# CHAPTER FOUR

# RESULT AND DISCUSSION

## 4.0 INTRODUCTION

This section presents the result and discussion of findings generated from the study. Section 4.1 shows the data analysis and model training, while 4.2 presented the performance evaluation of the ML algorithms.

## 4.1 DATA ANALYSIS AND MODEL IMPLEMENTATION

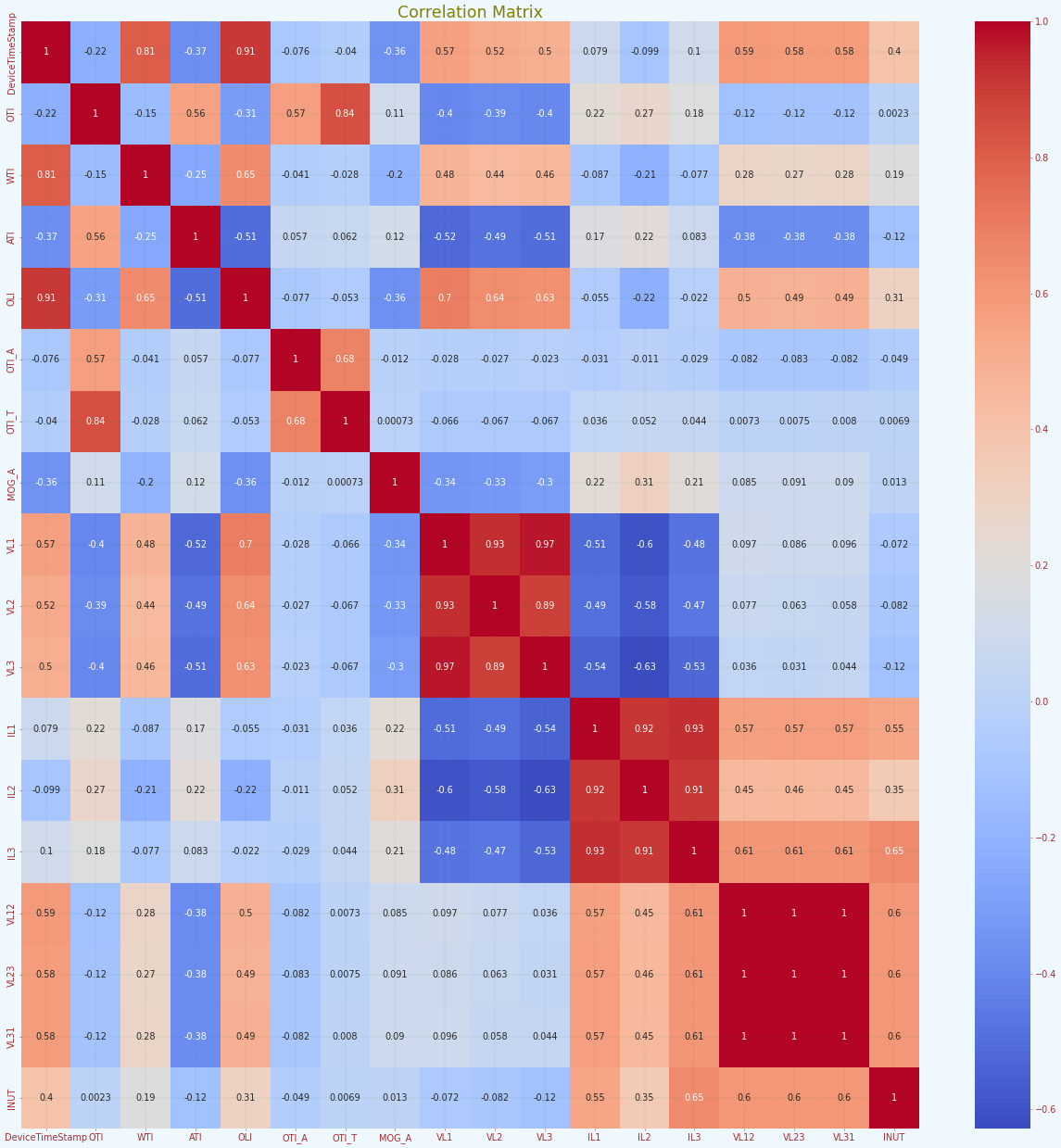
A transformer dataset containing 21174 instances and 11 attributes was extracted from a CSV file and loaded into a Jupyter Notebook environment. After data collection, an exploratory data analysis (EDA) is carried out on the dataset to  
evaluate and classify the data's key features by means of visualizations, then data cleaning and preparation is carried out before the models are implemented.

