Random Forest Regression to Predict Catalyst Deactivation in Industrial Catalytic Process

Wisnu Hafi Hanif, Fergyanto E Gunawan

Industrial Engineering Department, BINUS Graduate Program – Master of Industrial Engineering, Bina Nusantara University, Jakarta, 11480, Indonesia

***Abstract* – Catalyst deactivation has become a great concern in an industry with heterogenous catalyst-based production. An accurate model to predict catalyst performance is needed to optimize the maintenance schedule, avoid an unplanned shutdown, and ensure reliable operation. This research work applies a machine learning model to predict catalyst deactivation based on actual data from relevant multitube-reactor sensors. The product conversion is a crucial indicator of the catalyst performance degradation over time. Random forest regression (RFR) algorithm is chosen to construct the model. Hyperparameter tuning is applied and shows improvement over the default model. The result showed that the RFR model could predict the conversion as a time series function. The feature importance analysis shows the most influencing factor and facilitates the model interpretation.**

***Keywords* – catalyst deactivation, heterogenous catalyst, petrochemical process, machine learning, random forest regression.**

1. Introduction

The rising demand for chemical synthesis products, petroleum refining, polymers, petrochemicals, and the environment has driven the industry to find ways to increase production capacity efficiently from an operational and cost perspective. Catalyst technology in the chemical industry is one way to speed up the production process and reduce energy consumption. Catalysts are substances that speed up the reaction process and reduce activation barriers [1]. The use of catalysts has a significant economic impact on the petroleum, chemical, health, and environmental industries. More than 85% of chemical products worldwide are produced with the help of catalysts [2].

Catalyst technology can be categorized as a green technology because of its impact on the environment. Industries with catalyst technology produce significantly less waste than traditional industries [2]. The United States Environmental Protection Agency (EPA) defines green technology as the design of chemical products and processes that reduce or eliminate the use or generation of hazardous substances and are formulated in 12 principles, including the use of the catalyst technology for more effective production and minimizing waste [3].

The phenomenon of catalyst deactivation often characterizes the catalyst replacement process. Catalyst deactivation is the degradation of catalytic activity/selectivity over time and has become a significant problem in industrial catalytic processes. The total cost for catalyst replacement and production shutdown loss annually could reach billions of dollars [4]. If the catalyst in the reactor is replaced while some sites still have active catalysts, there will be a waste of costs. On the other hand, if the catalyst is deactivated prematurely, the production unit must be shut down early [5].

Previous related studies have been conducted. Several works predict catalyst deactivation on an industrial scale using a process simulator [6] and laboratory experiments [7]. A combination of laboratory measurements and online analyzers was used to monitor the decrease in catalyst activity [8]. Others used soft sensors to identify variations in catalyst activity using the data-based modeling (DBM) philosophy [9].

DOI: 10.18421/TEM111-12
https://doi.org/10.18421/TEM111-12

Corresponding author: Fergyanto E Gunawan,
Industrial Engineering Department, BINUS Graduate Program – Master of Industrial Engineering, Bina Nusantara University, Jakarta, 11480, Indonesia.
Email: fgunawan@binus.ac.id

Received: 01 October 2021.
Revised: 07 December 2021.
Accepted: 12 December 2021.
Published: 28 Februrary 2022.

Several studies have focused on operational monitoring and detection of reactor failures using predictive maintenance or machine learning. A study by [10] utilized a machine learning prediction model to support the development of autonomous control of small-scale reactors, such as Transportable Fluoride-salt-cooled High-temperature Reactor (TFHR). A machine learning method was also used to diagnose batch reactor failure for providing early fault detection to minimize the risk of thermal runaway [11] and diagnose nuclear reactor cores [12].

Meanwhile, researchers had explored factors influencing the activity of heterogeneous metal catalysts by using a machine learning method [13], [14], and understanding the structure of the catalyst with Artificial Neural Networks [15]. Machine learning algorithms and chemo-informatics are considered capable of accelerating the progress of catalyst development by recognizing patterns that cannot be understood in large data sets compared to traditional methods during the development of a catalyst [16]. Other studies have focused on predicting the yield and potential output of components resulting from catalytic reactions [17], [18], [19].

