Chemical Engineering Science 283 (2024) 119386

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Data-driven prediction of product yields and control framework of hydrocracking unit

Zheyuan Pang [a], Pan Huang [a], Cheng Lian [a,b,*], Chong Peng [c,d,*], Xiangcheng Fang [d], Honglai Liu [a,b]

[a] State Key Laboratory of Chemical Engineering, Shanghai Engineering Research Center of Hierarchical Nanomaterials, and School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China
[b] School of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai 200237, China
[c] State Key Laboratory of Fine Chemicals, School of Chemical Engineering, Dalian University of Technology, Dalian 116024, China
[d] Dalian Research Institute of Petroleum and Petrochemicals, SINOPEC, Dalian 116024, China

ARTICLE INFO

Keywords:
Hydrocracking
Machine learning
Yield prediction
Process control

ABSTRACT

In this study, the relationship between the operating conditions and the product yields and a control framework of the hydrocracking process was developed. The data were collected from a hydrocracking unit in a Chinese refinery. Principal component analysis was used to decrease the number of input variables. Then support vector machine, Gaussian process regression (GPR), and decision tree regression models were developed to establish the relationship above. The best model is GPR, whose Pearson correlation coefficient between the prediction value and the actual value is greater than 0.97 for all the product yields. Shapley additive explanations were performed to interpret the results of the GPR models. A control framework of the hydrocracking unit was then proposed based on the results above. The results show that the machine learning method is a valuable tool for predicting the yield of hydrocracking products, and the control framework proposed helps optimize hydrocracking product yields.

1. Introduction

Hydrocracking is an improved catalytic cracking technology that can inhibit dehydrogenation condensation reaction and reduce coke formation. Fig. 1 is the flowchart of the hydrocracking unit studied in this paper. After mixing with the make-up hydrogen and the recycled hydrogen, the feedstock is fed into three reactors where the hydrocracking reactions happen. The outlet of R1 and R3 are fed in R2. The outlet of R2 is fed into the separator. In the separator, the products are separated with hydrogen. The hydrogen is sent to the HDS tower to recycle after cleaning, while the product is sent to the next separator. In the next separator, the products are separated from each other. The feedstock of hydrocracking is vacuum gas oil and other heavy oil. The products of hydrocracking are fuel gas (FG), light naphtha (LN), propane (PRO), liquefied petroleum gas (LPG), heavy naphtha (HN), diesel fuel (DF), and other light oil. The product yield can be manipulated flexibly by adjusting operating conditions. In China, the composition of the feedstocks of hydrocracking units fluctuates significantly with time and refinery location, which makes the product yield fluctuate remarkably. Hence, the relationship between the operating conditions and the yield of the products, which is hard to acquire by empirical methods, is needed. Therefore, a more accurate and effective method to establish the relationship above as well as a control framework of the unit are required to improve the robustness of the hydrocracking unit and meet the quality of products.

The modeling method is an important and widely used approach to establishing the relationship above. The commonly used models are the mechanism model and the data-driven model (Song et al., 2020). Typical mechanism models include the structured lumping model, the discrete lumping model, the continuous lumping model, and the single-event model (Ancheyta et al., 2005). In the discrete lumping model, components are lumped according to the boiling point, and each lump is considered a pseudo-component (Umana et al., 2016). The continuous lumping model considers the mixture of hydrocarbons as a continuous distribution (such as the boiling point) (Becker et al., 2016b). In the structured lumping model, a vector is used to describe the hydrocarbon molecule, of which the elements represent the structural features sufficient to build the molecule (Basak et al., 2004). First proposed by

* Corresponding authors.
E-mail addresses: liancheng@ecust.edu.cn (C. Lian), pengchong@dlut.edu.cn (C. Peng).

https://doi.org/10.1016/j.ces.2023.119386
Received 24 July 2023; Received in revised form 17 September 2023; Accepted 9 October 2023
Available online 12 October 2023

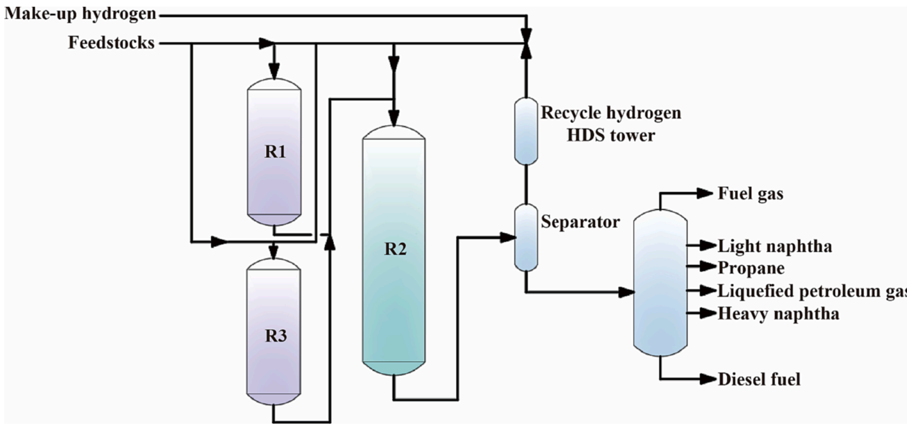

Fig. 1. The flow chart of the hydrocracking process.