### DESCRIPTIVE STATISTICS OF DATA

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| OTI | 21174 | 30.17786 | 11.961138 | 0 | 26 | 30 | 34 | 250 |
| WTI | 21174 | 0.259564 | 0.438406 | 0 | 0 | 0 | 1 | 1 |
| ATI | 21174 | 27.740059 | 5.750776 | 0 | 24 | 28 | 32 | 44 |
| OLI | 21174 | 69.661519 | 27.79258 | 36 | 40 | 64 | 100 | 100 |
| OTI\_A | 21174 | 0.00477 | 0.068902 | 0 | 0 | 0 | 0 | 1 |
| OTI\_T | 21174 | 0.00222 | 0.047063 | 0 | 0 | 0 | 0 | 1 |
| MOG\_A | 21174 | 0.101681 | 0.302236 | 0 | 0 | 0 | 0 | 1 |
| VL1 | 20652 | 241.023455 | 9.392606 | 0 | 235.8 | 242.4 | 247.3 | 261.2 |
| VL2 | 20652 | 240.490538 | 9.784313 | 0 | 235.5 | 241.9 | 246.4 | 261.3 |
| VL3 | 20652 | 239.923107 | 8.712857 | 0 | 235.5 | 241 | 245.1 | 261.3 |
| IL1 | 20652 | 70.56517 | 42.963096 | 0 | 43.6 | 67.5 | 98.4 | 224.1 |
| IL2 | 20652 | 56.522187 | 41.311167 | 0 | 28.7 | 48.7 | 80.3 | 253.6 |
| IL3 | 20652 | 79.705825 | 45.816941 | 0 | 53.1 | 77.7 | 111.9 | 247.3 |
| VL12 | 20652 | 363.387391 | 140.051283 | 0 | 397.4 | 416.2 | 427.1 | 446.5 |
| VL23 | 20652 | 362.494984 | 139.488304 | 0 | 398.2 | 415.3 | 424.9 | 444.8 |
| VL31 | 20652 | 363.871165 | 140.200657 | 0 | 399 | 416.5 | 427.3 | 447.3 |
| INUT | 20652 | 25.170497 | 15.705378 | 0 | 15.3 | 24.8 | 35.1 | 145.8 |

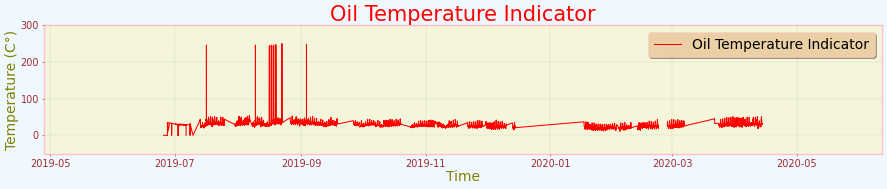
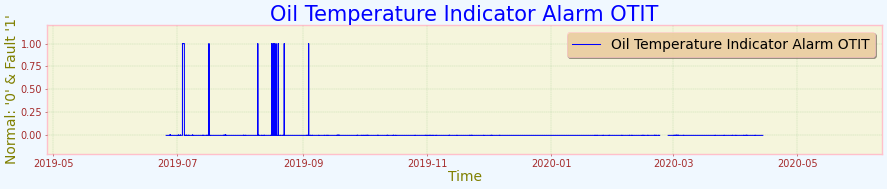
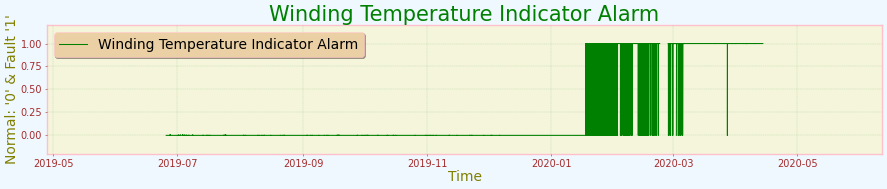
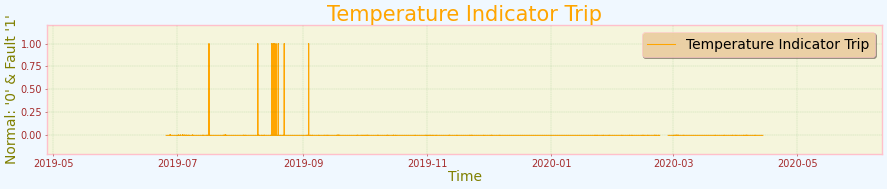
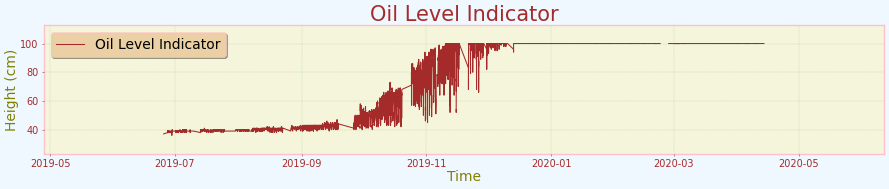
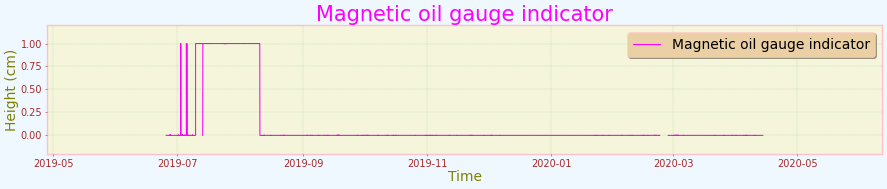
### CORRELATION MARTRIX