This paper is designed to predict catalyst deactivation by observing the degradation of catalyst conversion using a machine learning method. Data from actual sensors from a multitube reactor in a petrochemical process will be involved to predict catalyst activity. The aim of this research is to demonstrate the use of the latest method to construct an advanced predictive maintenance system in process engineering.

2. Literature Review

Catalyst Deactivation Mechanism

Catalyst deactivation or catalyst decay is the loss of catalyst activity or selectivity over time [4]. Catalyst deactivation may cause losses in conversion and selectivity degradation over time and inhibit production rates, which require replacement or regeneration [20], [21]. Details of factors influencing catalyst deactivation are provided in Table 1.

Table 1. Factors influencing catalyst deactivation [4]

Category	Description
Poisoning	Strong absorption of a chemical compound on the catalyst surface blocks its active site.
Fouling	Physical deposition of fluid on the catalyst surface blocks the sites and pores, and leads to activity loss. It may destroy catalyst particles and blockage of the reactor cavity.
Thermal degradation	The crystallite growth in the catalytic phase reduces the catalytic surface or buffer area and destroys the pore in the active phase crystallites (sintering). In general, sintering takes place at high reaction temperatures and is deteriorated by the presence of moisture.
Gas/Vapor–Solid & Solid-State Reactions	The vapor phase and the catalyst surface react and produce a set of inactive phases generated by sintering the adsorbate interaction.
Mechanical degradation	The stacking process destructs monolithic catalysts. The turbulent flow results in erosion due to particle collisions. The catalyst debris accumulates at the reactor bed, increasing the pressure drop.

Random Forest Regression (RFR)

The random forest is a supervised learning algorithm that uses ensemble learning techniques. Ensemble learning is a random forest combining several aggregate decision trees to predict or classify the output of variables [22]. The random forest method begins by selecting several samples at random (bootstrapping) from a subset of training data through a bagging process. Bagging is a technique of replacing training data by re-sampling selected randomly without deleting the selected data from its input of the following subset [23].

The random forest method developed by [24], involves a set of tree classifiers $\{h(\mathrm{x}, \Theta_k), k = 1,...\}$ from the training set, where Θ_k is an identically distributed random vector of trees that run as many as k, and x is the input vector. After several tree nodes are formed sequentially, the voting process is carried out. The tree classifier that gains the highest score is the model output. The decision functions for majority voting [25] is:

$$H(x) = \arg\max_Y \sum_{i=1}^{k} I(h_i(x) = Y),$$

where $H(x)$ is combinations of model classifier, h_i is each node tree, Y is output variable, and I is indicator function.

Random sample switching allows the data to be used more than once in another training subset sequence, resulting in more stable and robust prediction, especially when dealing with high variations in the input data. The non-selected data in the bagging process for training in each node tree becomes part of another subset in the form of out-of-bag (OOB), which is used to predict the performance of the classifier [23].

Apart from classification problems, random forest can also be used for regression problems, which is called random forest regression (RFR). In contrast to the classification task which produces categorical and binary variables, the estimator of the RFR produces a continuous-value output [26]. RFR uses the same steps as the random forest classifier. The RFR model works by forming a tree that depends on the random variable Θ, relative to its categorical class so that the tree predictor $h(\mathrm{x}, \Theta_k)$ produces a continuous value output.

3. Research Method

The framework of the research is presented in Figure 1. The research involves data acquisition, pre-processing, RFR model deployment, and RFR model development to predict catalyst deactivation. The detail of each step is described in Figure 1.

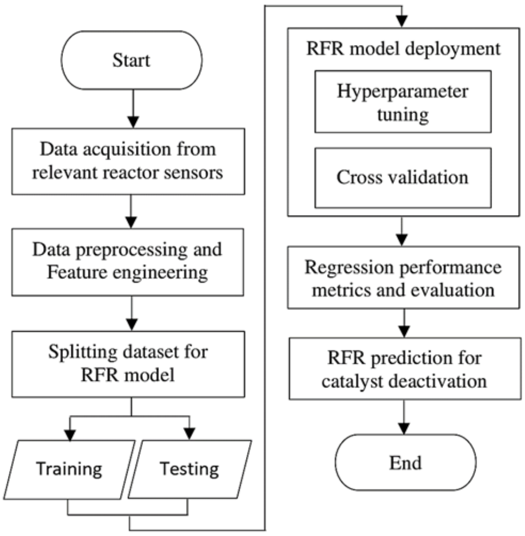

Start
Data acquisition from relevant reactor sensors
Data preprocessing and Feature engineering
Splitting dataset for RFR model
Training
Testing
RFR model deployment
Hyperparameter tuning
Cross validation
Regression performance metrics and evaluation
RFR prediction for catalyst deactivation
End

