Data Driven Machine Learning For Fault Detection And Classification In Binary Distillation Column

Silvester Bennys jakes *, M. Mythily, D. Vasanthi , D. Manamalli

Department of Instrumentation, MIT Campus, Anna University, Chennai, India

Emails: jakesbenjamin22@gmail.com ; mythily_eie@yahoo.co.in; vasanthi_d1@rediffmail.com; manamalli_m@yahoo.com

Abstract

Mathematical programming can express competency concepts in a well-defined mathematical model for a particular Any system that runs is always be expected to experience faults in different ways. Any change in the physical state of numerous components, control machinery, as well as environmental factors, might result in these problems. In process industries, where prompt detection is crucial in maintaining high product quality, dependability, and safety under various operating situations, finding these flaws is one of the most difficult tasks. The goal of this project is to implement several machine learning techniques for fault identification and classification in a binary distillation column. A pilot binary distillation unit (UOP3CC) is utilized for this purpose. The set up is run under normal operating conditions and the real time data is collected. Three common faults namely reboiler fault, feed pump fault and sensor fault are introduced one at a time and the faulty data is collected. These data are then introduced in to different machine learning algorithms like Logistic Regression, KNN, Naive Bayes, Decision Tree, Gradient Boosting, X Gradient Boosting, SVC and Light Gradient Boosting for model development. 70% of the data samples used for training and 30% of data samples are used for testing. It is found the Decision tree algorithm gives the best accuracy possible with 99.9%. Using decision tree algorithm, fault classification is performed for different datasets and is found that the algorithm was able to classify accurately even for new untrained datasets.

Keywords: UOP3CC binary distillation column; Normal and faulty data; Machine learning algorithm; Fault classification.

1. Introduction

A fault is characterised as an abnormal activity that leads the system to diverge in an unacceptable way from its intended course of operation [3]. Chemical process industries all over the world have frequent accidents brought on by system flaws, which have an impact on both the plant's performance and system components. It is crucial to find flaws as soon as possible to avoid problems like plant shutdown and safety concerns [5]. By reducing disruption, fault detection in any chemical process will keep the system trustworthy and secure. These faults can arise any time, such as sudden failure of any equipment at the plant, or evolve over time with gradual wear and tear within the process or sensor drift. For complex processes like distillation column, fault detection is essential for safe and productive operation. These goals can be met by contrasting the process actual behavior with a model of typical or desirable process behavior. The monitoring of the discrepancy between actual processes and those anticipated by the model forms the basis for the detection of process problems [5]. The classification of faults can be done using the deviation.

Physical model-based, reliability-based, and data-driven methods can all be used to find faults. Techniques based on physical models rely on mathematical representations of the research objects. The reliability-based techniques

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

adapt probability theory and knowledge-based statics, but it necessitates system-specific prior information [6]. For extremely dynamic and non-linear processes, it is inappropriate. Data-driven approaches, on the other hand, simply need the prior data for the model training and don't require any prior knowledge of the process. This prompted the use of machine learning approaches for highly dynamic, nonlinear processes like distillation columns [2].

2.Data Collection From Realtime Setup

UOP3CC is a computer- controlled distillation column with a 50 mm periphery that has eight sieve plates and downcomers. A temperature detector is erected into each plate and is deposited to take an accurate reading of the liquid's temperature. A central feed portion divides the columns, which are placed vertically for custom liquid/ vapor inflow. The P&ID diagram of a binary distillation column is shown in Fig. 1. The distillation column consists of 8 seive plates mounted by using the central iron rod. Each 8 seive plates correspond of one thermocouples, in order to measure the temperature of the column. By using this setup can suitable to handle the batch and nonstop process. Corresponding temperature samples are collected using the Armfield software. The setup is operated in batch mode with distilled water and ethanol fed into the feed tank in 3:1 ratio.UOP3CC Binary Distillation Column consists of 2 feed tanks each of 4 litres capacity. The feed of ethanol water mixture is given to the reboiler by using the feed pump with the help of viton tube. The feed is given to the reboiler through the centre of the distillation column.