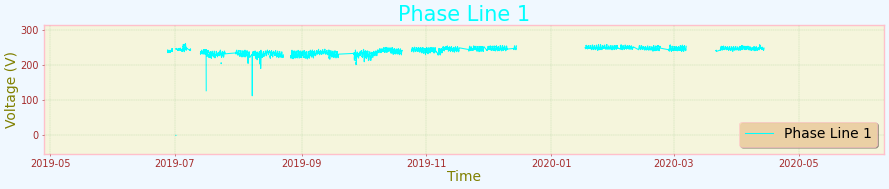
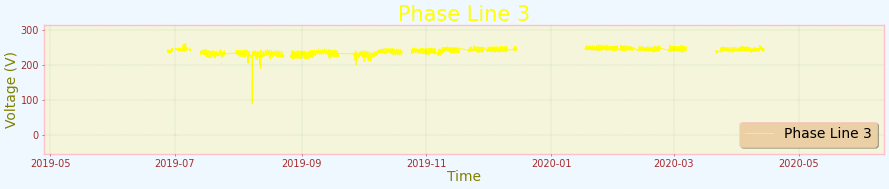
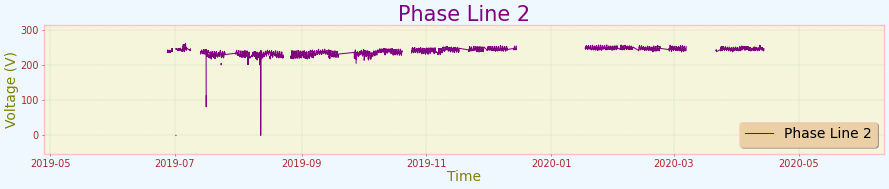
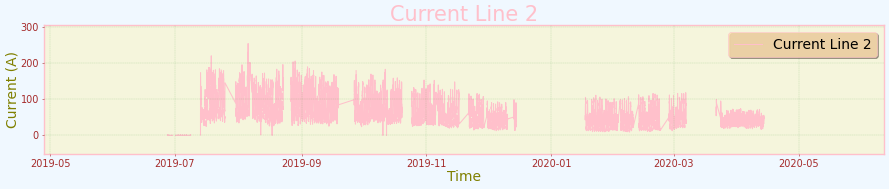
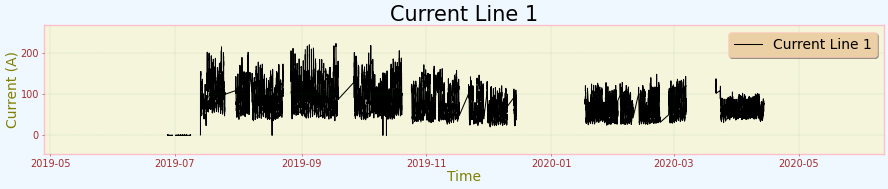
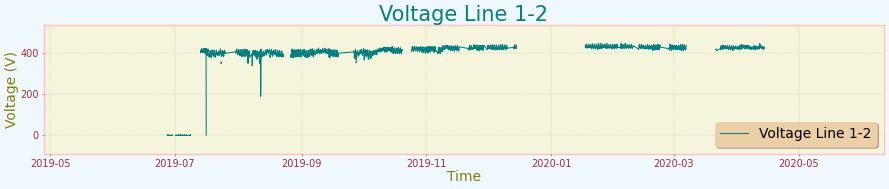
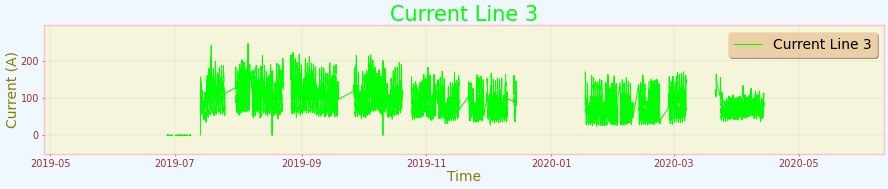
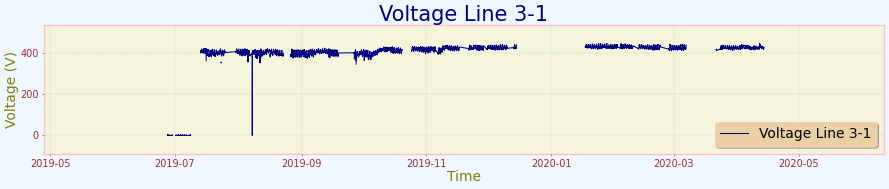
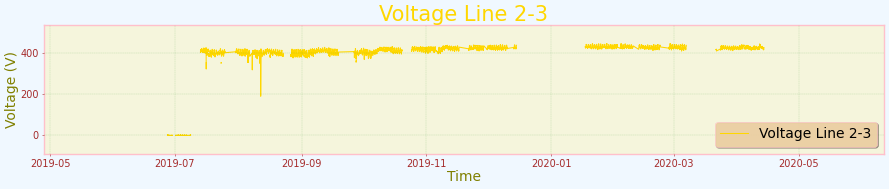
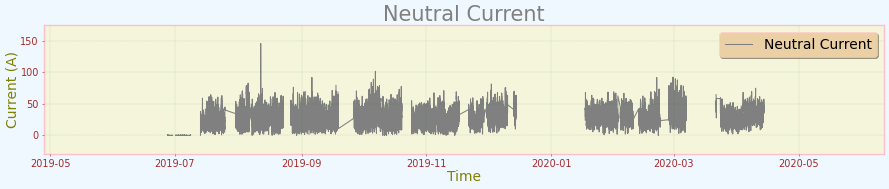
To further understand the data, we take a look to see the correlations between the features. This is done to ensure that the correlation between them is not too high and that they are suitable for machine learning algorithms and avoid overfitting or underfitting the models. Figure below shows the heatmap of the features in the dataset.