Figure 1. Research framework

Data acquisition and pre-processing

The datasets contain historical records of the continuous hydrocarbon oxidation process from a petrochemical company. The process is a catalytic gas phase carried out in a multi-tubular fixed-bed reactor. Typical conversion in a reactor is around 99.8% for the fresh catalyst and 88.7 % for the spent catalyst. After reaching a predetermined threshold, the spent catalyst is usually regenerated every 5–6 months. Catalyst is also regularly replaced after 48 months of use, and this is where catalyst has the highest performance. Indicators that suspect the changes in catalyst activity are influenced by the following factors [4]:

- The total mass flow of reactor feed;
- Reactant component concentration of reactor feed;
- The changes in pressure of inlet and outlet reactor;
- The changes in the pressure difference between catalyst beds inside the multi-tubular reactor;
- The gas recycle ratio to the reactor;
- The changes in temperature of inlet and outlet reactor; and
- The changes in temperature difference between catalyst bed inside the multi-tubular reactor.

Based on the above indicators, we selected relevant sensors and instrumentation as predictor variables in Table 2. Multi tubular fixed-bed reactor observed in this research is equipped with multiple sensors. The dataset is historical data from 68 relevant reactor sensors/instrumentations downloaded from the Plant Information Management System (PIMS) software. The data are time series with 8 hours intervals from 192 months of reactor operation, totaling 18,723 data.

The range of data for each column is highly varied, from decimals to hundreds. So, to make the algorithm work faster and perform better, the data are scaled with the normalization method to a range of 0 and 1. At the end of predicting phase, the target variable scale is inverted back to the original value for better visualization.

The last pre-processing is dividing datasets into the training set, validation set, and testing set. The latest datasets are assigned for testing, while the oldest datasets are for training and validation. The percentage of the training set is 70%, while the percentage of both validation and testing set is 15%.

Our goal is to predict catalyst performance by estimating the degradation of catalyst conversion using a machine learning model based on actual input sensors.

Table 2. Data features

ID	Description	Value range	Units	Type
AI-xxx	Organic component concentration analyzer	0.005–12.900	%	Predictor
DP-xxx	Differential pressure between catalyst bed	16.383–44.775	kPa	Predictor
DT-xxx	Differential reactor temperature	22.402–80.546	°C	Predictor
FI-xxx	Mass flow rate	668.779–836.911	kg/hr	Predictor
PI-xxx	Pressure inlet reactor	0.133–70.812	kPa	Predictor
RX-xxx	Gas recycle ratio to a reactor	19.220–37.630	%	Predictor
TI-xxx	Catalyst temperature sector *i*	277.009–392.321	°C	Predictor
CONV.	Product conversion	88.948–99.839	%	Target

Model Hyperparameter Tuning

Hyperparameter tuning of RFR seeks the best value to describe the dataset and develop the best fitting model to be deployed. We considered five parameters to be defined in the tuning model. Those are the number of estimators, max features, max depth, min samples split, min samples leaf, and bootstrap. Detail of hyperparameter component is delivered in Table 3.

Table 3. Hyperparameter value

Parameters	Value
N estimator	[10,…,500], step 10
Max features	'auto', 'sqrt'
Max depth	[10,…,200], step 10
Min samples split	[2,…,10], step 1
Min samples leaf	[1,…,10], step 1

The hyperparameter tuning for random forest regression method was carried out with Randomized Search CV, which hyperparameters values are drawn randomly with given number of iterations. This study runs 500 random iterations. The best hyperparameter output is determined based on mean squared error metrics.

Model Validation and Accuracy

Cross-validation was applied to assess the effectiveness of the model that has been built by evaluating several available inputs. The purpose of cross-validation is to find the best fitting model and mitigate overfitting.

