 February 2022
Available online 16 February 2022

assumption that structurally similar compounds have similar flammable activities [3]. These physical properties are relative to the normal boiling point (NBP) and the enthalpy of vaporization (Hv), given their relationship to the volatility and hence flammability of the fuel [4]. Measurements are performed indirectly by taking samples from the process and analyzing them in the laboratory using experimental measurement methods. However, nowadays, new developments are focused on the group contribution models (GCM) and quantitative structure-property relationship (QSPR) models [5,6]. (i) The GCM approach is used to predict the thermodynamic properties of organic compounds from their molecular structure. (ii) The QSRP approach uses information from molecular descriptors to represent the characteristics of numerous organic compounds. These descriptors numerically represent various chemical structural properties such as constitutional, topological, geometrical, thermodynamic, quantum chemical, and charge-related characteristics [7]. Both GCM and QSRP approaches use the molecular information of organic compounds as input predictor variables to the models [8].

In parallel, GCM and QSPR models have been developed using various Machine Learning algorithms, increasing the predictability of these models [7] through non-linear techniques mainly for the selection of the most relevant predictors. Algorithms such as, e.g., artificial neural network (ANN), support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), and non-linear regressions are the most popular [7]. A good literature review on FT calculation can be found in the following literature reviews [9–11].

However, in the process industry, these methods have several operational limitations. Experimental methods, although always preferable to any other inference technique in terms of reliability, require sample analysis tasks in the laboratory that can take hours from collection to flash temperature measurement [12], and sometimes they are extremely difficult processes [6]. As an alternative to obtaining the FT experimentally, MCQ and QSPR-based methods often use estimation methods based on available data in public databases, the most popular are DIPPR, Merck, NIOSH, and the chemical database of Akron University. In these cases, the methods refer only to pure compounds, where molecular information is indispensable [13,14]. In some other cases, studies recognize the limitations of the physical models used (Liaw's model) especially when this molecular information does not correctly define some properties such as isomers [15]. The lack of real-time FT information, either due to the delay of experimental methods or to the lack of knowledge of the chemical structural properties [13] of the compounds, can lead to economic losses.

The author is not aware of any previous work in the literature for the inference of the FT value in real-time. Moreover, at least in this use case, there is no physical sensor capable of measuring the flash temperature continuously in this refinery environment. In the reality of this use case, the measurement is done indirectly by taking a process sample and analyzing it in the laboratory using an experimental measurement method known as Pensky's "closed cup" method [4,16]. It is also controlled with an inference based on classical rigorous models.

In this work, we propose for the real-time inference of the FT value a procedure based on soft sensors. Based on machine learning techniques, soft sensors can infer the value of a certain magnitude from the indirect measurement of other magnitudes. In other words, a data-driven soft sensor is defined as an inference scheme capable of learning certain multi-parametric and highly non-linear causality relationships from a set of historical data [17]. Its main requirement is the existence of [18] data. In this use case, the variables that determine the natural variability of the process are variables related to temperature, pressure and hydraulic flow. Its use is particularly suitable for the operation of certain industrial processes, e.g. the measurement of the chemical composition of certain compounds in petrochemical distillation processes, chemical companies, cement plants, paper industry, nuclear power plants, among others. A complete list of applications based on soft sensors or virtual sensing can be found in Ref. [19]. The main competitive advantages of these soft sensing methods are [20]:

- They do not require specific knowledge of the parametric equations governing the physical relationships of the problem to be addressed. Therefore, they do not require prior knowledge of the characteristics of diesel and related chemical reactions.
- They are schemes with high inference capacity in highly non-linear multi-parametric relationships. In this case, although the molecular compositions of diesel are unknown, this information is indirectly implicit in the operating variables. The scheme is able to infer TP through these process variables.
- They are systems that offer, a relatively low design cost, high generalisability.

When the industrial process to be modeled responds to stable, non-volatile behavior over time, a non-adaptive strategy is sufficient for soft sensor modeling. But real-world industrial behaviors tend to be unstable and volatile and the data is often dirty, noisy, and contains outliers, irrelevant or unnecessary features, and null or non-standardized values [21]. To solve these problems, the authors [22] introduce the term adaptive soft sensor as a concept to assimilate the behavioral changes related to the dynamic transformation of the process. Soft sensors can adopt adaptive strategies based on small training fragments of reduced dimensions capable of delimiting a sufficiently stable caustic (a term known as window) to train with such a group of instances and to predict with them the closest immediate instants. The windows make up for the lack of knowledge of the chemical composition of the diesel because this information is indirectly implicit in the process variables. The size of the window determines the generalisability of the model and the level of redundancy of the data, as well as the fast or slow adaptation to sudden changes in the model: the smaller the window, the less the data is affected before the change and the better the adaptation [23]. On the other hand, though, it can produce model estimates with high variance (especially in the presence of a large number of collinear process variables). The big advantage is that by composing a set of local linear models, one can approximately describe a non-linear process [17, 22].