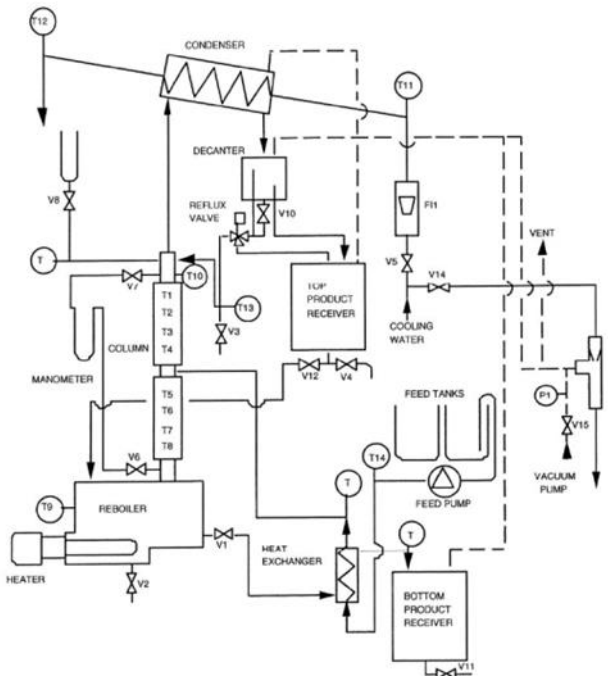

Figure 1: P&ID Diagram of UOP3CC Distillation Column

The Reboiler power range is set at 1.15kw and feed pump at 200 RPM. Before the set up is powered on, the colling water circulation around the condenser is ensured so as to avoid overheating and the flowrate is set at 3500cc/min. Around 10000 data were collected from the set up under normal operation. The data were acquired through Armfield software. Three different faults namely faults in reboiler, feed pump and sensor were introduced one at a time and the faulty data were collected. The data from four temperature sensors were collected namely column's feed tray temperature, reboiler liquid temperature, column top temperature and temperature of top product/reflux [1] [2] [3]. Data samples were grouped as normal data, reboiler fault data, sensor fault data, feed pump fault data. Table 1 explains the types of data samples collected.

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

Table1**:** Types Of Data Collection

S.NO	DATA TYPE	DATA COLLECTION	PROCESS
1.	NORMAL DATA	Normal data collected from T9, T10, T13, T5 temperature sensors.	Reboiler Condition -Works at 1.15 kw. Feed pump Condition -Works until reaches the Reboiler level of 4.5litres at200RPM. Sensor Condition -Works Properly. Condenser Condition -Inlet to Condenser also properly sent at 3500cc/min.
2.	REBOILER FAULT	Faulty data collected from T9, T10, T13, T5 temperature sensors.	Reboiler Condition -Off state Feed pump Condition -Off state Sensor Condition -works properly Condenser Condition -Inlet to Condenser also properly sent at 3500cc/min.
3.	SENSOR FAULT	Normal data collected from T5, T13 and faulty data collected from T9 and T10 temperature sensors.	Reboiler Condition -Works at 1.15 kw Feed pump Condition -Off state Sensor Condition -T9 and T10 at Off state Condenser Condition -Inlet to Condenser also properly sent at 3500cc/min.
4.	FEEDPUMP FAULT	Normal data collected from T9, T10, T13, and faulty data collected from T5	Reboiler Condition -Works at 1.15 kw Feed pump Condition -2litres of Feed is given to Reboiler at 200 RPM. Sensor Condition -works properly Condenser Condition -Inlet to Condenser also properly sent at 3500cc/min.

3. Fault Detection And Classification

Exploratory Data Analysis For Normal And Faulty Data

Fig. 2. shows the count plot indicating total number of counts for normal, reboiler fault, sensor fault and feed pump fault [9]. It is found that there are 11,238 normal data, 6448 reboiler fault data, 3037 sensor fault data, 901 feed pump fault data present in dataset.

Strip plot is also used for datasets in order to see the range and data distribution of samples for each orders. Fig.3 explains the range and data distribution for T9 sample. It is found that the normal data ranges between 40 and 100^0C , reboiler fault data ranges between 70 and 100^0C, sensor fault ranges between 10 and 120^0C and feed pump range between 95 and 100 ^{0}C.

