COS40007- Artificial Intelligence for Engineering

Portfolio Assessment – 3

Name – Rupayan Banerjee
Student Number – 103538229
Studio Class – 1 – 3

This week's portfolio tasks mainly aimed at developing various ML models and analysing the data at hand and the accuracies of the models involved.

I carried out the required data shuffling and dropping of the 1000 data points as instructed in step 1 of the portfolio activities. The code for it can be found in my GitHub repository, which has been linked to at the end of this document. The answers to the questions are noted below.

1. Yes, the dataset did have 2 constant value columns, 'TFE Steam temperature SP' and 'TFE Product out temperature'. These were removed accordingly.
2. Yes, the dataset does have the 'Class' column, which can be a column with few integer values since it only consists of 0, 1 and 2, however this is already in a categorical order so there won't be any conversion required.
3. No, the classes did not have a balanced distribution overall. Necessary actions have been taken in the form of SMOTE to oversample the minority classes.
4. I did find composite features to create by charting out a correlation heatmap. I have selected the pairs that had co-relation higher than 0.9. The columns that I have added are 'FFTE Production solids PV * FFTE Discharge solids', 'FFTE Pump 2 * FFTE Pump 1', 'FFTE Temperature 2 - 1 * FFTE Temperature 1 - 1', 'FFTE Temperature 3 - 2 + FFTE Temperature 1 - 1' and 'FFTE Temperature 3 - 2 + FFTE Temperature 2 - 1'.

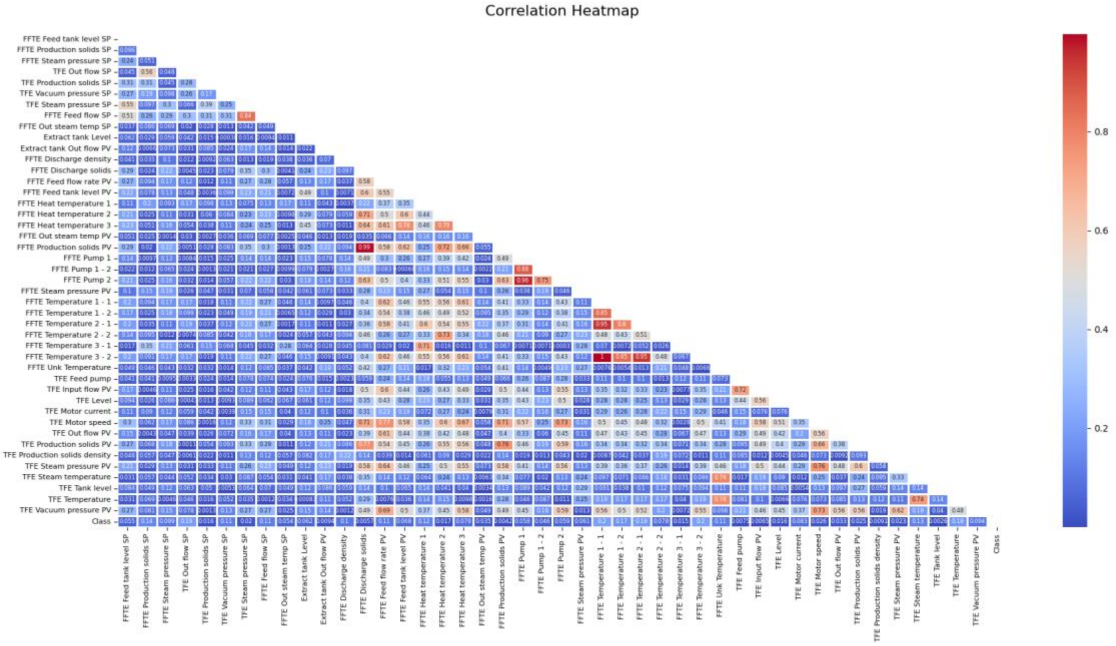

Figure 1. Correlation heatmap

5. There are 49 features in my dataset up till this point, which is not including the ‘Class’ column since it is the target variable.

After the first step, I have carried out the required feature selection and ML training processes as instructed in the step 2 of this portfolio task. The answers to the questions are noted below.

6. The training process doesn’t need all features, since using the features that have already been converted to composite features because of their high co-relativity would prove to be redundant and resource consuming. So, I have made sure to drop these features from the dataset, which include 'FFTE Production solids PV', 'FFTE Discharge solids', 'FFTE Pump 2', 'FFTE Pump 1', 'FFTE Temperature 2 - 1', 'FFTE Temperature 1 - 1' and 'FFTE Temperature 3 - 2'.
7. I have trained the data using 5 different ML models which include 'Decision Tree', 'Random Forest', 'SVM', 'MLP Classifier' and 'SGD'.
9. The comparison among all the models is shown in the form of a table as below.

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	Confusion Matrix Deductions
Decision Tree	0.97	0.97	0.97	0.97	High accuracy, balanced performance across classes.
Random Forest	1.00	1.00	1.00	1.00	Nearly perfect classification.
SVM	0.47	0.53	0.47	0.42	Significant misclassification in class 1.
MLP Classifier	0.37	0.53	0.37	0.25	Significant misclassification in classes 1 and 2.
SGD	0.44	0.47	0.43	0.39	Significant misclassification in class 0.