In other industrial disciplines, come works propose novel empirical expressions for process modeling. Such as the one suggested by the authors in Ref. [24] concerning metallurgical processes and their properties. In this case, such expressions allow correlating thermo-physical properties based on temperature and tin molar composition. Or as suggested by the authors in Ref. [25] to estimate the fan width in the paint spray application process. In this case, the fan width is often determined in a trial and error method but now the authors propose a linear regression model based on process parameters such as velocity and flow rate ejection and shapping air. And even to infer non-linear knowledge learned through physical laws in the field of civil engineering, as proposed by the authors in Ref. [26]. Finally [27], proposed to model the conditions governing the complex hydro-hydraulic flow process in a cathodic cell through a novel device through a numerical scheme.

According to the author, this work is the first known approach capable of inferring FT using experimental measurements and operational data. This work evaluates the relationship between the complexity of the machine learning methods used and the quality of the prediction through a statistical study of how they affect relevant feature selection techniques, dimensional reduction techniques, and signal processing techniques, in combination with different regressive algorithms. From regressive algorithms such as Ridge -regularised linear regressive algorithm-to tree assembly algorithms such as RandomForest and XGBoost in bagging or boosting and without the need to implement complex deep techniques. This paper presents a methodology based on Artificial Intelligence applied to real plant operation data.

2. Case study: Flash point temperature inference in a realtime process

2.1. Dataset

A study period of 3 years (2017, 2018, and 2019) was considered for the conceptualization of the model. The average sampling frequency of the process sensors was 1 min.

For the time-lag analysis that studies the dynamics of the process, it was agreed to consider process lags of no more than 4 h before sampling. This was because it is considered that the hydraulic process lag was below 4 h so any earlier value should not influence the flash temperature measured in the laboratory.

Data from all the units defined by refinery technicians was collected. This data contained all the physical parameters (temperature, pressure, and hydraulic flow) able to characterize the lighter components of each of the streams that affect the flash temperature. Access to the data was done through web services (client/server REST request mechanism [28]) that provide information from each of the streams. All data was stored in a single serializable object for the subsequent study described in the following sections.

2.2. Cross-validation

This use case deals with a dynamically changing process with a significant time effect. In order to provide a more robust model performance, the validity of the solution was tested with split time series cross-validation [29], a method based on the expanding training window concept. The size of the training window w+1 was continuously expanding with new instances while the testing size h remained constant. The training and testing data were daily instances (every 24 h).

The RMSE metric was used as the measure of success and calculated for each window. If the RMSE value was below a certain threshold (based on the reproducibility stated by the standard ASTM D97 method), that instance was included in the training set of the next training window. Otherwise, it was discarded.

The 10 min instances output value (flash temperature) was inferred with the respective daily model. However, the output value of these 10-min instances was unknown and it was not possible to quantify the error. Fig. 1 illustrates this procedure.

The training, testing, and inference time was in the range of 1–4s on an Intel i7 processor running at 2.6 GHz with 32 GB of RAM.

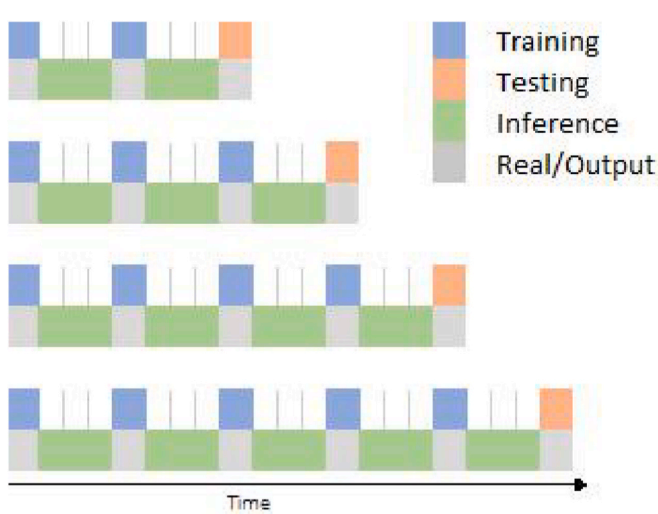

Fig. 1. Minutal inference and timeseries cross-validation.

2.3. Methodology

2.3.1. Data collection and Pre-processing techniques

The data collected directly from the production units had to be conditioned for use in the design of the soft sensor, and the following steps were taken:

- If the unit was operating in a different mode of operation than desired, data were discarded. For example, instances, where the unit was in startup or recirculation mode, were eliminated and only those instances where the main feed was kerosene, were considered.
- Inputs with a single constant value were discarded. Those inputs whose value did not vary over time and remained constant, did not provide information on the process dynamics, were completely superfluous, and should be eliminated.
- Inputs with a high percentage (>90%) of null values were discarded.