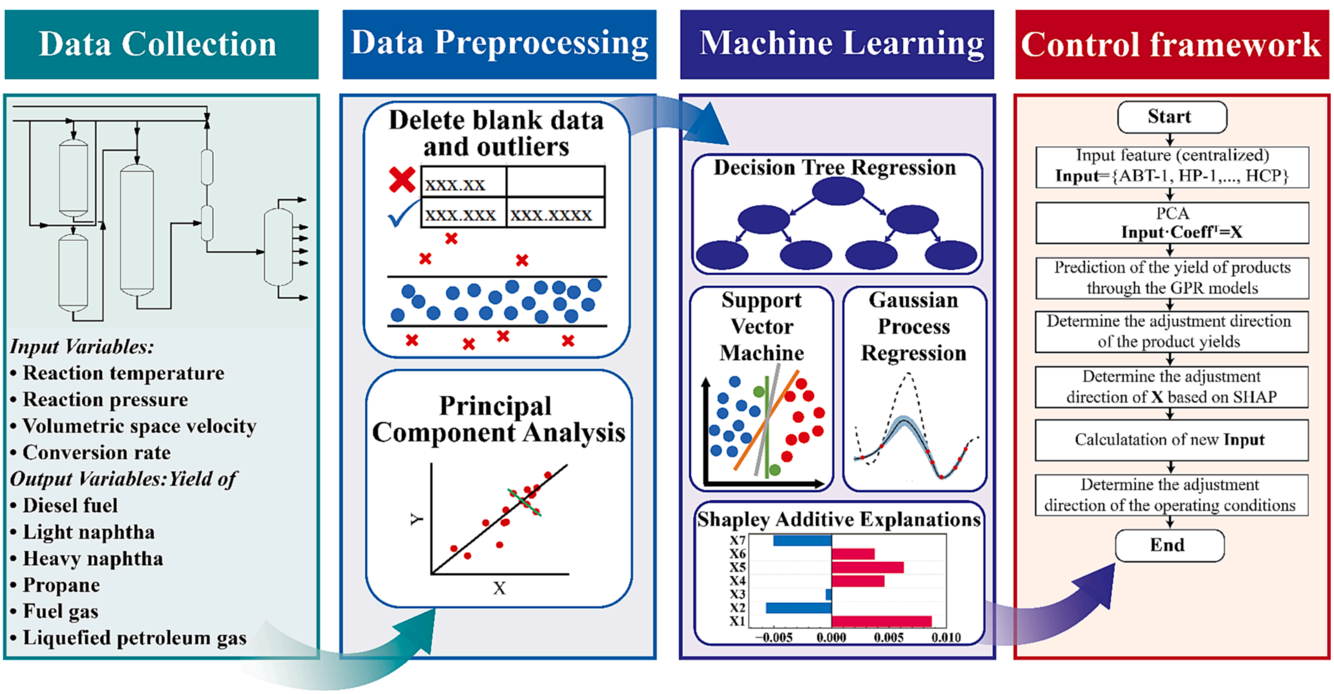

Fig. 2. Framework of data-driven prediction and control framework of hydrocracking product yields.

Froment et al.(Baltanas and Froment, 1985; Baltanas et al., 1989; Froment, 1987; Hillewaert et al., 1988; Vynckier and Froment, 1991), the single-event model restructures the feed into individual molecules and then builds enormous reaction networks with thousands of possible reaction path (Becker et al., 2016a). In the lumping method, the estimation and adjustment of the parameters need enormous calculation and time, which makes it difficult to find optimal operating conditions (Elizalde et al., 2009; Lababidi and AlHumaidan, 2011). Single-event and other microkinetic models require an understanding of chemical kinetics and feed composition, which limits the application of the model (Ancheyta et al., 2005; Becker et al., 2016a; Elizalde and Ancheyta, 2011; Iplik et al., 2020).

The data-driven model aims to provide less model complexity and quicker prediction, which makes the model readily available online (Iplik et al., 2020). In data-driven models, machine learning algorithms are commonly used. The machine learning models that are used extensively include decision tree regression (DTR), support vector machine (SVM), and Gaussian process regression (GPR). SVM is a supervised learning method with an excellent theoretical foundation and little need for data amount (Wu et al., 2008). Sharifi (Sharifi et al., 2019) used the data from the Tehran oil refinery in Iran and developed a SVM model to establish the relationship between the operating conditions and the yields of hydrocracking products. Based on Bayesian probability theory, GPR is a useful nonlinear regression model (Deringer et al., 2021). Iapteff (Iapteff et al., 2021) established the GPR model to study the hydrocracking process and predict diesel density. Fadzil (Fadzil et al., 2021) developed decision tree regression and other models to predict the base oil product's kinematic viscosity using the feedstock and process conditions. However, the existing machine learning models of the hydrocracking process rarely use the data of the Chinese hydrocracking units as input data. Additionally, the composition of the feedstocks used in Chinese refineries is not as stable as that used in refineries in other countries. Therefore, the operating conditions used in Chinese refineries fluctuate remarkably to meet stable product yields and may have a wider range compared with those used in refineries in other countries. As a result, the machine learning models that are suitable for the hydrocracking units in other countries are trained with operating conditions that have a narrower range than those used in Chinese refineries. Therefore, the models above may not be suitable for the Chinese hydrocracking process. Hence, machine learning models suitable for the

Table 1
All variables used in this paper and the abbreviations of input variables.

	Name	Abbreviation
Input variables	Adiabatic reactor temperature −1	*ABT-1*
	The space velocity of hydrocracking pretreatment-1	*HP-1*
	Adiabatic reactor temperature −2	*ABT-2*
	The space velocity of hydrocracking pretreatment −2	*HP-2*
	Adiabatic reactor temperature −3	*ABT-3*
	The space velocity of hydrocracking catalyst	*HC*
	The space velocity of hydrocracking pretreatment −3	*HP-3*
	VGO (R1)	*VGO (R1)*
	VGO (R3 feed)	*VGO (R3)*
	Conversion rate	*CR*
	Hydrogen consumption	*HCO*
Output variables	DF yield	–
	FG yield	–
	HN yield	–
	LN yield	–
	LPG yield	–
	Pro yield	–

Chinese hydrocracking units and a control framework are urgently needed to develop the Chinese hydrocracking process.

Fig. 2 shows the framework of data-driven prediction and control framework of hydrocracking product yields. First, the data were acquired from the hydrocracking unit of a Chinese refinery. The data was cleaned by deleting the outliers using the 3-σ method and the blank data. Then principal component analysis (PCA) was performed to decrease the number of input variables. Three machine learning models (SVM, GPR, and DTR) were then developed to establish the relationship between the operating conditions and the yield of the products. Shapley additive explanations (SHAP) were performed to interpret the results of the best model of the three models above. A control framework of the hydrocracking unit was then proposed based on PCA, the best machine learning model above, and SHAP.

2. Methods

2.1. Data collection

The adjustment of the hydrocracking process is generally accomplished by the reaction temperature, reaction pressure, volumetric space velocity, and hydrogen-to-oil volume ratio. The reaction temperature increases from the inlet to the outlet in the industrial adiabatic reactor, and the average reaction temperature of the entire reactor is used to characterize the temperature of the whole reactor. The reactor inlet pressure generally characterizes the reaction pressure. The volume space velocity is the volume of feedstock processed per unit volume of catalyst. It is usually characterized by the feedstock processing capacity as the catalyst loading in the industrial reactor is fixed. The hydrogen-to-oil volume ratio refers to the volume ratio of the amount of circulating hydrogen and fresh hydrogen to the fresh feed. Additionally, the conversion rate is also an important feature. Conversion rate is the ratio of heavy oil to light oil and is the primary product distribution control method in hydrocracking units.

In summary, eleven important features (the adiabatic reactor temperature measured at three sites, the space velocity of commercial hydrocracking pretreatment catalyst measured at three locations, hydrogen consumption, the space velocity of commercial hydrocracking catalyst, VGO flow rate in two reactors, and conversion rate) are used as input variables. The yields of six products are used as output variables.

Table 1 lists all the variables used in this paper and the abbreviations of the input variables.

2.2. Features preprocessing

2.2.1. Kaiser-Meyer-Olkin (KMO) test

The KMO test was proposed by Kaiser (Kaiser, 1970), and could judge whether the input variables matrix is suitable for PCA. For j, the measure of sampling adequacy can be calculated by (Dziuban and Shirkey, 1974):

$$KMO = \frac{\sum_{k \neq j} r_{jk}^2}{\sum_{k \neq j} r_{jk}^2 + \sum_{k \neq j} q_{jk}^2} \tag{1}$$

where q is the square of the off-diagonal elements of the anti-image correlation matrix $SR^{-1}S$ and r is the square of the diagonal elements of the original correlations.