10. Based on the analysis of the models' performances, Random Forest is the best model to use. Random Forest achieved an accuracy of 1.00 (100%) on the test data, which means it correctly classified all the samples without any errors. Also, the confusion matrix for Random Forest shows minimal to no misclassifications. This is crucial as it suggests that the model is not only accurate overall but also consistent in correctly predicting all classes.

After the tasks in step 2, I morphed the existing data set of 1000 rows to be like the one we have been working on. I undertook the same composite feature building and feature selection techniques. The answers to the questions are noted below.

16. The performance of the Random Forest model in this case is almost identical to its performance with the actual feature set. The accuracy, precision (avg), recall (avg) and f1-score (avg) all come up to 1.00 (100%). The confusion matrix is also almost flawless.
17. The comparison among all the models is shown in the form of a table as below. From this comparison, we can deduce that the performance of the models remains almost the same for the 1000 unseen data points as well.

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	Confusion Matrix Deductions
Decision Tree	0.99	0.99	0.99	0.99	Almost as accurate at classification as Random Forest.
Random Forest	1.00	1.00	1.00	1.00	Nearly perfect classification.
SVM	0.48	0.51	0.48	0.43	Significant misclassification in class 1.
MLP Classifier	0.38	0.54	0.38	0.26	Significant misclassification in classes 1 and 2.
SGD	0.45	0.49	0.45	0.40	Significant misclassification in class 0.

Now that step 3 is completed, we can move on to step 4 of this portfolio activity. The decision tree generated in this case is too big to be included in full here, however I will have a text file in my GitHub repository folder that consists of it entirely and a link will be provided at the end of this document. However, this is the starting bit of the decision tree just for reference purposes.

```
|--- FFTE Feed flow SP <= 10199.47
|   |--- FFTE Feed flow SP <= 9230.17
|   |   |--- TFE Out flow SP <= 2168.40
|   |   |   |--- FFTE Feed tank level SP <= 49.87
|   |   |   |   |--- TFE Production solids SP <= 67.84
|   |   |   |   |   |--- FFTE Steam pressure SP <= 117.02
|   |   |   |   |   |   |--- FFTE Steam pressure SP <= 115.79
|   |   |   |   |   |   |   |--- FFTE Feed flow SP <= 8965.00
|   |   |   |   |   |   |   |   |--- TFE Vacuum pressure SP <= -58.69
|   |   |   |   |   |   |   |   |   |--- class: 2
|   |   |   |   |   |   |   |   |--- TFE Vacuum pressure SP >  -58.69
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- FFTE Feed flow SP >  8965.00
|   |   |   |   |   |   |   |   |--- FFTE Production solids SP <= 40.51
|   |   |   |   |   |   |   |   |   |--- TFE Production solids SP <= 63.75
|   |   |   |   |   |   |   |   |   |   |--- FFTE Steam pressure SP <= 95.00
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- FFTE Steam pressure SP >  95.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- TFE Production solids SP >  63.75
|   |   |   |   |   |   |   |   |   |   |--- class: 2
|   |   |   |   |   |   |   |   |--- FFTE Production solids SP >  40.51
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- FFTE Steam pressure SP >  115.79
|   |   |   |   |   |   |   |--- class: 2
|   |   |   |   |   |--- FFTE Steam pressure SP >  117.02
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- TFE Production solids SP >  67.84
|   |   |   |   |   |--- FFTE Steam pressure SP <= 110.00
|   |   |   |   |   |   |--- TFE Out flow SP <= 2015.90
| | | | | | | |--- class: 1
| | | | | | |--- TFE Out flow SP >  2015.90
| | | | | | | |--- class: 0
```

A few decision rules that can be deduced from the decision tree are as follows.

- For class 2 (TFE Vacuum pressure SP <= -58.69, FFTE Feed flow SP <= 8965.00, FFTE Steam pressure SP <= 115.79, TFE Production solids SP <= 67.84, FFTE Feed tank level SP <= 49.87 and TFE Out flow SP <= 2168.40)
- For class 1 (TFE Vacuum pressure SP > -58.69, FFTE Feed flow SP <= 8965.00, FFTE Steam pressure SP <= 115.79, TFE Production solids SP <= 67.84, FFTE Feed tank level SP <= 49.87 and TFE Out flow SP <= 2168.40)
- For class 0 (TFE Out flow SP > 2015.90, FFTE Steam pressure SP <= 110.00, TFE Production solids SP > 67.84, FFTE Feed tank level SP <= 49.87, FFTE Feed flow SP <= 9230.17)

Appendix

Link to the GitHub repository – https://github.com/rupayan-banerjee/Portfolios/tree/193c2bc82cad46e6d394857c6e4922d0f6a0ca71/Week4
Link to the Decision Tree model – https://github.com/rupayan-banerjee/Portfolios/blob/193c2bc82cad46e6d394857c6e4922d0f6a0ca71/Week4/decisionTree.txt