2.3.2. Input selection techniques

To identify the relevant inputs, the Permutation-based Importance (PIMP) technique [30,31] and algorithms for feature selection were applied among all the characteristics. The PIMP technique is based on a gradient boosting model capable of determining the importance of each input and it allowed to discard inputs with no effect on the flash point temperature. The Random Forest and Gradient Boosting algorithms, through their "feature importance" function, make it possible to quantify the effect of each input. From the combination of both (technique and algorithms), it was calculated the relevant inputs. Likewise, a weighting process was established between the chosen inputs and those inputs that were frequently chosen as relevant by the different algorithms. On the other hand, with the help of refinery technicians and their knowledge of the process, the inputs that should be most important from a physical point of view were identified.

After performing several tests with the different inputs identified by means of the feature selection techniques and the list provided by refinery technicians, the list of the most relevant inputs that provided the best results was established. Those variables represent different stream flows, pressuresand temperatures that can be measures of the process or set points and controller outputs.

In order to avoid redundancies and co-linearities, principal component analysis technique (PCA) [32,33] as the dimensional reduction was used and it was estimated that just a few were the main inputs needed to explain at least 95% of the value of the native variables.

In addition, it was performed an analysis of the relevant inputs for different time ranges to check if they remained constant or not. The different studies performed (adversarial validation and feature consistency over time) concluded that the relevant inputs varied over time. This fact suggested that it was not enough to train a non-adaptive model with a prefixed data set, but that it was necessary to opt for an adaptive scheme (moving or extensive window strategy) that would be retrained as new data arrived.

2.3.3. Time-lag analysis

Time lagged cross-correlation (TLCC) [34] allowed to study the synchrony between each input and the target output by calculating the instant when the correlation was maximum. Fig. 2 shows the correlation values between a temperature input and the flash temperature and indicates $3\frac{1}{2}$ hour temporary decalage (209 min lag) as the maximum correlation.

2.3.4. Strategy selection and model training

Soft sensor systems must adapt to gradual changes in behavior and must provide optimal performance even in the presence of process variations. This is achieved by updating and adapting the model of the soft sensor with each new instance iteration [17]. Two adaptive learning strategies arise to mitigate the effects of concept drift, giving rise to what

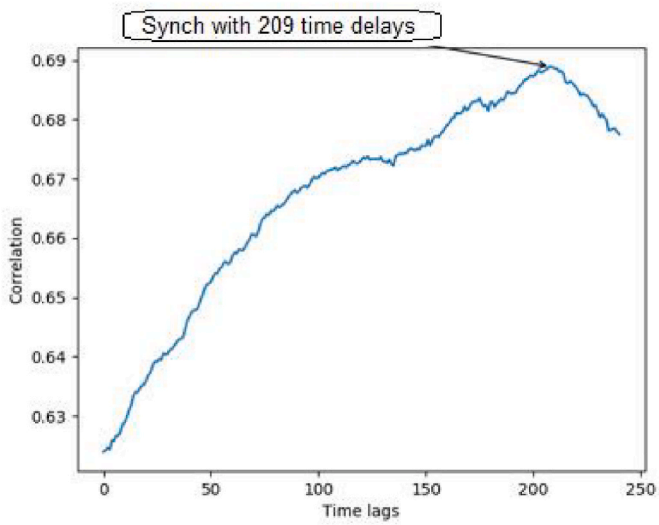

Fig. 2. Time lagged cross-correlation analysis for a temperature input.

is known as adaptive soft sensor learning [35]:

- Moving Window (MW) strategy. This method uses a moving learning dataset formed of a fixed number of points, also known as the window size. This moving dataset is used to continuously retrain the model by incorporating new knowledge and removing old ones. After each iteration, once the model is retrained the next time step is forecasted. The performance scheme has a stronger dependency and correlation relationship with the new instances than with the older ones [36–38] even when the process starts to change gradually [39]. The scheme is depicted in Fig. 3a.
- Extending Window (EW) strategy. It is a walk forward scheme an incremental training window that expands after each iteration with a new instance. Adaptation to new behaviors is slow precisely because of the weight of past behaviors that are not forgotten, see Fig. 3b.

Each of these strategies requires an algorithm on each of the windows. The overall study has been implemented based on four reliable approaches, including:

- Ridge Regression, a linear algorithm based on the classical linear regression model but regularising the impact on non-relevant features;
- Support Vector Regression (SVR) [40], the non-linear algorithm that transforms the data into a higher dimensional feature space to make it possible to perform linear separation;
- Random Forest Regression, a non-linear algorithm that has the ability to act as an ensemble algorithm by bagging individual trees; and
- XGBoost [41] as efficient implementation of Gradient Boosting regression. A linear or non-linear algorithm depends on the kernel used. The present work compares a non-linear kernel *gbtree* as the ensemble algorithm for boosting individual trees.

2.3.5. Validation

Model performance evaluation remains a subjective matter for a data scientist, since the performance evaluation is closely related to the chosen learning strategy, to the algorithm, and to the aspects that refinery technicians need to prioritize. In order to use a common jargon with refinery technicians, on the present work the evaluation metric was based on "reliability", "repeatability" and "reproducibility" [42]. But, for technical metric, the Root Mean Square Error (RMSE) loss function is adopted. The RMSE formula is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{1}$$

where y and $\hat{y}$ are the actual and predicted values of the validation data respectively and n is the number of elements of the validation data.