Table S1 shows the relationship between the KMO result and the suitability for PCA.

2.2.2. PCA

PCA is performed to reduce the dimensionality of the input variables matrix. PCA was first proposed by Hotelling (Hotelling, 1933) and is likely the most popular multivariate statistical technique (Abdi and Williams, 2010). If the input matrix is $N \times d$ (N is the number of data and d is number of feature), PCA could reduce the dimensionality of the matrix to $N \times k (k \leq d)$ without lose of information, and the k features are called the principal components. The principal components are the linear combination of the original features. Generally, k should make the cumulative contribution greater than 0.9.

2.2.3. Data preprocessing

396 pieces of data from January 2019 to January 2020 were collected from a Chinese refinery. The operating data of the reactor and other towers are acquired by the distributed control system (DCS), and the properties of the products are obtained by analyzing the sample of the corresponding product at a specific time daily.

During production, missing data exists because of equipment damage, human negligence, or parking. 377 data remain after deleting the blank data and cleaning the data by the 3-σ method. KMO test and PCA are then performed. Afterward, the dataset is randomly divided into the test set and training set (2:8).

2.3. Machine learning methods

Three machine learning methods were used in this paper.

SVM is an effective machine learning method based on the structural risk minimization principle, VC dimension theory, and statistical learning theory (Wu et al., 2008). It was developed by Vapnik and Cortes (Vapnik, 1999).

SVM regression projects training samples onto a high-dimensional plane to find a suitable hyperplane to divide the training samples (Cortes and Vapnik, 1995). The objective optimization function of SVM is (Li et al., 2022):

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_\varepsilon(f(x_i) - y_i) \tag{2}$$

where C is the regularization constant, w is the normal vector that determines the direction of the hyperplane, represents the maximum margin for dividing the hyperplane, and l_ε is the insensitive loss function.

Developed by Rasmussen and Williams (Williams and Rasmussen, 1995), GPR is a non-parametric supervised machine learning model that could generate results in Gaussian distribution. Compared with other machine learning methods, GPR has the advantages of easy implementation, strong generalization ability and fewer adjustable

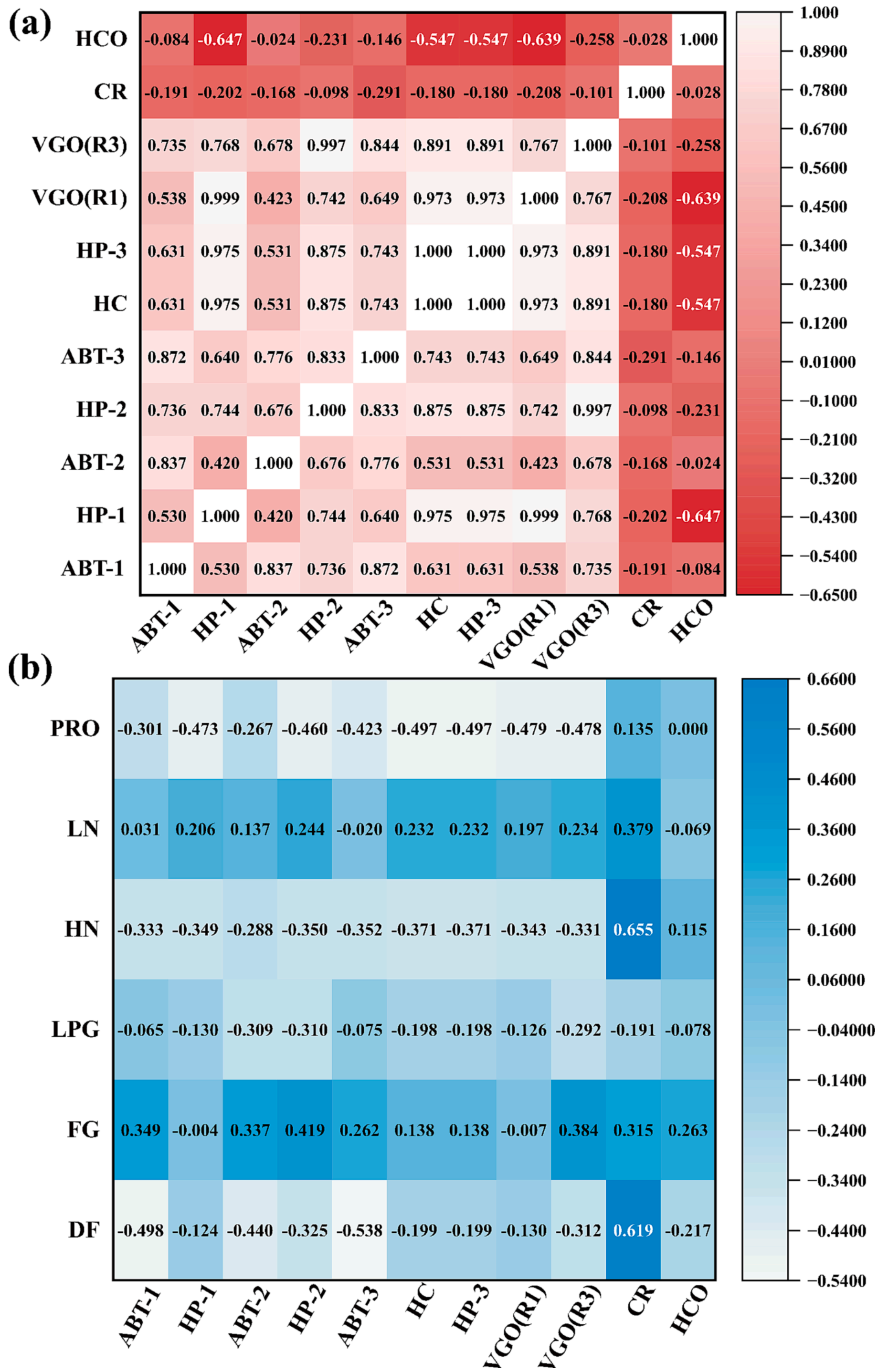

Fig. 3. Correlation coefficient heatmap: (a) input variables with input variables, (b) input variables with output variables.

parameters, and the result of GPR is of probability significance. The Gaussian process is as follows (Xing et al., 2023):

$$f(x) \sim GP(m(x), kf(x, x')) \tag{3}$$

where $m(x)$ is the mean function and $kf(x, x')$ is the covariance function.

Based on the binary tree structure, the decision tree is a non-parametric supervised machine learning method that has wide applications for both classification and regression tasks. The decision tree model has a clear structure and is simple to understand (Sun et al., 2022).

A decision tree model consists of three different nodes (Balogun and Tella, 2022). The root node is the first node of the tree. The interior node is split from the root node and represents the data feature of the model and the rules of decision. The leaf node is the result of the decision.

2.4. Machine learning process

PCA is performed using MATLAB to reduce the number of features. Then the dataset is randomly divided into the test set and training set (2:8).