The heat map is obtained for collected samples in order to show whether it is positive correlation or negative correlation or not a correlation [2]. Fig. 4. explains about the correlation between T9, T10, T5, T13 in which all have positive correlation.in different forms like low positive correlation , medium positive correlation and high positive correlation.

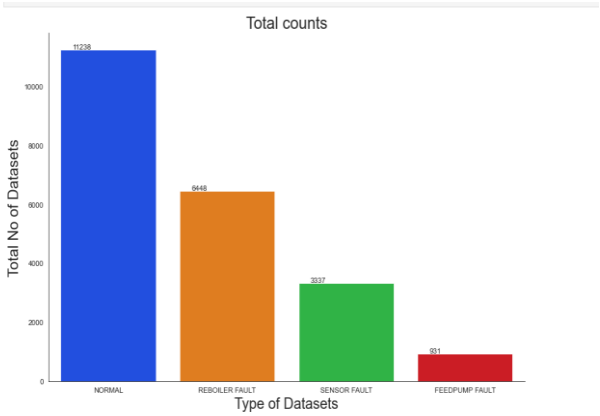

Figure 2: Count plot for normal and faulty data

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

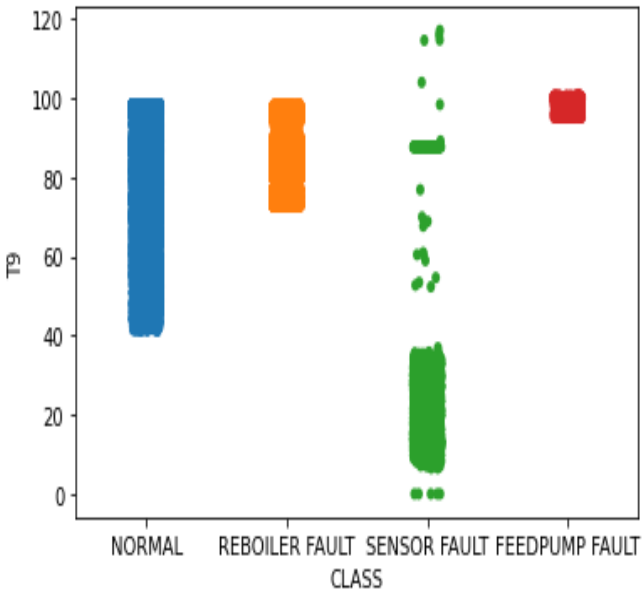

Figure 3: Strip plot for normal and faulty data

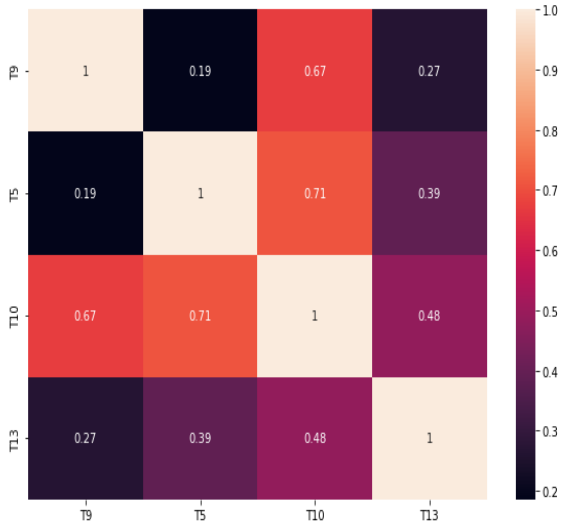

Figure 4: Heat map for normal and faulty data

The Bar plot in Fig. 5. show that the data is of imbalanced type which means that all the 4 types of data are not equal in range and distribution. It is observed that the normal and faulty data are represented by numerical variables. 0 for normal data, 1 for reboiler fault, 2 for sensor fault, 3 for feed pump fault.