2.4. Methods comparison and discussion

The study compares the soft sensor validation through different pre-processing techniques, application of regressors, and learning strategies. The pre-processing techniques are probabilistic data cleaning, normalization between the maximum and minimum values of each feature (minmax-scaler), noise removal by smoothing the input with the Savitzky-Golay filter (savgol-filter), and dimensionality reduction with PCA (pca-decomposition). The regressors used are Ridge, RF, XGB, and SVR. The adaptive learning strategies are MW and EW. Only the results computed on the daily test dataset are reported.

The result of testing the different combinations of pre-processing techniques over the relevant inputs and the reaction of each of the above algorithms is shown in Table 1. The values of the RMSE error metric are worse/greater in the linear Ridge algorithm ($RMSE_{relevantfeatures}$, $Ridge,minmax,pca = 3.905$) compared to non-linear algorithms ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$, $RMSE_{relevantfeatures,XGB,probabilistic,\ minmax} = 3.653$ and $RMSE_{relevantfeatures,SVR,minmax,pca} = 3.646$) where they are better/lesser. The table shows that the best approaches are obtained in the application of non-linear assembly algorithms. Specifically, the best performance is obtained with the RF algorithm when the non-representative values of each of the input processes are removed and when PCA is used as the dimensional reduction method for the relevant features ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$). This approach (contrary to the best RMSE-XGB and RMSE-SVR solutions shown in blue) gets worse when scaling the inputs (minmax-scaler). This is because, with the exception of the tree-based models, in the rest of the cases, the objective function of the algorithms assumes (wrongly) that each of the inputs follows a normal distribution and therefore, as shown in Table 1, the best approaches are obtained when each of the inputs is monotonically transformed using normalization/scaling techniques such as minmax-scaler. The case of XGBoost is a bit peculiar because being a tree-based boosting algorithm it should not require any scaling. However, when optimizing the objective function using the gradient method, normalization tends to improve the results.

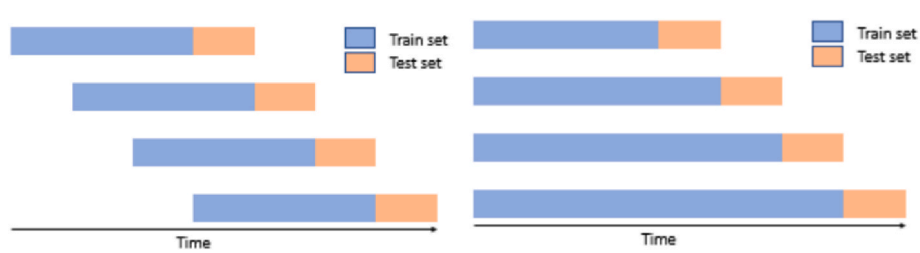

Fig. 3. Adaptive learning strategies.

Table 1
RMSE comparison: pre-processing techniques, relevant inputs and EW-strategy.

Pre-processing techniques				Algorithms			
probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	Ridge	RF	XGB	SVR
✓	✓	✓	✓	4.529	4.345	4.985	4.448
✓	✓	✓		4.462	4.319	4.982	4.451
✓	✓		✓	4.001	3.705	3.808	3.726
✓	✓			5.627	3.638	**3.653**	3.721
✓		✓	✓	4.512	4.399	4.851	4.741
✓		✓		4.462	4.319	4.978	4.744
✓			✓	4.038	3.565	3.728	4.748
✓				16.01	3.639	3.655	4.748
	✓	✓	✓	4.43	4.372	4.649	4.35
	✓	✓		4.352	4.352	4.967	4.357
	✓		✓	**3.905**	3.743	3.897	**3.646**
	✓			12.37	3.583	3.676	3.648
		✓	✓	4.443	4.309	4.554	4.631
		✓		4.352	4.353	4.961	4.634
			✓	3.944	3.653	3.822	4.653
				15.978	3.979	4.077	4.652

Furthermore, in order to quantify the suitability of the selection of the relevant features technique versus the treatment of all process features, we have compared the approaches in both cases. Table 2 shows the approaches of the best combination of pre-processing techniques when performed only on the relevant inputs versus all inputs. The study demonstrates that correct input selection minimizes the value of the RMSE error metric and therefore improves the prediction result in validation ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$ vs $RMSE_{allfeatures,SVR,probabilistic,pca} = 4.836$); it also prevents over-fitting and reduces model complexity.