SVM, GPR, and DTR are completed using MATLAB R2021a. All eligible hyperparameters (as shown in Tables S2–S4) are optimized

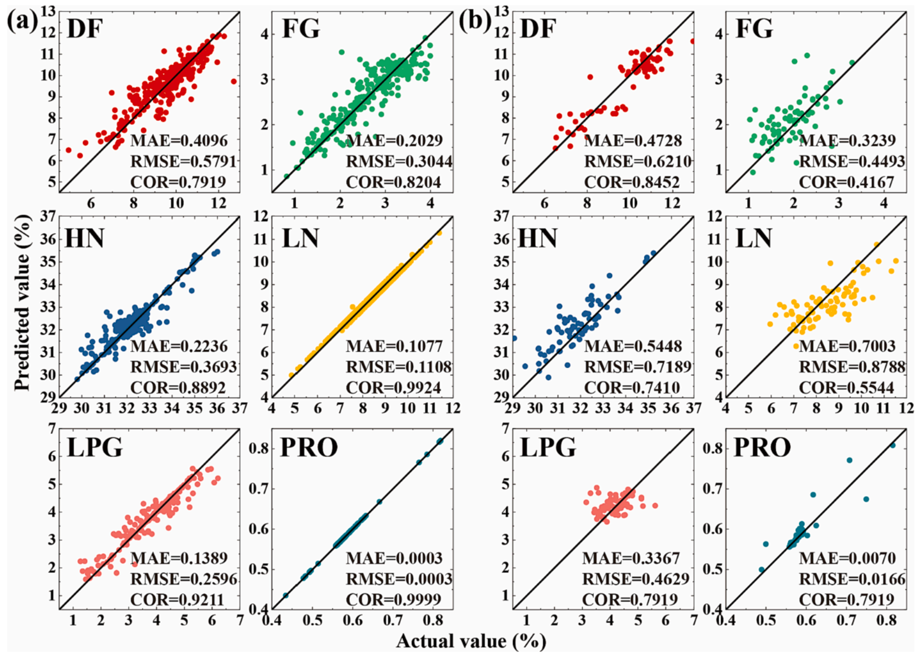

Fig. 4. results of SVM: (a) on the training set; (b) on the test set.

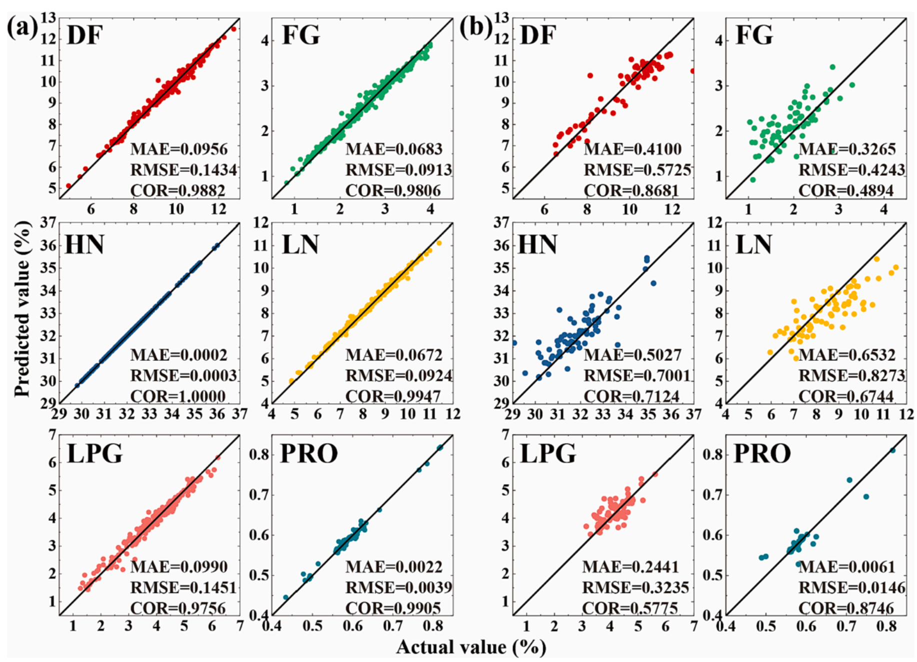

Fig. 5. results of GPR: (a) on the training set; (b) on the test set.

using the grid search method for SVM, GPR, and DTR models. Additionally, 10-fold cross-validation is applied when establishing the models. For SVM and GPR, standardization of data is a hyperparameter that can be optimized among “True” and “False”. When the flag is “True”, the input data would be standardized using the algorithm that is shown in supporting information. The input data is not standardized when training DTR. However, standardization or not has little impact on the results of DTR (Lakshmi et al., 2016; Shreyas et al., 2016).

Pearson correlation coefficient (COR) is calculated to plot a heatmap and evaluate the model performance. Additionally, mean absolute error (MAE) and root mean square error (RMSE) are used to assess the model performance.

The formula of COR, MAE, and RMSE is shown in supporting information.

2.5. SHAP

To establish a control framework, a deep understanding of the influence of each input variable on the output variables is needed, which is difficult in data-driven models. Hence, a powerful and reliable tool is needed to interpret the process inside the machine learning models.

Established by Lundberg and Lee (Lundberg and Lee, 2017), the SHAP method could interpret the results of the black box models and help researchers who have no knowledge about machine learning to

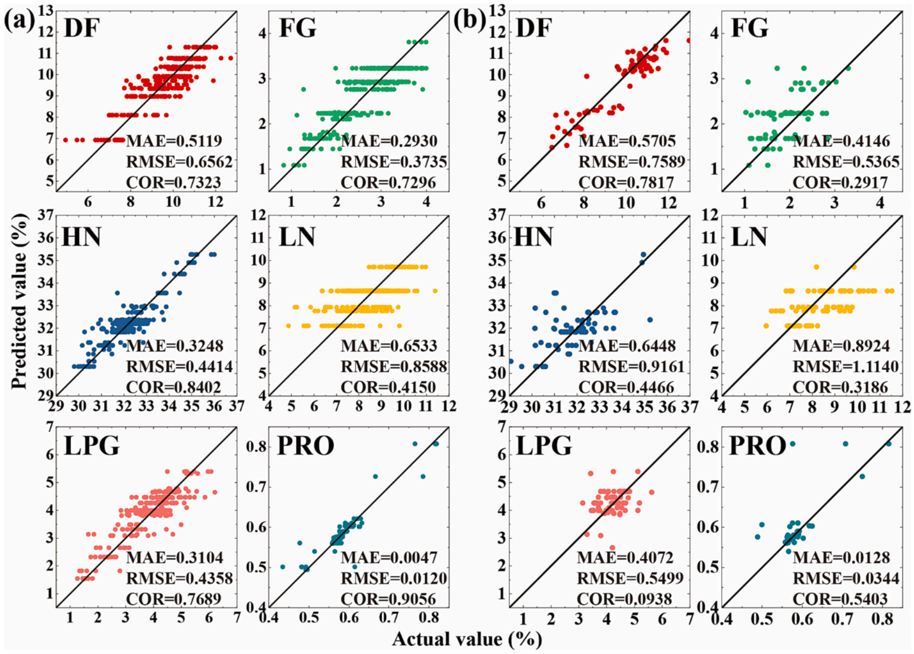

Fig. 6. results of DTR: (a) on the training set; (b) on the test set.