For the regression task expanding windows cross-validation approach was used in this research. The test /validation data is fixed in number by using a different number of training inputs. The cross-validation scheme of the expanding windows technique with a total of five splits is illustrated in Figure 2. Cross-validation splits for time series using the TimeSeriesSplit function from Scikit-learn library.

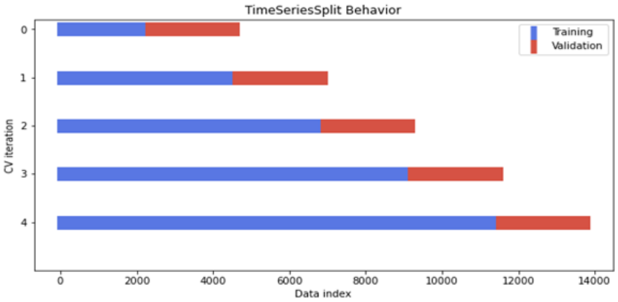


Figure 2. Time series split scheme for cross validation

Firstly, in developing the model, the RFR model prediction is evaluated using regression metrics. The model with the lowest error is then applied for predicting catalyst conversion using a testing dataset. In both training and testing stages, the models are evaluated by mean squared error (MSE), root mean squared error (RMSE), Akaike information criterion (AIC), and coefficient of determination (R^2), such as described by Eqs. (1) to (4).

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

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

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{3}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{4}$$

where y_i is the actual observed value, $\bar{y}$ is mean of y, $\hat{y}_i$ is the predicted value, k is the number of estimated parameters in the model, and $\hat{L}$ is the maximum value of the likelihood function for the model.

Models are developed in Jupyter Notebook version 6.1.3 which is supported by package Python Scikit-learn [27] and it is processed in 64-bit operation system with 12 GB of RAM.

4. Result and Discussion

Dataset and Model Development

Preprocessing started with converting the obtained data into a CSV file and imported it into Jupyter Notebook. We decided to omit the plant shutdown condition from the dataset to not read as learning input. We also remove outliers for all columns and replace the missing value with imputation techniques. These processes leave 16,213 observations as input.

Features and target variables must be separated. Features are sensors used to predict catalyst performance as targets. The original data contains 67 predictor variables and one target variable. However, to reduce redundancy and make model learning faster, features with low correlation ($-0.4 >$ or < 0.4) towards the targets variable are eliminated. We also performed a multicollinearity check to detect high correlation among predictor variables and remove it from the dataset—the processes resulting in 17 features used as model predictors such as shown in Table 4.

Table 4. Selected features for model predictor

Features	Correlation to conversion
AI-007	−0.46
AI-009	−0.80
DP-001	0.81
DP-002	0.53
DP-004	0.68
DT-001	0.54
PI-001	0.43
PI-002	0.47
TI-005	0.47
TI-008	−0.61
TI-012	0.54
TI-013	0.76
TI-018	0.70
TI-019	0.65
TI-022	0.64
TI-024	−0.41
TI-037	−0.68

Hyperparameter Tuning

Five hyperparameters are considered in this study to improve the performance of the RFR model: number of decision trees (*n*_estimators), the minimum number of samples required to split an internal node (min samples split), the minimum number of samples required to be at a leaf node (min samples leaf), maximum depth of the tree (max depth), number of features (max features) and number of splits in each tree (max depth). Due to many parameter combinations, such as in Table 3., we used a randomized search CV with 500 iterations. The advantage of this method is computationally faster than a grid search since it does not explore the entire combinations. The result of hyperparameter tuning is provided in Table 5.

The performance before tuning means that the model is using a default value for each parameter. In the case of the RFR model in Scikit-learn, the default value for tree estimators is 100, the min samples leaf is 1, min samples split 2, max features is 'auto', and none for max depth. While bootstrap mode in default is 'True', and is also applied in hyperparameter tuning. A comparison of performance scores before and after tuning is provided in Table 6.