The analysis also includes, in addition to the previous one on the application of EW-strategy, the comparison with MW-strategy. Table 3 shows the combination of the pre-processing and dimensional reduction techniques for RF, with MW-strategy, when choosing to select the relevant inputs and when choosing to work on all inputs. In this case, better results are obtained when RF is applied on all the features ($RMSE_{allfeatures,RF,minmax,pca} = 4.599$) than when it is only applied on the relevant ones ($RMSE_{relevantfeatures,RF,minmax,pca} = 4.653$). However, these approaches do not improve the values of the error metric obtained with the EW-strategy ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$).

This study shows that the best results are obtained when relevant feature selection techniques, pre-processing techniques (as probabilistic data cleaning and PCA), and RF as the algorithm for the EW as adaptive learning strategy are applied. The best result obtained is $RMSE = 3.565$.

Fig. 4 illustrates the values of the real output and the predicted output, the residual error and the residual histogram. The Figure shows in the first graph, and in blue color, the output value of the flash point temperature (Real) and in orange color, the forecast value (Prediction); the second graph represents the difference between both, the residual being $Residual = Real - Prediction$; and finally, the third graph shows the histogram of the residual of the error. For a model to be considered adequate, it is necessary that the residuals follow a Gaussian distribution with mean value 0 and minimum variance. The histogram does not exactly reproduce the Gaussian curve, but these deviations are usually attributed to the not too large number of residuals. If the number of residuals was larger, one tends to think that the Gaussian representation would be more evident. The graph shows that the estimated regression model maintains equality of variance (no heteroscedasticity) and that it complies with the assumption of normality of the residuals.

Table 3
RMSE comparison: pre-processing techniques, relevant inputs andMW-strategy.

Pre-processing techniques				Algorithms	
probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	RF relevant features	RF all features
✓	✓	✓	✓	5.216	5.026
✓	✓	✓		4.996	4.948
✓	✓		✓	4.743	4.619
✓	✓			4.742	4.804
✓		✓	✓	5.140	5.004
✓		✓		4.996	4.948
✓			✓	4.944	5.070
✓				4.742	4.804
	✓	✓	✓	5.288	4.933
	✓	✓		4.898	4.852
	✓		✓	**4.653**	**4.599**
	✓			4.682	4.812
		✓	✓	5.028	4.938
		✓		4.898	4.851
			✓	4.864	4.996
				5.079	5.012

As the daily typing of the laboratory variable is manual and prone to human errors when the model is daily trained with the arrival of each new laboratory measurement, the error between the output prediction and the real laboratory measurement is checked. If the difference between output and real exceeds a certain threshold, the model does not consider that laboratory measurement. This is the case of three points in Fig. 5 that are identified in the scatter diagram, one yellow point and two green points that stand out from the rest. In this case, the 3 real laboratory measurements are considered as outliers (13 °C, 76 °C y 78 °C).

Fig. 5 shows the scatter plot between the real value of the flash point temperature and the predicted value when the algorithm with the lowest RMSE is used ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$). The points are distributed around the regression line, normally distributed, with a mean value 0.

Table 2
RMSE comparison: pre-processing techniques, all/relevant inputs and EW-strategy.

Features	Pre-processing techniques				Algorithms			
	probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	Ridge	RF	XGB	SVR
Relevant features	✓			✓	4.038	3.565	3.728	4.748
All features	✓			✓	4.888	4.928	5.162	**4.836**

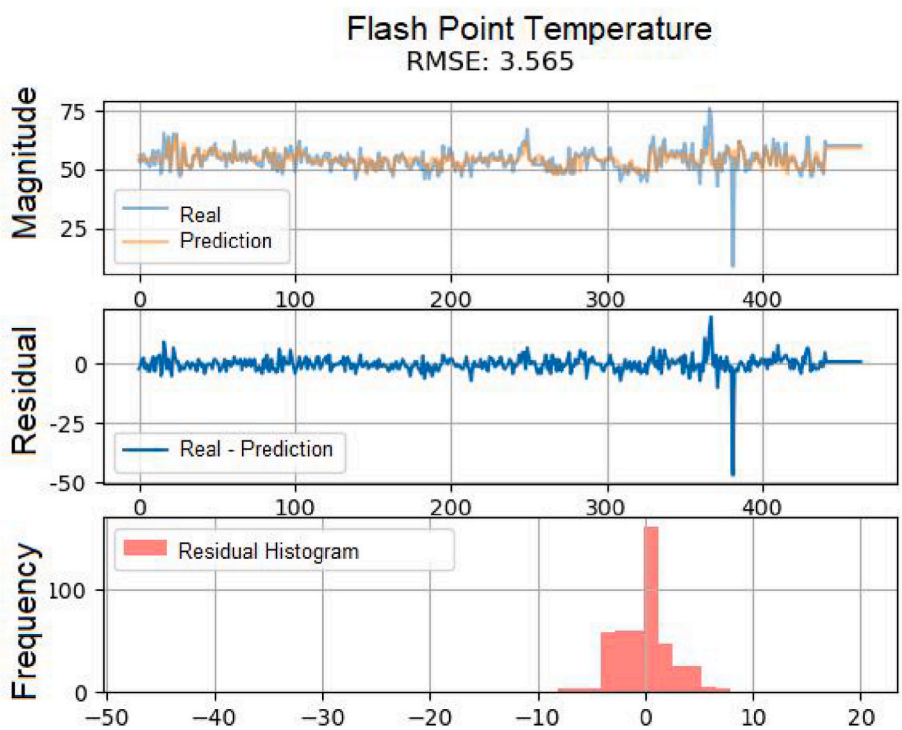

Fig. 4. Flash point temperature Process estimates and its residual signals. **a)** Real flash point temperature and prediction values. **b)** Residual signal. **c)** Residual histogram.