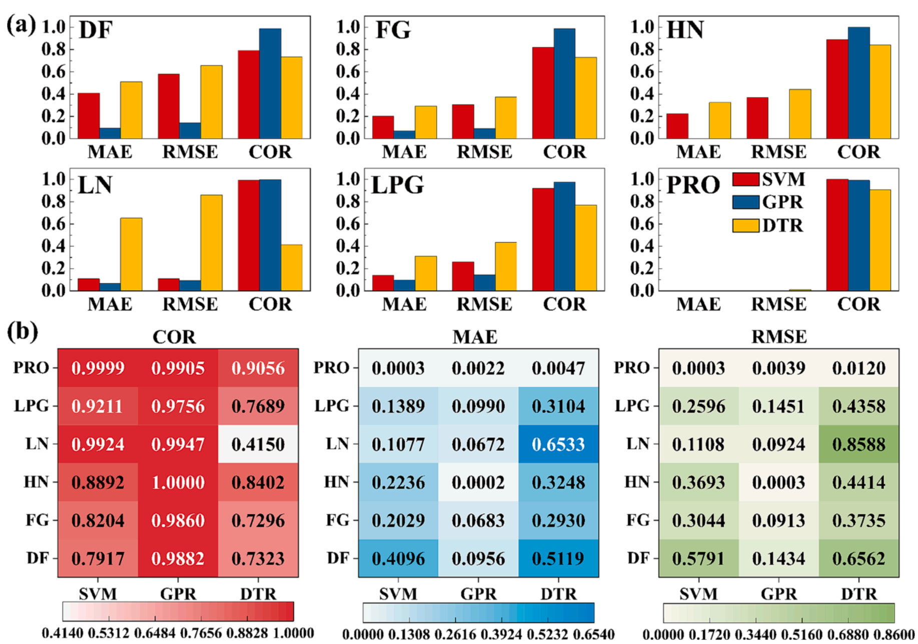

Fig. 7. Comparison of the three models on the training set: (a) histogram, (b) heatmap.

understand the relationships established by the models. SHAP has been successfully applied to interpret machine learning models in the petrochemical and coal chemical field without a benchmark sensitivity analysis method (Chakkingal et al., 2022; Steurtewagen and Van den Poel, 2021).

The key step of the SHAP method is the calculation of the *Shapley values* in cooperative/coalitional game theory (Jas and Dodagoudar, 2023). For a cooperative game with M players, the Shapley value of player j can be calculated as follows (Aas et al., 2021):

$$\phi_j(v) = \phi_j = \sum_{S\subseteq M\setminus\{j\}} \frac{|S|!(M-|S|-1)!}{M!}(v(S\cup\{j\}) - v(s)),\ \mathrm{j}\ =\ 1,\ .\ .\ .\ ,\ \mathrm{M}, \tag{4}$$

where M is the number of players, $S \subseteq M = \{1, ..., M\}$ is a subset containing |S| players, $v(S)$ is a contribution function. This contribution function maps subsets of players to the real number. It is also called the contribution of coalition S.

To explain the difference between the global average prediction and the prediction value $y^* = f(x^*)$ of a machine learning model $f(x)$ that is trained using a set $\{\mathrm{y^i}, \mathrm{x^i}\}_{\mathrm{i}=1,\cdots,\mathrm{n_{train}}}$ whose size is $\mathrm{n_{train}}$, $f(x^*)$ need to be presented using Shapley values (Aas et al., 2021):

$$f(x^*) = \phi_0 + \sum_{j=1}^{M} \phi_j^* \tag{5}$$

where $\phi_0 = \mathrm{E}[\mathrm{f(x)}]$ and ϕ_j^* is the ϕ_j for the prediction $x = x^*$.

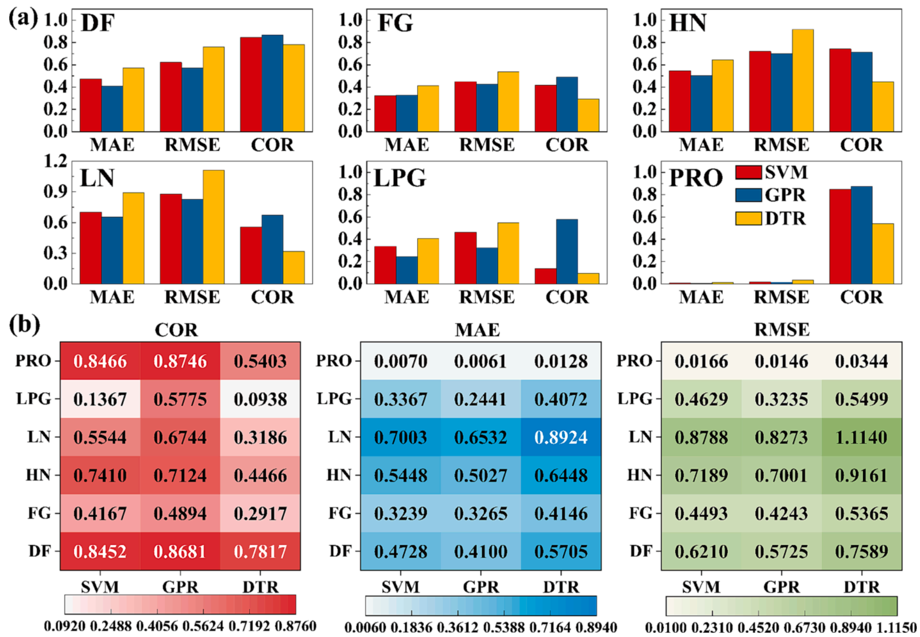

Fig. 8. Comparison of the three models on the test set: (a) histogram, (b) heatmap.

After the calculation, SHAP can explain the global average prediction and consider the linear and nonlinear interactions based on the local interpretation of each record (Chehreh Chelgani et al., 2023).

The SHAP analysis is completed through MATLAB.

3. Result and discussion

3.1. Statistical analysis between input variables and output variables

Fig. 3 shows the correlation coefficient heat map. According to Fig. 3 (a), the correlation coefficient of some input variables is high (such as HP-1 and HC, HP-3, VGO (R1)). The correlation coefficient between HP-1 and HC, HP-3 is greater than 0.97, which may be caused by the order of the process. The feedstock of HC and HP-3 is HP-1, which causes a strong linear relationship between them. HP-1 and VGO (R1) provide two reactants for the hydrocracking reaction. Considering factors such as economy, the consumption of HP-1 and VGO (R1) should meet the stoichiometric number of the reaction equation. Therefore, the correlation coefficient between HP-1 and VGO (R1) is greater than 0.9. Therefore, it is reasonable to combine multiple input variables with strong linear correlation into one input variable and reduce the dimensionality of the input variables matrix.

