**Comparative Analysis of Machine Learning Algorithms for Detecting Electrical Line Faults**

Shreyas Pakhare

**Abstract**

**This project focuses on the binary classification of faults in electrical transmission lines using machine learning algorithms based on line voltages and line currents.**

**Accurate fault detection is imperative for maintaining the robustness of power systems. Leveraging machine learning models, specifically designed for binary classification tasks, allows for efficient identification of normal and faulty conditions in transmission lines. By utilizing features extracted from line voltages and currents, these algorithms can distinguish between fault-free operation and the presence of anomalies, such as short circuits or line failures.**

**The urgency to deploy such binary classification models arises from the critical need to swiftly isolate and address faults, thereby minimizing downtime and preventing potential cascading effects on the power grid. The real-time analysis of line voltages and currents enhances the responsiveness of the system, ensuring a rapid and targeted response to abnormal conditions.**

**I. Introduction**

In the domain of electrical fault detection in transmission lines, this study conducts a comprehensive comparative analysis of four prominent machine learning algorithms: K-Means clustering, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Logistic Regression. The primary objective is to assess the efficacy of each algorithm in the binary classification of electrical faults using line voltages and line currents as input features.

K-Means clustering is employed to discern distinctive clusters within the dataset, showcasing its unsupervised learning capability in identifying fault patterns. Meanwhile, ANN capitalizes on its neural network architecture to capture complex relationships and non-linearities, offering a powerful tool for fault classification. SVM, known for its effectiveness in high-dimensional spaces, provides a discriminative framework for precise fault identification, and Logistic Regression brings a probabilistic approach, estimating the likelihood of faults based on input features.

The comparative analysis involves evaluating the performance metrics, such as accuracy, precision, recall, and F1 score, to gauge the strengths and weaknesses of each algorithm in the context of fault detection. The study aims to contribute insights into the optimal choice of algorithm for specific scenarios, thereby enhancing the understanding of the applicability and performance nuances of these methodologies in real-world electrical transmission line fault detection scenarios.

**II. Dataset and Preprocessing**

1. **Quick View at Dataset**

The dataset comprises a total of ten columns, with the first six columns containing numerical values representing line currents and line voltages in an electrical transmission system. The remaining four columns are binary indicators (0 or 1), with "G," "C," "B," and "A" denoting ground, phase C, phase B, or phase A faults, respectively. A binary value of 1 indicates the presence of the corresponding fault, while 0 signifies the absence of faults.

In the interest of simplifying the binary classification task, a new column is added to the dataset,isFault, serving as a binary indicator.

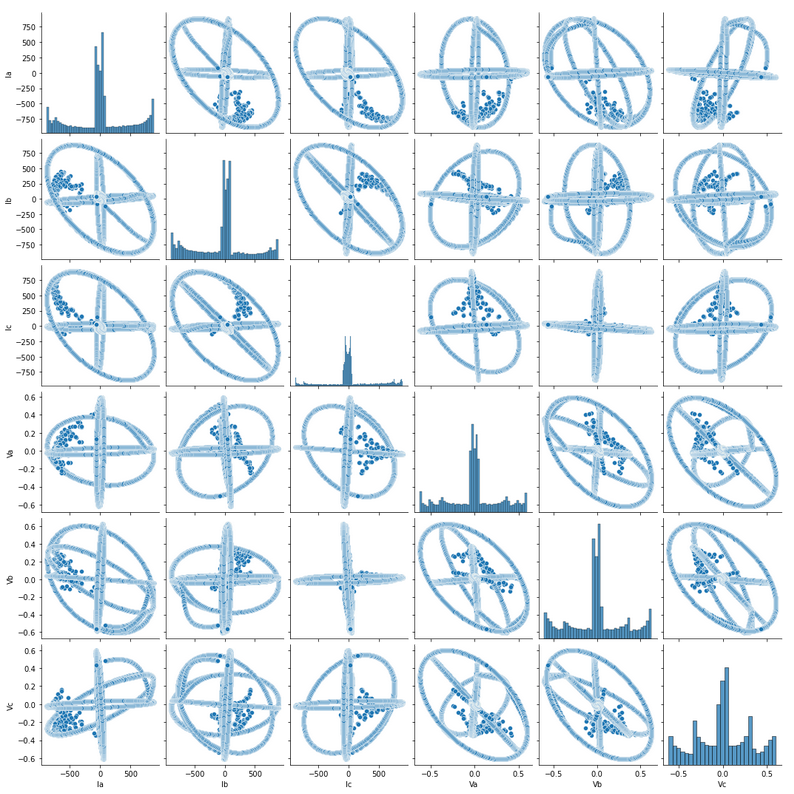
This additional column takes on the value of 1 if any of the faults labeled as G, C, B, or A is present (indicated by a 1 in the corresponding fault columns), and 0 otherwise.

This aggregated binary indicator allows for a unified classification approach, aiming to detect the presence of any fault within the system.

The inclusion of this consolidated column streamlines the binary classification task, providing a holistic perspective on fault occurrence and facilitating the development of machine learning models for efficient fault detection in electrical transmission lines.

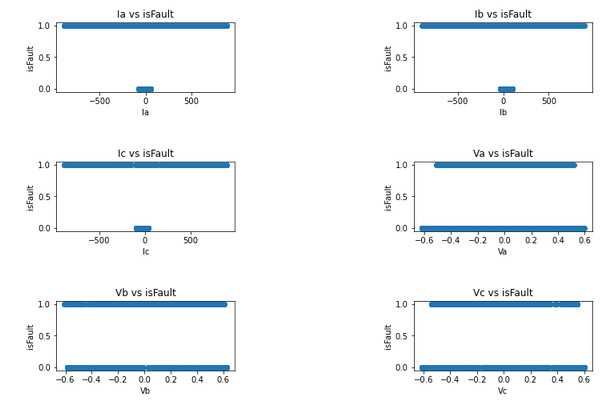
With the goal of streamlining the binary classification task, the dataset has undergone a refinement process where columns corresponding to specific fault indicators, namely G, C, B, and A, have been dropped.