The box plot in Fig. 6 explains the distribution and standard for T9 samples and also shows the present of outlier [5]. For T9 sample, the normal data ranges from 40 to 100 and it has no outliers. Reboiler fault ranges from 70 to 100 and it has no outliers whereas sensor fault ranges from 10 to 40 and it has some outlier. Feed pump fault ranges between 90 to 100 which has no outlier. From the figure, the normal and faulty data are represented by numerical variable like 0 for normal data, 1 for reboiler fault, 2 for sensor fault, 3 for feed pump fault.

Elbow plot is obtained by plotting the graph for number of clustering vs Total Variances for each cluster. At the particular point, the curve will bend. This corresponds to the number of cluster. Fig. 7. explains about the number of clustering present in the data samples which is almost equal to 4.

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

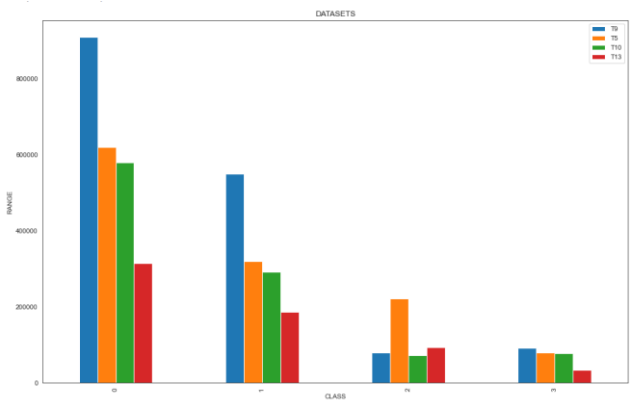

Figure 5: Bar plot for normal and faulty data

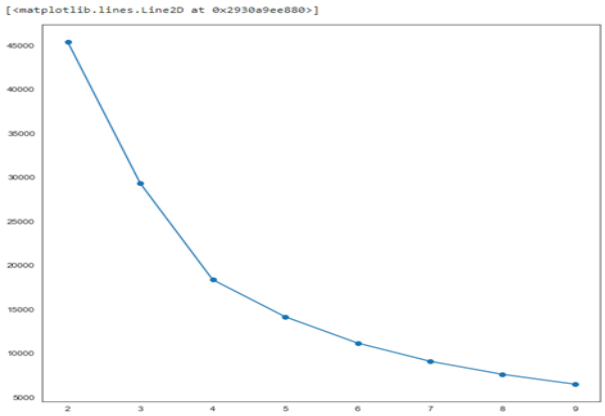

Figure 6: Box plot for normal and faulty data

The Silhouette analysis was performed for samples in order to show the number of clusters present in the dataset. Table 2 explain about silhouette score for each cluster . The silhouette analysis are founded by for each cluster the medians are founded based on that neighbours are grouped and corresponding scores are displayed[5]. Based on the calculation, cluster 4 gives the high score. So it also prove that the dataset contains 4 different samples. Fig. 8. shows groupings for 4 different samples.

Table 2: Silhoutte Score

NUMBER OF CLUSTER	SILHOUTTE SCORE
2	0.4856280730903792
3	0.5229241350426569
4	0.5500188599354087
5	0.49060953305287464
6	0.5079527432485337
7	0.49930366475888205
8	0.4750220132812603
9	0.507286793654344
10	0.5065030721043929
11	0.5110629141523253
12	0.5280507174012989
13	0.542952717613189
14	0.5382952717613189