According to Fig. 3(b), the correlation coefficient between input and output variables is not high enough. Therefore, the nonlinear relationship between input and output variables is needed. The result of the KMO test is 0.8214, which means the use of PCA is meritorious. Then the PCA is implemented, and the dimensionality of the input variables matrix is reduced from 11 to 7.

3.2. Results of machine learning

Fig. 4 shows the results of the SVM models. According to Fig. 4(a), the performance of the SVM models of the yield of LN and PRO is excellent on the training set but relatively bad for other products. The order of the six products from light to heavy is FG, LPG, PRO, LN, HN and DF. Therefore, the SVM model is more suitable for middle products rather than those that are too light or too heavy in this study.

Fig. 5 shows the results of the GPR models. The GPR models show a

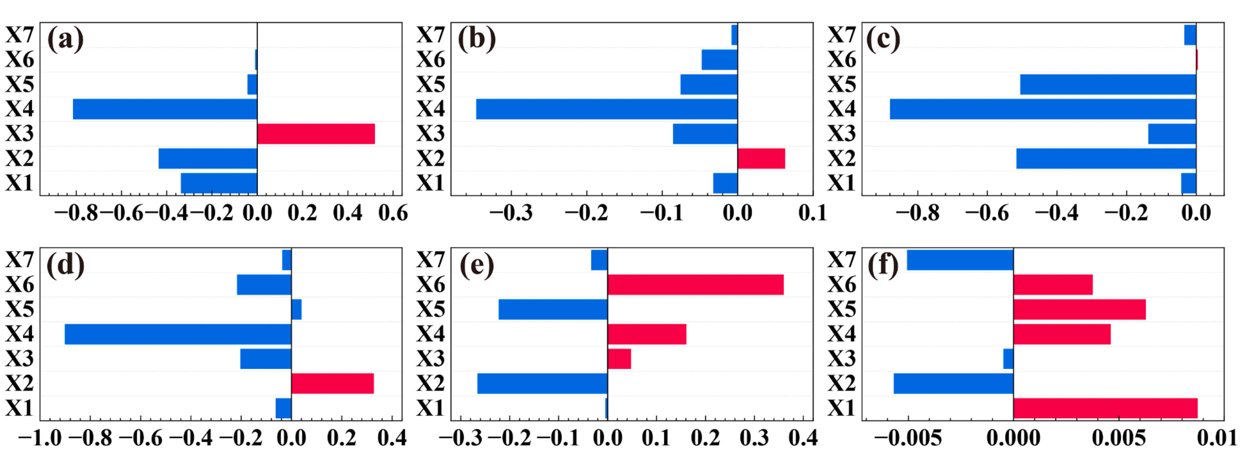

Fig. 9. SHAP values of the first day after the maintenance of the hydrocracking unit: (a) the yield of DF, (b) the yield of FG, (c) the yield of HN, (d) the yield of LN, (e) the yield of LPG, (f) the yield of PRO.

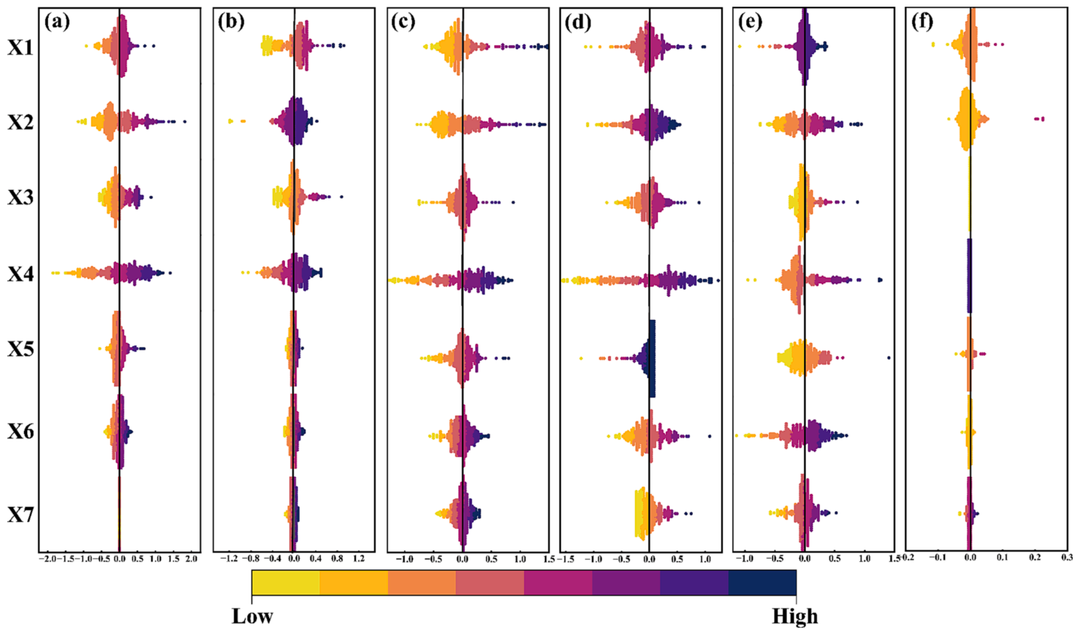

Fig. 10. SHAP values on the whole training set: (a) the yield of DF, (b) the yield of FG, (c) the yield of HN, (d) the yield of LN, (e) the yield of LPG, (f) the yield of PRO.

marvelous performance of all product yields on the training set according to Fig. 5(a). In contrast, the DTR models show poor performance both on the training set and the test set as shown in Fig. 6.

Fig. 7 and Fig. 8 show the COR, MAE and RMSE comparison histogram and heatmap of the three models on the training set and test set, respectively. The GPR models show a better performance than the other two models, and the DTR method is not suitable for the hydrocracking process studied in this paper.

The performance differs significantly when the different output variables are used in the same machine learning method. The nonlinear relationship between the input variables and different output variables may be different. One machine learning model may be suitable to describe some nonlinear relationships but unsuitable to describe others.

The hyperparameter optimization results are shown in Tables S5–S7.

3.3. SHAP analysis and the control framework of the hydrocracking unit

3.3.1. SHAP analysis

SHAP analysis is performed on the training set based on the GPR models. The hydrocracking unit studied in this paper was maintained from July 11 to 21, 2019. Fig. 9 shows the SHAP values on July 22, 2019, the first day after maintenance. The main negative impact on the yield of DF, FG, HN and LN comes from X4, which indicates the four product yields could be controlled simultaneously in the same direction. Additionally, X1 has a great positive impact on the yield of PRO while having little impact on the yield of other products, which indicates the adjustment of X1 could control the yield of PRO with little interruption to other products. A similar relationship exists between X6 and the yield of LPG.

Fig. 10 shows the summary of the local SHAP value on the whole training set. X4 has a great impact on the yield of DF, FG, HN and LN, which agrees with the results in Fig. 9. However, different from Fig. 9, X1 has a great negative impact on the yield of HN and LN on average.