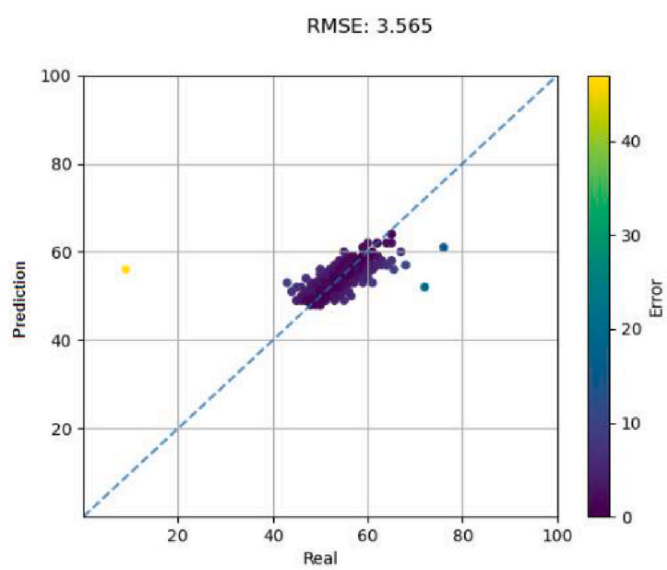

Fig. 5. Scatter plot (Real flash point temperature vs prediction values).

The following conclusions were drawn from the above studies:

- It is observed that a non-adaptive strategy model is not adequate due to the emergence of new behaviors and its inability to adapt to these new behaviors. Therefore, it is concluded that it is necessary to work with an MW-strategy that allows learning new behaviors.
- The relationship between the relevant characteristics and the flash point temperature is multivariable and also evolves over time. In this study, an MW-strategy approach is the best way to generate the process inference model: the difference between EW-strategy and the different ensemble algorithms is minimal ($RMSE_{relevantfeatures,\ Ridge,\ minmax,pca} = 3.905$, $RMSE_{relevantfeatures,XGB,probabilistic,\ minmax} = 3.653$ and $RMSE_{relevantfeatures,SVR,minmax,pca} = 3.646$), obtaining the best result with RF ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$).
- The study shows better results when a selection of the relevant characteristics is made with the help of the refinery technicians than when working on the totality of the variables.
- The results obtained are better in the EW-strategy than in the MW-strategy ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3.565$ vs $RMSE_{allfeatures,RF,minmax,pca} = 4.599$).

2.5. Economic quantification of cost savings

The quality variable flash temperature indicates the quality of the product at the exit of the desulfurization process. Currently, it is obtained on a daily basis in a laboratory process. Once the flash temperature value is known, the operator modifies the rest of the process variables in order to obtain a temperature value that meets the specifications. If the temperature value is higher than the specifications, it impacts an unnecessary overrun for the refinery: a higher quality product is produced that does not result in higher profits, since the selling price is maintained. This higher quality product implies indirect cost overruns in the process, e.g. operating at higher temperatures shortens the lifecycle of the catalysts used in the desulphuration reaction. In this desulfurization process and in general, in the high-volume processes of refineries, the limitation of not having real-time information causes great economic impacts. The tighter the process variables are tuned, the less they deviate from specification quality. The soft sensor provides refinery operators real-time information in order to adjust operating conditions, maximizing the stability of the desulfurization unit and producing diesel to specification.

According to the refinery's estimates and with the current procedure, 50% of the time, despite the delay between taking the samples and obtaining the experimental measurements in the laboratory, realistic information is provided on the temperature of the process in progress. But in 12% of the cases, there is a slight uncertainty in the temperature value information, and in the 38% of the cases, is it not known if the temperature information is sufficiently realistic. When the information

is not sufficiently realistic (38%), it stops generating profits worth 497, 306*$/semester*, and when the information is slightly uncertain (38% + 12%), it stops generating profits worth 654, 350*$/semester*. The soft sensor proposed in this study provides reliable flash temperature measurements 94% of the operating time. Thus, a considerable reduction of the generated losses is estimated, from 497, 306*$/semester* to 29, 838 *$/semester*, in the first scenario where the represented reality is not sufficiently realistic, whereas in the second scenario the reduction goes from 654, 350*$/semester* to 39, 261*$/semester*.