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

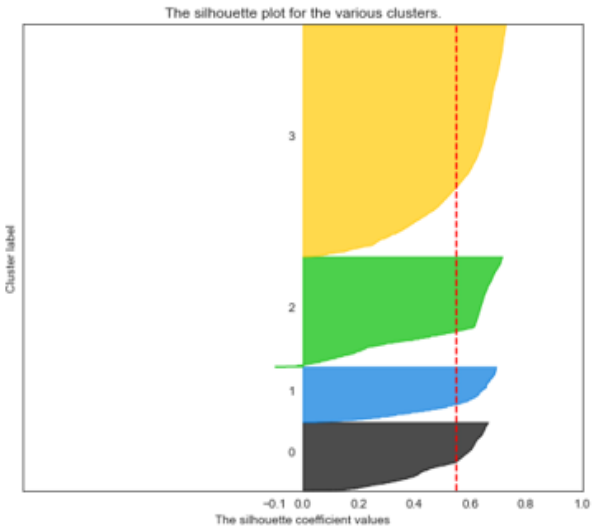

Figure 7: Silhoutte Analysis for normal and faulty data

4.Model Development Using Machine Learning Algorithms

Logistic regression is one form of supervised learning which is used to determine or forecast the likelihood that a binary (yes/no) event will occur. Figure 8 explains the model accuracy or classification score for the datasets in which hyperparameters of the logistic regressions are c, penalty, solver, tol etc. The hyper parameters are tuned in order to get the precision, recall, f1-score, support scores. It shows model accuracy around 69% in which the model is moderately trained.

```
Best Parameters : {'C': 10, 'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.1}

 [[ 166   20    0    0]
 [   0 1845  403    0]
 [   0  894  396    0]
 [   1   21    3  642]]

Model Accuracy : 69.437
                precision    recall  f1-score   support

FEEDPUMP FAULT       0.99      0.89      0.94       186
        NORMAL       0.66      0.82      0.73      2248
REBOILER FAULT       0.49      0.31      0.38      1290
  SENSOR FAULT       1.00      0.96      0.98       667

      accuracy                           0.69      4391
     macro avg       0.79      0.75      0.76      4391
  weighted avg       0.68      0.69      0.68      4391
```

Figure 8: Accuracy using Logistic Regression

The Naive Bayes algorithm is based on the Bayes theorem which is a supervised learning method for classification issues. Figure 9 explains about the model accuracy developed by naive bayes algorithm which also tells about the precision, recall, f1, support scores. It shows the model accuracy around 88.2 % in which the model is moderately trained [1]. One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. Figure 10 shows the model accuracy and also shows the Accuracy, recall, f1-score, support score. It shows the model accuracy of 99.3% in which the model is very well trained for KNN algorithm [2].

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

```
[[ 182    4    0    0]
 [ 109 1951  188    0]
 [  36  156 1098    0]
 [   0   24    1  642]]


 The Accuracy is : 88.2
                precision    recall  f1-score   support

FEEDPUMP FAULT       0.56      0.98      0.71       186
        NORMAL       0.91      0.87      0.89      2248
REBOILER FAULT       0.85      0.85      0.85      1290
  SENSOR FAULT       1.00      0.96      0.98       667

      accuracy                           0.88      4391
     macro avg       0.83      0.92      0.86      4391
  weighted avg       0.89      0.88      0.89      4391
```

Figure 9: Accuracy using Naive Bayes

Decision tree, non-parametric supervised learning technique for classification and regression is implemented. The objective is to learn straightforward decision rules derived from the data features in order to build a model that predicts the value of a target variable. Fig. 11. shows the accuracy of the datasets and also explains about the precision, recall, f1-score, support accuracy. It shows the model accuracy of 99.9% in which among 8 algorithms it gives the highest accuracy. The model is trained very well [5].

A machine learning method called gradient boosting is used for classification and regression tasks. It provides a prediction model in the form of an ensemble of decision trees-like weak prediction models.Fig.12. shows the hyperparameters of gradient boosting and model accuracy for datasets.It shows the model accuracy of 84%.

```
[[ 186    0    0    0]
 [   0 2223   25    0]
 [   0    4 1286    0]
 [   0    1    0  666]]

Model Accuracy : 99.317
                precision    recall  f1-score   support

FEEDPUMP FAULT       1.00      1.00      1.00       186
        NORMAL       1.00      0.99      0.99      2248
REBOILER FAULT       0.98      1.00      0.99      1290
  SENSOR FAULT       1.00      1.00      1.00       667

      accuracy                           0.99      4391
     macro avg       0.99      1.00      1.00      4391
  weighted avg       0.99      0.99      0.99      4391
```

Figure 10: Accuracy using KNN

```
                precision    recall  f1-score   support

FEEDPUMP FAULT       1.00      1.00      1.00        77
        NORMAL       1.00      1.00      1.00      1126
REBOILER FAULT       1.00      1.00      1.00       649
  SENSOR FAULT       1.00      1.00      1.00       344

      accuracy                           1.00      2196
     macro avg       1.00      1.00      1.00      2196
  weighted avg       1.00      1.00      1.00      2196
```

```
DT.score(X_test, Y_test)
```

```
0.9995446265938069
```