### EXPLORATORY DATA ANALYSIS

Visualizations relating to the distribution of the data to be used for the modeling are provided below.

### DATA PREPARATION

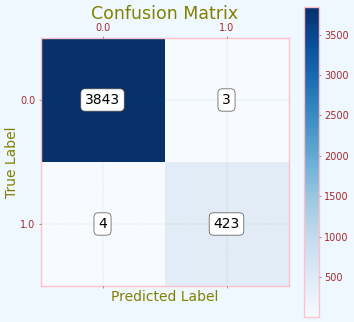
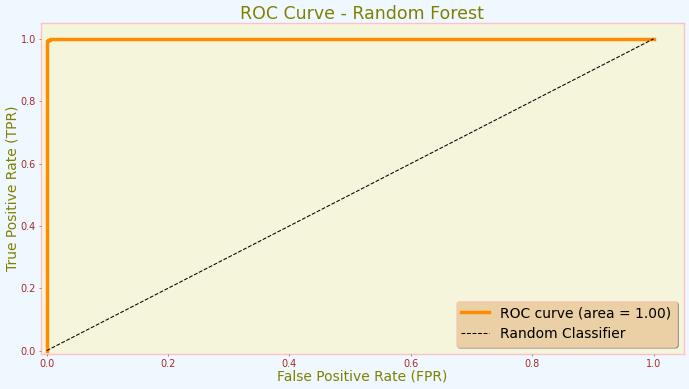
In preparing the data for machine learning, we began by examining the data types and checking for any missing values to ensure the dataset’s integrity. We then split the data into training (80%) and test (20%) subsets to facilitate robust and unbiased model evaluation. The training set was used for fitting the models and tuning hyperparameters, while the test set provided an independent assessment of model performance. To address the variance in attribute scales, we normalized the data using Scikit-Learn’s RobustScaler, which is less sensitive to extreme values, ensuring that all features were scaled between 0 and 1 for optimal model performance.