3. Conclusions

Historically, the calculation of flash temperature has been done basis on laboratory measurements or based on molecular information values of pure components. But these approaches have their limitations. This is especially true for real-time monitoring and applications. The present study proposes a complete data-driven procedure definition for real-time flash temperature inference through a novel adaptive soft sensor. Three years of real 10-min sampled data from an important international refinery is used for training and testing the defined procedure. The procedure is analyzed based on two key aspects:

- Soft sensor scheme (based on Machine Learning) capable of learning the non-linear relationship between the easily measured process variables (i. e. temperature, pressure, and hydraulic Flow) and the value of the flash point temperature. The proposed methodology demonstrates its validity to provide flash point temperature in real-time, when there is no laboratory information related to the target variable, through machine learning inference. Different schemes are benchmarking showing that RandomForest outperforms Ridge regression, XGBoost, or SVR.
- Mechanisms to automatically adapt its behavior dynamically, adapting it to the behavioral changes that the process tends to undergo in real situations. It is demonstrated that a non-adaptive strategy model is not adequate due to the emergence of new behaviors and its necessity to adapt to these new behaviors.

The best way to generate this dynamical adaptation capability is obtained using RF scheme by applying probabilistic-data-cleaning and pca-decomposition techniques over the most relevant features when EW approach is implemented (RMSE = 3.565).

The accuracy of the proposed adaptive soft sensor allows us to infer the flash temperature in real-time, and even also can be adapted to work as an anomaly detection tool tracking the difference between the predicted value and the real value, alarming when this difference exceeds a certain threshold. The soft sensor proposed in this study provides reliable flash temperature measurements 94% of the operating time. Thus, a considerable reduction of the generated losses is estimated, from 497, 306*$/semester* to 29, 838*$/semester*, in the first scenario where the represented reality is not sufficiently realistic, whereas in the second scenario the reduction goes from 654, 350*$/semester* to 39, 261 *$/semester*.

Author contributions

Conceptualization, I.M., E.M., S.G-L., I.L-T and L.O; methodology, I. M., and E.M.; software, I.M.; validation, I.M., E.M., I.L-T. and L.O; formal analysis, I.M.; investigation, I.M.; resources, I.L-T and L.O.; writing—original draft preparation, I.M.; writing—review and editing, I.M, E.M, S.G-L, I.L-T and L.O; visualization, I.M.; supervision, S.G-L.; project administration, E.M. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received funding support from the SPRI-Basque Government through the ELKARTEK program (OILTWIN project, ref. KK-2020/00052). We also thank the anonymous reviewers for the careful and positive review enabling us to improve the paper.