Figure 11: Accuracy using Decision Tree

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

```
gbr = GradientBoostingClassifier(n_estimators=37,
                                 learning_rate=0.01,
                                 max_depth=1)
```

```
Model Accuracy is 0.8475942858674607
```

Figure 12: Accuracy using Gradient Boosting

```
svc = SVC(C= 0.5, kernel='linear')
```

```
Model accuracy score : 0.9738
```

Figure 13: Accuracy using Support vector classifier

```
xgb = XGBClassifier(n_estimators=10,
                    learning_rate=0.01,
                    colsample_bytree=0.45,
                    max_depth=1,
                    gamma=0,
                    reg_alpha=0,
                    reg_lambda=0,
                    objective='reg:squarederror')
```

```
Model Accuracy is 0.9178875567498863
```

Figure 14: Accuracy using X Gradient Boosting

SVM is also used to solve Classification and Regression problems. Fig. 13. explain about the hyperparameter of SVC and Model accuracy for datasets. It shows the model accuracy of 97% and in which model is trained very well [6-10]. An efficient and effective implementation of the gradient boosting technique is offered by the open-source package known as Extreme Gradient Boosting (XGBoost). Fig.14. shows the hyperparameter of algorithms and also shows the Model accuracy of an algorithms around 91% and it is trained very well.

The Train With AutoML tool employs LightGBM, a gradient boosting ensemble technique that is based on decision trees. LightGBM is a decision tree-based technique that may be applied to both classification and regression problems. For excellent performance with dispersed systems, LightGBM has been specially

```
lgb = LGBMClassifier(num_leaves=4,
                     learning_rate=0.01,
                     n_estimators=29,
                     max_bin=200,
                     bagging_fraction=0.8,
                     bagging_freq=3,
                     bagging_seed=5,
                     feature_fraction=0.5,
                     feature_fraction_seed=5,
                     min_sum_hessian_in_leaf = 11,
                     verbose=-1,
                     random_state=42)
```

```
Model Accuracy is 0.9822023820000491
```

Figure 15: Accuracy using Light Gradient Boosting

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

```
DT.predict([[42.0, 26.8, 26.6, 26.5]])
array(['NORMAL'], dtype=object)
DT.predict([[87.9, 42.7, 37.8, 28.4]])
array(['REBOILER FAULT'], dtype=object)
DT.predict([[15.5, 97, 12.7, 27.6]])
array(['SENSOR FAULT'], dtype=object)
DT.predict([[100.3, 86.6, 96.5, 39.3]])
array(['FEEDPUMP FAULT'], dtype=object)
```

Figure 17: Fault classification for normal and faulty data

designed. Fig. 15. shows the Hyperparameters of algorithms and model accuracy of the algorithms. It shows the model accuracy of 98% and the model is trained very well. The fault classification can be done by using the high accuracy model in which the Decision tree algorithm trained model very well with high accuracy. Figure 16 shows the fault classification for the given Datasets it was done by Decision tree algorithm. Figure 16 shows the fault classification for the New samples that was not present in the Datasets.

```
DT.predict([[41.0, 25.8, 25.6, 25.5]])#NEW
array(['NORMAL'], dtype=object)
DT.predict([[86.9, 41.7, 36.8, 27.4]])#NEW
array(['REBOILER FAULT'], dtype=object)
DT.predict([[14.5, 96, 11.7, 26.6]])#NEW
array(['SENSOR FAULT'], dtype=object)
DT.predict([[99.3, 85.6, 95.5, 37.3]])#NEW
array(['FEEDPUMP FAULT'], dtype=object)
DT.predict([[2.0, 8.0, 8.5, 8.7]])#NEW
array(['SENSOR FAULT'], dtype=object)
DT.predict([[21.0, 19.5, 0, 15.3]])#NEW
array(['SENSOR FAULT'], dtype=object)
DT.predict([[21.0, 9.5, 0, 5.3]])#NEW
array(['SENSOR FAULT'], dtype=object)
```

Figure 18: Fault classification for new normal and faulty data

The comparison between all the implemented machine learning techniques is given in Table 6.