Table 5. RFR model parameters for data testing

Parameter	Value
N estimators	70
Min samples split	4
Min samples leaf	5
Max features	'sqrt'
Max depth	50

Table 6. RFR model performance

RFR Model	MSE	RMSE	AIC	R^2
Before tuning	1.288	1.135	330.175	0.916
After tuning	0.698	0.835	-1183.975	0.956

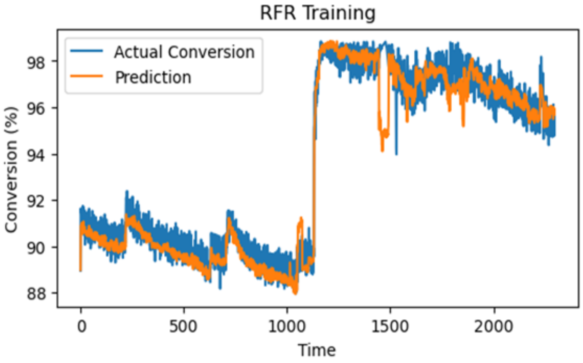


Figure 3. RFR training prediction and actual conversion

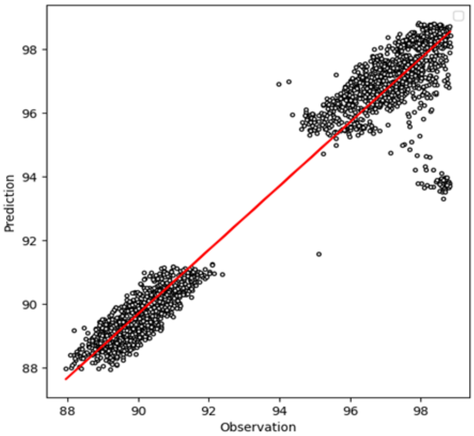


Figure 4. The relationship between observed and predicted conversion using training datasets

Illustration of the RFR model prediction using validated training dataset is presented in Figure 3., while the relationship between observed training data and predicted is shown in Figure 4.

To fully explore the effect of each hyper-parameters in predicting conversion, we use the grid search method to visualize the best hyper-parameters in the prescribed range, which is visualized in Figure 5. It is worth noting that max depth, min samples split, and min samples leaf does not significantly affect the RFR model.

Feature Importance

Feature importance in random forest is variable ranking according to respective contributions that construct the trees. It is measured as to how well variables reduce the impurity of a node during the learning, which is indicated by a percent increase in mean squared error. The higher the importance value means the more important feature.

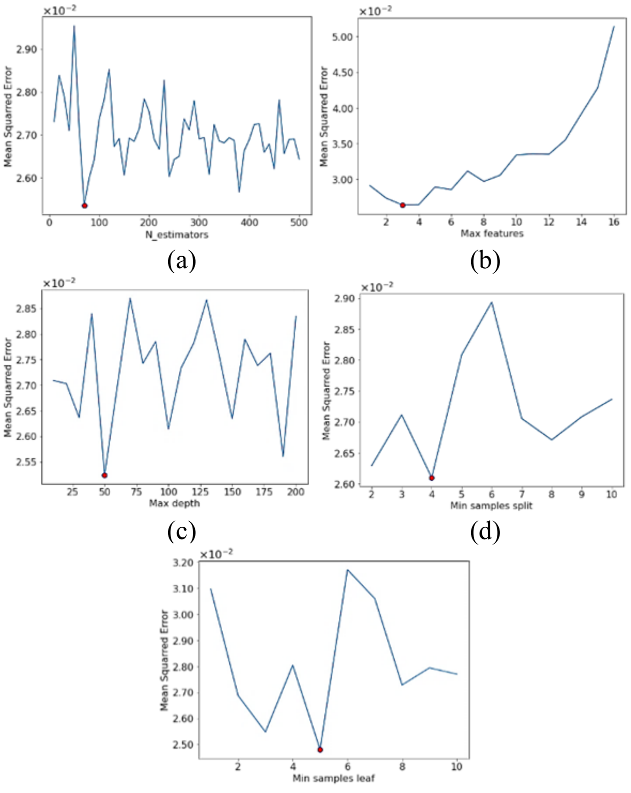
The 16 predictor variables were measured and ranked based on their weight. The result of feature importance is shown in Figure 6. It seems that differential pressure of bed catalyst DP-001 is the essential feature to construct the tree, followed by catalyst temperature TI-018 and TI-038. At the same time, the least important feature is feed component analyzer AI-007.