References

[1] B. Karun, R. VR, S. Elayidom, Application of fuzzy logic and machine learning techniques to improve inherently safer design in process safety management: a brief study, Process Saf. Prog. (2021).
[2] H. Ahmed, I.F. Tahoun, S. Zakel, Development of decahydronaphthalene reference material for low flash point measurements, Egypt. J. Petrol. 30 (2021) 7–10.
[3] R.W. Prugh, Estimation of flash point temperature, J. Chem. Educ. 50 (1973) A85.
[4] A. Alibakhshi, H. Mirshahvalad, S. Alibakhshi, Prediction of flash points of pure organic compounds: evaluation of the DIPPR database, Process Saf. Environ. Protect. 105 (2017) 127–133.
[5] F. Gharagheizi, A new molecular-based model for prediction of enthalpy of sublimation of pure components, Thermochim. Acta 469 (2008) 8–11.
[6] L.Y. Phoon, A.A. Mustaffa, H. Hashim, R. Mat, A review of flash point prediction models for flammable liquid mixtures, Ind. Eng. Chem. Res. 53 (2014) 12553–12565.
[7] S.M. Santos, D.C. Nascimento, M.C. Costa, A.M. Neto, L.V. Fregolente, Flash point prediction: reviewing empirical models for hydrocarbons, petroleum fraction, biodiesel, and blends, Fuel 263 (2020) 116375.
[8] T.N.G. Borhani, M. Saniedanesh, M. Bagheri, J.S. Lim, QSPR prediction of the hydroxyl radical rate constant of water contaminants, Water Res. 98 (2016) 344–353.
[9] Z. Jiao, H.U. Escobar-Hernandez, T. Parker, Q. Wang, Review of recent developments of quantitative structure-property relationship models on fire and explosion-related properties, Process Saf. Environ. Protect. 129 (2019) 280–290.
[10] Z. Jiao, P. Hu, H. Xu, Q. Wang, Machine learning and deep learning in chemical health and safety: a systematic review of techniques and applications, ACS Chem. Health Saf. 27 (2020) 316–334.
[11] Q. Sun, L. Jiang, M. Li, J. Sun, Assessment on thermal hazards of reactive chemicals in industry: state of the art and perspectives, Prog. Energy Combust. Sci. 78 (2020) 100832.
[12] W.F. McClure, Near-infrared spectroscopy the giant is running strong, Anal. Chem. 66 (1994) 42A–53A.
[13] S. Hu, others, A general framework for incorporating molecular modelling into overall refinery optimisation, Appl. Therm. Eng. 21 (2001) 1331–1348.
[14] S. Park, J.P. Bailey, H.J. Pasman, Q. Wang, M.M. El-Halwagi, Fast, easy-to-use, machine learning-developed models of prediction of flash point, heat of combustion, and lower and upper flammability limits for inherently safer design, Comput. Chem. Eng. 155 (2021) 107524.
[15] D.J.L. Prak, G.R. Simms, M. Hamilton, J.S. Cowart, Impact of low flash point compounds (hydrocarbons containing eight carbon atoms) on the flash point of jet fuel and n-dodecane, Fuel 286 (2021) 119389.
[16] A. International, Standard Test Methods for Flash Point by Pensky-Martens Closed Cup Tester, ASTM International, 2015.
[17] P. Kadlec, R. Grbić, B. Gabrys, Review of adaptation mechanisms for data-driven soft sensors, Comput. Chem. Eng. 35 (2011) 1–24.
[18] H. Albazzaz, X.Z. Wang, Historical data analysis based on plots of independent and parallel coordinates and statistical control limits, J. Process Control 16 (2006) 103–114.
[19] Y. Dote, S.J. Ovaska, Industrial applications of soft computing: a review, Proc. IEEE 89 (2001) 1243–1265.
[20] P. Kadlec, B. Gabrys, S. Strandt, Data-driven soft sensors in the process industry, Comput. Chem. Eng. 33 (2009) 795–814.
[21] I.F. Ilyas, X. Chu, Data Cleaning, Association for Computing Machinery, 2019.
[22] L. Yao, Z. Ge, Moving window adaptive soft sensor for state shifting process based on weighted supervised latent factor analysis, Control Eng. Pract. 61 (2017) 72–80.
[23] D. Markudova, E. Baralis, L. Cagliero, M. Mellia, L. Vassio, E.G. Amparore, R. Loti, L. Salvatori, Heterogeneous Industrial Vehicle Usage Predictions: A Real Case, EDBT/ICDT Workshops, 2019, pp. 3–8.
[24] R. M'chaar, N. Ouerfelli, M.K. Ammar, B. Hafez, M. Singh, A. Messaâdi, H. Elmsellem, H. Arslan, Surface tension and viscosity-temperature dependence and mutual causal correlation in tin-silver alloys, Surface. Interfac. 26 (2021) 101444.
[25] F. Tanzim, E. Kontos, D. White, Generating prediction model of fan width by optimizing paint application process for Electrostatic Rotary Bell atomizer, Result. Eng. 13 (2022) 100302.
[26] S.R. Vadyala, S.N. Betgeri, J.C. Matthews, E. Matthews, A review of physics-based machine learning in civil engineering, Result. Eng. 13 (2022) 100316.

[27] A. Latifi, S. Saeidijam, M. Derakhshandi, M. Heydari, Numerical assessment of Electrokinetic Barrier with coupled flow modeling approach, Result. Eng. 13 (2022) 100325.
[28] A. Neumann, N. Laranjeiro, J. Bernardino, An Analysis of Public REST Web Service APIs, IEEE Transactions on Services Computing, 2018.
[29] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, Elsevier, 2021.
[30] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.
[31] A. Fisher, C. Rudin, F. Dominici, All Models Are Wrong but Many Are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, Using Model Class Reliance, 2018, pp. 237–246, *arXiv preprint arXiv:1801*.01489.
[32] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417.
[33] I.T. Jolliffe, Principal components in regression analysis, in: Principal Component Analysis, Springer, New York, 1986, pp. 129–155.
[34] T. Helleseth, Some results about the cross-correlation function between two maximal linear sequences, Discrete Math. 16 (1976) 209–232.
[35] G. Ditzler, M. Roveri, C. Alippi, R. Polikar, Learning in nonstationary environments: a survey, IEEE Comput. Intell. Mag. 10 (2015) 12–25.
[36] Y. Du, Y. Liang, J. Jiang, R.J. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.
[37] H. Kaneko, K. Funatsu, Smoothing-combined soft sensors for noise reduction and improvement of predictive ability, Ind. Eng. Chem. Res. 54 (2015) 12630–12638.
[38] C. Kneale, S.D. Brown, Small moving window calibration models for soft sensing processes with limited history, Chemometr. Intell. Lab. Syst. 183 (2018) 36–46.
[39] W. Ni, S.K. Tan, W.J. Ng, S.D. Brown, Moving-window GPR for nonlinear dynamic system modeling with dual updating and dual preprocessing, Ind. Eng. Chem. Res. 51 (2012) 6416–6428.
[40] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[41] T. Chen, C. Xgboost Guestrin, A scalable tree boosting system, in: Proceedings of the 22nd acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[42] J. Bartlett, C. Frost, Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables, Ultrasound Obstet. Gynecol.: Off. J. Int. Soc. Ultrasound Obstetr. Gynecol. 31 (2008) 466–475.