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

Table 6: Comparision Of Machine Algorithms

ALGORITHMS	ACCURACY
Logistic Regression	69.4%
Naive Bayes	98.3%
Decision Tree	**99.8%**
Gradient Boosting	84.2%
X Gradient Boosting	91.7%
SVC	47.5%
Light Gradient Boosting	96.7%

From the table, it is inferred that among the algorithms, it is found that decision tree algorithm shows better result with 99.8% accuracy.

5. Conclusion

The common anomalies that occur in a UOP3CC Binary Distillation column was studied. The binary distillation column was run at optimized operating conditions and the data were collected (Normal data). Three commonly occurring faults were introduced one at a time which constitutes faulty data. Machine learning methods namely Logistic Regression, Gaussian Naive Bayes, KNN, Decision Tree, SVM, Gradient Boosting, X Gradient Boosting, Light Gradient Boosting for attaining best accurate model in binary distillation column. Finally Decision tree algorithm was implemented on the data sets for detecting and classifying the faults. Table 6 shows the accuracy comparison between 8 different machine learning algorithms.

References

[1] Wang, G.-Y.; Yang, Z.-H.; Zhang, Y.; Wang, H.-H.; Zhang, Z.- X.; Gao, B.-J. A Preliminary Fault Detection Methodology for Abnormal Distillation Column Operations Using Acoustic Signals. *Appl. Sci.* **2022**, *12*, 12657. https://doi.org/10.3390/app122412657.

[2] Taqvi, SAA, Zabiri, H, Uddin, F, et al. Simultaneous fault diagnosis based on multiple kernel support vector machine in nonlinear dynamic distillation column. *Energy Sci Eng*. 2022; 10: 814– 839. doi:10.1002/ese3.1058

[3] Taqvi, S.A., Tufa, L.D., Zabiri, H. *et al.* Fault detection in distillation\ column using NARX neural network. *Neural Comput & Applic* **32**, 3503–3519 (2020). https://doi.org/10.1007/s00521-018-3658-z.

[4] Neema Davis, Gaurav Raina, krishna P.Jagannathan: Aframework for End- to-End Deep Learning-Based Anomaly Detection in Transportation Networks. CoRR abs/1911.08793(2020).

[5] E. Lithoxoidou, C. Ziogou, T. Vafeiadis, S. Krinidis, D. Ioannidis, S. Voutetakis, D. Tzovaras, Towards the behavior analysis of chemical reactors utilizing data-driven trend analysis and machine learning techniques, Applied Soft Computing, Volume 94,2020, 106464, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.106464.

[6] Souza, Davi & Granzotto, Matheus & De Almeida, Gustavo & Oliveira Lopes, Luis Cláudio. (2014). Fault Detection and Diagnosis Using Support Vector Machines -A SVC and SVR Comparison. Journal of Safety Engineering. 2014. 18-29. 10.5923/j.safety.20140301.03.

[7] Goldstein, Markus & Uchida, Seiichi. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PloS one. 11. e0152173. 10.1371/journal.pone.0152173.

[8] Alcala, Carlos F., and S. Joe Qin. "Analysis and generalization of fault diagnosis methods for process monitoring." *Journal of Process Control* 21.3 (2011): 322-330.

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023

[9] Amer, Mennatallah, Markus Goldstein, and Slim Abdennadher. "Enhancing one-class support vector machines for unsupervised anomaly detection." Proceedings of the ACM SIGKDD workshop on outlier detection and description. 2013.

[10] Zhang K, Shardt YAW, Chen Z, Yang X, Ding SX, Peng K. A KPI-based process monitoring and fault detection framework for large-scale processes. ISA Trans. 2017 May;68:276-286. doi: 10.1016/j.isatra.2017.01.029. Epub 2017 Feb 9. PMID: 28190565.

Doi : https://doi.org/10.54216/JCIM.110105
Received: October 27, 2022 Revised: December 20, 2022 Accepted: January 18, 2023