Data Testing

Optimized parameter based on previous section is applied to predict catalyst deactivation in catalytic rector process in petrochemical industry. As aforementioned, the dataset is split into training, validating, and testing. In this stage, the number of data testing are 2,432 observations. Illustration of the RFR model prediction using test dataset is presented in Figure 7. The catalyst deactivation is determined by observing degradation of conversion overtime.

The relationship between observed testing data and predicted is shown in Figure 8. The result shows that the model has a good performance in predicting conversion with respect to actual conversion.

The metrics for the testing result after rescaled to original values indicated that mean squared error (MSE) is 1.232, and root mean squared error is 1.110, Akaike information criterion (AIC) is -463.890, and coefficient of determination (R^2) is 0.803.



(a) (b)

(c) (d)

(e)

Figure 5. Grid Search for RFR model. (a) N estimators, (b) max features, (c) max depth, (d) min samples split, (e) min samples leaf

The model and framework proposed in this paper can be applied not limited to downstream hydrocarbon processes such as in this study but also to petroleum refining, oleochemical, pharmacy, and other petrochemical products. However, some features or parameters need to be adjusted or validated for specific industrial processes.

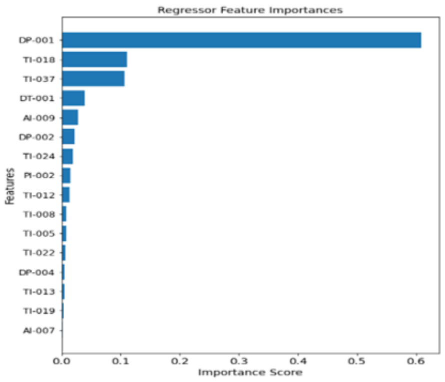


Figure 6. Feature importance

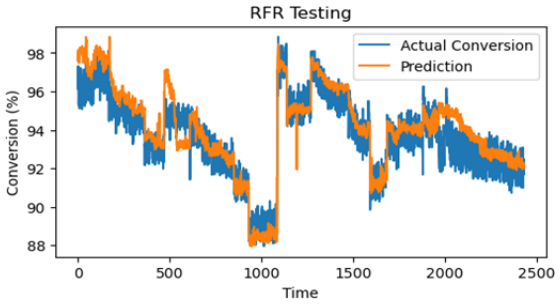


Figure 7. RFR testing prediction and actual conversion

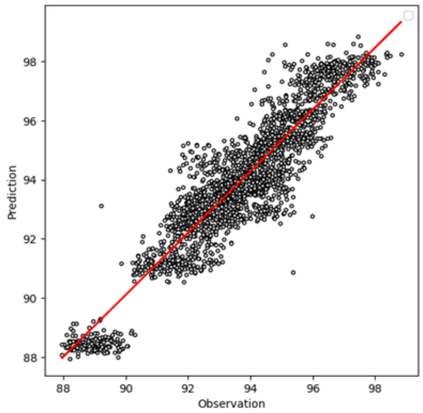


Figure 8. The relationship between observed and predicted conversion using testing datasets

5. Conclusion

This paper demonstrated machine learning methods based on actual data from catalytic reactor sensors to predict catalyst deactivation. The degradation of reactor conversion over time is a target to determine catalyst deactivation. Hyperparameter tuning worked well to improve model performance. Based on the feature importance analysis results with random forest regression algorithm, it shows that differential pressure and catalyst reactor temperature sensors are the most contributing to the model for predicting catalyst deactivation.

For further research, we seek to compare a similar study with another machine learning algorithm, i.e., ARIMA, support vector regression (SVM), kernel ridge regression (KRR), multilayer perceptron (MLP), or long short term memory (LSTM), to find the best model to predict catalyst deactivation.

Acknowledgments

The research is supported by Direktorat Riset dan Pengabdian Masyarakat Direktorat, Jenderal Riset dan Pengembangan, Kementerian Riset, Teknologi, dan Pendidikan Tinggi, Republic of Indonesia, via Penelitian Terapan Unggulan Perguaran Tinggi dan Penelitian Terapan with the main contract number: 309/E4.A/AK.04.PT/2021 and the derivative contract number: 3507/LL3/KR/2021 with the title: "Pengembangan Sistem Pemantauan Struktur - Tahap 1: Model Matematis untuk Prediksi Kerusakan."

References

[1]. Hanefeld, U., & Lefferts, L. (Eds.). (2018). *Catalysis: An integrated textbook for students*. John Wiley & Sons.