3.3.2. The control framework of the hydrocracking unit

The control framework is developed based on the results of PCA, GPR model and SHAP. The relationship between the original input variables and the results of PCA is as follows:

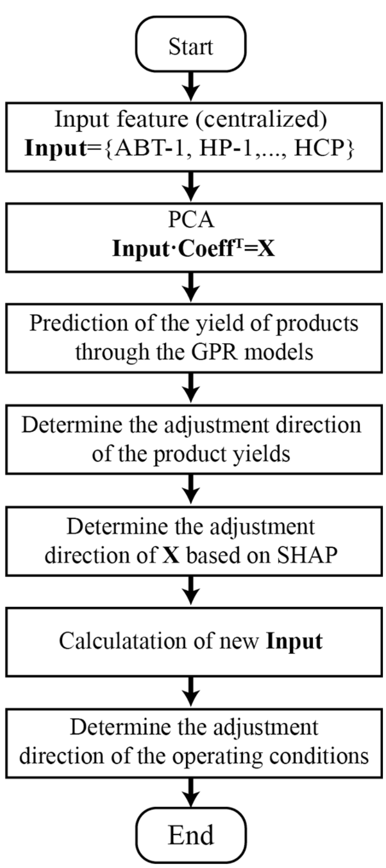

Fig. 11. The control framework of the hydrocracking unit.

$$\textbf{\textit{Input}} \bullet \textbf{\textit{Coeff}}^T = \textbf{\textit{X}}$$

$$\textbf{\textit{Coeff}} = \begin{vmatrix} -0.034 & 0.1598 & 0.3321 & -0.0899 & -0.4072 & 0.141 & 0.818 \\ -0.0045 & 0.0053 & -0.004 & 0.0008 & -0.0005 & 0.0013 & -0.0004 \\ -0.0233 & 0.1411 & 0.3615 & -0.0764 & -0.3378 & 0.71 & -0.4742 \\ -0.0015 & 0.0046 & 0.0054 & 0.0042 & 0.01 & 0.0031 & 0.0034 \\ -0.0809 & 0.3214 & 0.5529 & -0.2474 & -0.1619 & -0.6436 & -0.2874 \\ -0.0051 & 0.0081 & 0.0003 & 0.0036 & 0.0067 & 0.0033 & 0.002 \\ -.00376 & 0.0597 & 0.002 & 0.027 & 0.0498 & 0.0247 & 0.015 \\ -0.541 & 0.6577 & -0.4902 & 0.056 & -0.166 & 0.008 & -0.0301 \\ -0.1212 & 0.3694 & 0.408 & 0.3149 & 0.7313 & 0.1825 & 0.1258 \\ 0.0084 & -0.0781 & 0.112 & 0.90545 & -0.357 & -0.1666 & -0.0793 \\ 0.8264 & 0.5289 & -0.1842 & 0.0448 & -0.0376 & -0.0024 & -0.0076 \end{vmatrix} \tag{6}$$

$$\textbf{\textit{Input}} = |ABT-1 \quad HP-1 \quad ABT-2 \quad HP-2 \quad ABT-3 \quad HC \quad HP-3 \quad VGO(R1) \quad VGO(R3) \quad CR \quad HCO|$$

$$\textbf{\textit{X}} = |X1 \quad X2 \quad X3 \quad X4 \quad X5 \quad X6 \quad X7|$$

The control framework of the hydrocracking unit is shown in Fig. 11. Based on Eq. (4) and the GPR models, the yield of the products can be predicted. To increase or decrease the yield of single or multiple products, the adjustment direction of matrix *X* is determined based on the results of SHAP. Then the adjustment direction of the matrix *Input* is determined based on Eq. (4).

An example of the adjustment of the yield of LPG is presented for better understanding. The centralized *Input* matrix on July 22, 2019, the first day after the maintenance of the hydrocracking unit is:

$$\textbf{Input} = |-0.4371 \quad 0.0153 \quad -1.0976 \quad 0.0116 \quad 1.485 \quad 0.0221 \quad 0.1637 \quad 1.2287 \quad 0.6716 \quad -4.1787 \quad 13.4594|$$

According to Eq. (4), *X* is:

$$\textbf{X} = |10.2551 \quad 8.7567 \quad -2.9956 \quad -3.1405 \quad 1.5899 \quad -0.9957 \quad 0.0155|$$

Based on the GPR model, the yield of LPG is predicted as 4.0183 %. The SHAP values on this point are shown in Fig. 9 (e). To increase the yield of LPG, *X2* should be decreased. Therefore, *X* is adjusted to *X'*:

$$\textbf{X}' = |10.2551 \quad 6 \quad -2.9956 \quad -3.1405 \quad 1.5899 \quad -0.9957 \quad 0.0155|$$

Based on *X'*, the yield of LPG is predicted as 4.1253 %. And the *Input'* is:

$$\textbf{Input}' = |-0.8776 \quad 0 \quad -1.4865 \quad 0 \quad 0.5991 \quad 0 \quad 0 \quad -0.5845 \quad -0.3193 \quad -3.9635 \quad 12.0014|$$

The adjustment direction of the operating condition is determined by the comparison between *Input* and *Input'*. For example, the adiabatic reactor temperature measured at three sites should be decreased, which could be realized by reducing fuel consumption.

4. Conclusion

COR of the input and output variables is calculated to study the linear correlation between them. COR of the input variables shows a strong linear correlation between them, which indicates that it is reasonable to reduce the dimensionality of the input variable matrix. COR between the input and output variables shows a weak linear correlation between them between the variables, which indicates the necessity of using nonlinear relationship tools. KMO test is performed to determine whether the input variables matrix is suitable for PCA. The KMO value is 0.8214, which indicates the input variables matrix is meritorious for PCA. Then PCA is performed to reduce the number of input variables.

Based on the result of PCA, SVM, GPR, and DRT models are developed to establish the relationship between the operating conditions and the yield of the products. By the comparison of MAE, RMSE and COR, the GPR models are identified as the best models, whose COR of the yield of heavy naphtha and light naphtha are 1.0000 and 0.9860 on the training set, respectively. Additionally, the DTR models are unsuitable for the hydrocracking unit studied in this paper. SHAP analysis is then performed based on the GPR models. On the first day after the maintenance of the hydrocracking unit, X4 shows a similar negative impact on the yield of diesel fuel, fuel gas, heavy naphtha and light naphtha, which indicates these four product yields could be controlled simultaneously in the same direction. The SHAP values on the whole training set agree with the result above. A control framework of the hydrocracking unit is then proposed based on PCA, the GPR models, and SHAP.

More accurate models can be developed through the expanding of the dataset and the practice of the proposed control framework can be performed in future work.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was sponsored by the National Key Research and Development Program of China (No. 2019YFC1906702), the National Natural Science Foundation of China (No. 22122807 and No. 22378038), the Fundamental Research Funds for the Central Universities (No. 2022ZFJH04 and JKJ01231806), the State Key Laboratory of Clean Energy Utilization (Open Fund Project No. ZJUCEU2021005) and the National Center for International Research on Intelligent Nano-Materials and Detection Technology in Environmental Protection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ces.2023.119386.