### MODEL TRAINING AND IMPLEMENTATION

The model training process involved standardizing the data, tuning hyperparameters for both XGBoost and Random Forest classifiers using grid search, and fitting the best models on the training data. Predictions were then made on the test data, and the accuracy of the models was evaluated to ensure they were well-tuned and capable of making accurate predictions on unseen data.

**RANDOM FOREST ALGORITHM**

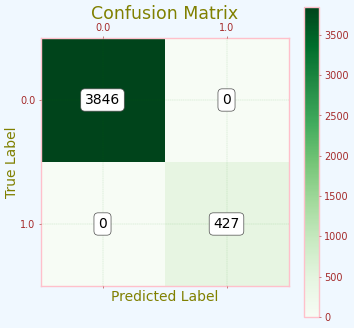
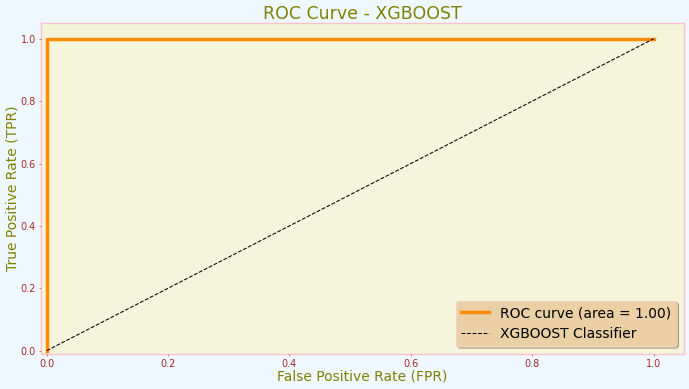
|  |  |
| --- | --- |
| **Measure** | **Value** |
| **Accuracy** | 99.8% |
| **F1 score** | 99.8% |
| **Recall** | 99.8% |
| **precision** | 99.8% |
| **Roc** | 99.99% |

The Random forest classifier had an accuracy score of 99.8% with a precision of  
99.8%. The recall for the model is at 99.8%, with an F1 score of 99.8% and roc of 99.99%.

**XGBOOST ALGORITHM**

|  |  |
| --- | --- |
| **Measure** | **Value** |
| **Accuracy** | 100% |
| **F1 score** | 100% |
| **Recall** | 100% |
| **precision** | 100% |
| **Roc** | 100% |
| **Execution time** | 50 Minutes |

The XGBOOST algorithm had an accuracy of 100% with an F1- score of 100%. The  
recall for the algorithm is at 100% and Roc of 100%.

## 4.2 PERFOMANCE EVALUATION OF THE MODEL

Accuracy, Confusion Matrix, AUROC (Area Under the Receiver Operating Characteristics), Precision, Recall and execution time are metrics used to evaluate the performance of both models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MODEL** | **PERFOMANCE SUMMARY OF ALGORITHMS** | | | | | |
| **ACCURACY (%)** | **PRECISION**  **(%)** | **RECALL (%)** | **F1-SCORE** | **ROC %** | **EXECUTION**  **TIME** |
| **RANDOM FOREST** | 99.8362 | 99.836 | 99.8362 | 99.8361 | 99.9961 | 8 min |
| **XGBOOST** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 50 min |

The table in figure above shows the performance of both algorithms. The Random Forest model achieved an accuracy of 99.8%, with precision, recall, and F1-score all hovering around 0.99836. This indicates a high level of reliability in its predictions. Additionally, the model completed its execution in a relatively short time of 8 minutes.

In comparison, the XGBoost model delivered flawless results, with an accuracy, precision, recall, and F1-score all at 100%. The ROC score was also perfect at 100%, showcasing its exceptional ability to distinguish between classes. However, this remarkable performance came with a significantly longer execution time of 50 minutes.

These results highlight the strengths and trade-offs of each algorithm in terms of accuracy and execution time. The figure below shows the ROC comparison of the models.