[2]. Heveling, J. (2012). Heterogeneous catalytic chemistry by example of industrial applications. *Journal of Chemical Education*, *89*(12), 1530-1536.

[3]. United States Environmental Protection Agency. (n.d). *Green Chemistry*. Retrieved from: https://www.epa.gov/greenchemistry [accessed: 20 September 2021].

[4]. Argyle, M. D., & Bartholomew, C. H. (2015). Heterogeneous catalyst deactivation and regeneration: a review. *Catalysts*, *5*(1), 145-269.

[5]. Robinson, P. R. (2004, September). Catalyst life management with a predictive deactivation model. In *2004 NPRA PADS Conference, PD-04-171*.

[6]. Mohaddecy, S. R. S., & SADIGHI, S. (2013). Predicting catalyst lifetime. *Petroleum Technol Q*, *14*, 85.

[7]. Abbas, A., & Sharifah, R. W. A. (2017). Prediction of Industrial Catalysts Deactivation Rate Using First Principle Model and Operating Data. *Malaysian Journal of Analytical Sciences*, *21*(1), 204-212.

[8]. Farsang, B., Németh, S., & Abonyi, J. (2015). Online Monitoring of Catalyst Deactivation Based on Data Reconciliation and Flowsheeting Simulator. *Periodica Polytechnica Chemical Engineering*, *59*(2), 145-150.

[9]. Gharehbaghi, H., & Sadeghi, J. (2016). A novel approach for prediction of industrial catalyst deactivation using soft sensor modeling. *Catalysts*, *6*(7), 93. https://doi.org/10.3390/catal6070093

[10]. Zeng, Y., Liu, J., Sun, K., & Hu, L. W. (2018). Machine learning based system performance prediction model for reactor control. *Annals of Nuclear Energy*, *113*, 270-278.

[11]. Subramanian, S., Ghouse, F., & Natarajan, P. (2014). Fault diagnosis of batch reactor using machine learning methods. *Modelling and Simulation in Engineering*, 1-15.

[12]. Kim, H., Yun, D., Shin, H., Moon, S., & Lee, D. (2020, July). Feasibility study on machine learning algorithm in nuclear reactor core diagnosis. In *Proceedings of the Transactions of the Korean Nuclear Society Virtual Spring Meeting, Korea (online)* (pp. 9-10).

[13]. Takigawa, I., Shimizu, K. I., Tsuda, K., & Takakusagi, S. (2018). Machine learning predictions of factors affecting the activity of heterogeneous metal catalysts. In *Nanoinformatics* (pp. 45-64). Springer, Singapore.

[14]. Jinnouchi, R., & Asahi, R. (2017). Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm. *The journal of physical chemistry letters*, *8*(17), 4279-4283.

[15]. Li, H., Zhang, Z., & Liu, Z. (2017). Application of artificial neural networks for catalysis: a review. *Catalysts*, *7*(10), 306.

[16]. Zahrt, A. F., Henle, J. J., Rose, B. T., Wang, Y., Darrow, W. T., & Denmark, S. E. (2019). Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*, *363*(6424).

[17]. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., & Jensen, K. F. (2017). Prediction of organic reaction outcomes using machine learning. *ACS central science*, *3*(5), 434-443.

[18]. Yada, A., Nagata, K., Ando, Y., Matsumura, T., Ichinoseki, S., & Sato, K. (2018). Machine learning approach for prediction of reaction yield with simulated catalyst parameters. *Chemistry Letters*, *47*(3), 284-287.

[19]. Rhone, T. D., Hoyt, R., O'Connor, C. R., Montemore, M. M., Kumar, C. S., Friend, C. M., & Kaxiras, E. (2019). Predicting outcomes of catalytic reactions using machine learning. *arXiv preprint arXiv:1908.10953*.

[20]. Beeckman, J. W. (2020). *Catalyst Engineering Technology: Fundamentals and Applications*. John Wiley & Sons.

[21]. Bartholomew, C. H., & Farrauto, R. J. (2011). *Fundamentals of industrial catalytic processes*. John Wiley & Sons.

[22]. Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*(1), 1-14.

[23]. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804-818.

[24]. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[25]. Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32

[26]. VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.".

[27]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.