1. **b. Visualize relationship between variables**



**Scatter Plot between feature variables in the dataset**

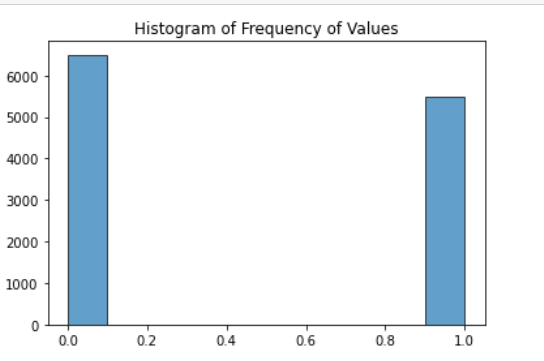
On plotting the relationship between the feature variables - line currents and line voltages, we can observe that there is no apparent causal or correlation between any two variables. Hence, all the variables are independent and need to be there to preserve the information of the dataset.



**isFault vs Feature variables**

From the plots above, it can be deduced that there is no linear relationship between the isFault variable and the features which we are using to predict the fault. Hence, as a result we can say that there doesn’t exist any direct relationship between the two and we need to move the variables into higher dimensions to learn the relations.

1. **Imbalance in the dataset**



The dataset exhibits no imbalance, mitigating the risk of errors associated with skewed class distribution. Absence of imbalance ensures that the machine learning models can effectively learn from both fault and non-fault instances, promoting robust binary classification in electrical transmission line fault detection without biased learning towards any specific class.

1. **Scaling of Data**

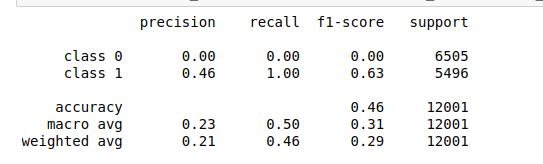
The dataset undergoes feature scaling using MinMaxScaler, ensuring that numerical values in the range of line currents and line voltages are normalized between 0 and 1. This preprocessing step enhances the performance of machine learning algorithms by preventing features with larger scales from dominating the learning process.

**III. Application of Algorithms**

1. **Logistic Regression**

The suboptimal performance of Logistic Regression on the dataset, evidenced by a low accuracy of 48%, can be attributed to the absence of a linear relationship between the variables.

Logistic Regression assumes a linear decision boundary, and when the underlying relationships in the data are non-linear or complex, the model struggles to capture the intricacies.



**Classification Report for Logistic Regression**

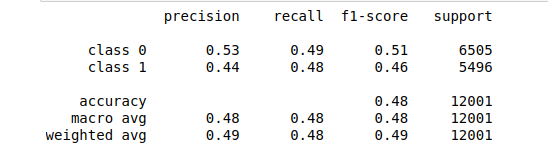
In the context of electrical fault detection, where fault patterns may not adhere to linear structures,non-linear algorithms like SVM and ANN perform better.

1. **K-Means**

The application of K-Means clustering on the dataset yielded suboptimal results, with a meager accuracy of 48%.

K-Means clustering performs poorly on non-linear datasets due to its inherent assumption of spherical or isotropic clusters.

This assumption is grounded in the algorithm's objective of minimizing the sum of squared distances between data points and the centroids of their assigned clusters.



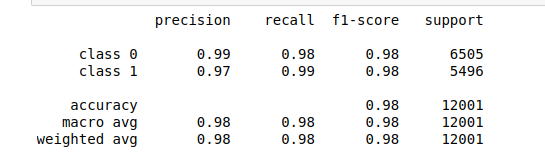
**Classification Report for KMeans**

1. **Support Vector Machines (SVM)**

The outstanding performance of Support Vector Machines (SVM) on the dataset, achieving an impressive accuracy of 98%, highlights the algorithm's capability to discern complex patterns and non-linear relationships.

SVM excels in binary classification tasks by effectively identifying the optimal hyperplane that maximally separates different classes.

This accuracy underscores SVM's adaptability to the nuanced interdependencies among line currents, line voltages, and fault indicators in electrical transmission line data. The robust performance of SVM showcases its efficacy in scenarios where the underlying relationships are intricate and non-linear, positioning it as a powerful tool for accurate fault detection in complex electrical systems.



**Classification Report for SVM**

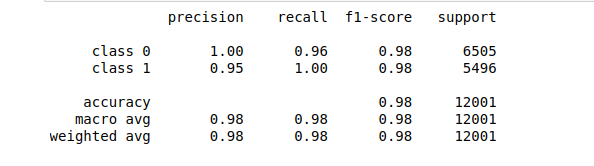
1. **Artificial Neural Network**

The impressive results obtained with Artificial Neural Networks (ANN) underscore their effectiveness as powerful tools for binary classification tasks.

Renowned for their ability to model complex relationships and adapt to non-linear patterns, ANNs excel in capturing the intricate interdependencies within the electrical transmission line dataset.

The success of ANN in achieving high accuracy reflects its capacity to learn and generalize from the numerical features, demonstrating its suitability for fault detection where relationships may be non-linear or intricate.

This positive outcome positions ANN as a valuable and versatile approach for accurate binary classification in the context of electrical fault detection.



**Classification Report for ANN**

**IV. Conclusion**

**References Used**