References

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif. Intell. 298, 103502.
Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433–459.
Ancheyta, J., Sánchez, S., Rodríguez, M.A., 2005. Kinetic modeling of hydrocracking of heavy oil fractions: A review. Catal. Today 109, 76–92.
Balogun, A.-L., Tella, A., 2022. Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. Chemosphere 299, 134250.
Baltanas, M.A., Froment, G.F., 1985. Computer generation of reaction networks and calculation of product distributions in the hydroisomerization and hydrocracking of paraffins on Pt-containing bifunctional catalysts. Comput. Chem. Eng. 9, 71–81.
Baltanas, M.A., Van Raemdonck, K.K., Froment, G.F., Mohedas, S.R., 1989. Fundamental kinetic modeling of hydroisomerization and hydrocracking on noble metal-loaded faujasites. 1. Rate parameters for hydroisomerization. Ind. Eng. Chem. Res. 28, 899–910.
Basak, K., Sau, M., Manna, U., Verma, R.P., 2004. Industrial hydrocracker model based on novel continuum lumping approach for optimization in petroleum refinery. Catal. Today 98, 253–264.
Becker, P.J., Celse, B., Guillaume, D., Costa, V., Bertier, L., Guillon, E., Pirngruber, G., 2016a. A continuous lumping model for hydrocracking on a zeolite catalysts: model development and parameter identification. Fuel 164, 73–82.
Becker, P.J., Serrand, N., Celse, B., Guillaume, D., Dulot, H., 2016b. Comparing hydrocracking models: Continuous lumping vs. single events. Fuel 165, 306–315.
Chakkingal, A., Janssens, P., Poissonnier, J., Barrios, A.J., Virginie, M., Khodakov, A.Y., Thybaut, J.W., 2022. Machine learning based interpretation of microkinetic data: a Fischer-Tropsch synthesis case study. React. Chem. Eng. 7, 101–110.
Chehreh Chelgani, S., Nasiri, H., Tohry, A., Heidari, H.R., 2023. Modeling industrial hydrocyclone operational variables by SHAP-CatBoost – A "conscious lab" approach. Powder Technol. 420, 118416.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Deringer, V.L., Bartók, A.P., Bernstein, N., Wilkins, D.M., Ceriotti, M., Csányi, G., 2021. Gaussian Process Regression for Materials and Molecules. Chem. Rev. 121, 10073–10141.
Dziuban, C.D., Shirkey, E.C., 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychol. Bull. 81, 358–361.
Elizalde, I., Ancheyta, J., 2011. On the detailed solution and application of the continuous kinetic lumping modeling to hydrocracking of heavy oils. Fuel 90, 3542–3550.
Elizalde, I., Rodríguez, M.A., Ancheyta, J., 2009. Application of continuous kinetic lumping modeling to moderate hydrocracking of heavy oil. Appl. Catal. A 365, 237–242.
Fadzil, M.A.M., Zabiri, H., Razali, A.A., Basar, J., Syamzari Rafeen, M., 2021. Base Oil Process Modelling Using Machine Learning. Energies 14, 6527.
Froment, G.F., 1987. Kinetics of the hydroisomerization and hydrocracking of paraffins on a platinum containing bifunctional Y-zeolite. Catal. Today 1, 455–473.
Hillewaert, L.P., Dierickx, J.L., Froment, G.F., 1988. Computer generation of reaction schemes and rate equations for thermal cracking. AIChE J 34, 17–24.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417.
Iapteff, L., Jacques, J., Rolland, M., Celse, B., 2021. Reducing the Number of Experiments Required for Modelling the Hydrocracking Process with Kriging Through Bayesian Transfer Learning. J. R. Stat. Soc. Ser. C. Appl. Stat. 70, 1344–1364.
Iplik, E., Aslanidou, I., Kyprianidis, K., 2020. Hydrocracking: A Perspective towards Digitalization. Sustainability 12.
Jas, K., Dodagoudar, G.R., 2023. Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP. Soil Dyn. Earthq. Eng. 165, 107662.
Kaiser, H.F., 1970. A second generation little jiffy.
Lababidi, H.M.S., AlHumaidan, F.S., 2011. Modeling the Hydrocracking Kinetics of Atmospheric Residue in Hydrotreating Processes by the Continuous Lumping Approach. Energy Fuel 25, 1939–1949.
Lakshmi, B.N., Indumathi, T.S., Ravi, N., 2016. A Study on C.5 Decision Tree Classification Algorithm for Risk Predictions During Pregnancy. Procedia Technol. 24, 1542–1549.
Li, J., Zhu, D., Li, C., 2022. Comparative analysis of BPNN, SVR, LSTM, Random Forest, and LSTM-SVR for conditional simulation of non-Gaussian measured fluctuating wind pressures. Mech. Syst. Sig. Process. 178, 109285.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.
Sharifi, K., Safiri, A., Asl, M.H., Adib, H., Nonahal, B., 2019. Development of a SVM model for Prediction of Hydrocracking Product Yields. Pet. Chem. 59, 233–238.
Shreyas, R., Akshata, D.M., Mahanand, B.S., Shagun, B., Abhishek, C.M., 2016. Predicting popularity of online articles using Random Forest regression, 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), pp. 1-5.
Song, W., Mahalec, V., Long, J., Yang, M., Qian, F., 2020. Modeling the Hydrocracking Process with Deep Neural Networks. Ind. Eng. Chem. Res. 59, 3077–3090.
Steurtewagen, B., Van den Poel, D., 2021. Adding interpretability to predictive maintenance by machine learning on sensor data. Comput. Chem. Eng. 152, 107381.
Sun, X., Opulencia, M.J.C., Alexandrovich, T.P., Khan, A., Algarni, M., Abdelrahman, A., 2022. Modeling and optimization of vegetable oil biodiesel production with heterogeneous nano catalytic process: Multi-layer perceptron, decision regression tree, and K-Nearest Neighbor methods. Environ. Technol. Innov. 27, 102794.
Umana, B., Zhang, N., Smith, R., 2016. Development of Vacuum Residue Hydrodesulphurization-Hydrocracking Models and Their Integration with Refinery Hydrogen Networks. Ind. Eng. Chem. Res. 55, 2391–2406.
Vapnik, V., 1999. The nature of statistical learning theory. Springer science & business media.
Vynckier, E., Froment, G., 1991. Modeling of the kinetics of complex processes based upon elementary steps. Kinetic and Thermodynamic Lumping of Multicomponent Mixtures 10, 131–161.
Williams, C., Rasmussen, C., 1995. Gaussian processes for regression. Advances in neural information processing systems 8.
Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14, 1–37.
Xing, J., Zhang, H., Zhang, J., 2023. Remaining useful life prediction of – Lithium batteries based on principal component analysis and improved Gaussian process regression. Int. J. Electrochem. Sci. 18, 100048.