Task 2 – Rohit Garg

1. Objective

Develop model to Identify real time failure.

2. Design of System

Combination of supervised and unsupervised algorithms are used to identify the reasons for failure.

2.1. Supervised Learning

- **Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.**
- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
- A classification problem is when the output variable is a category, such as “erro1” or “error2” or “error2” and “no error3”.

2.2. Unupervised Learning

- **Unsupervised learning is where you only have input data (X) and no corresponding output variables.**
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
- A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

3. Supervised Learning

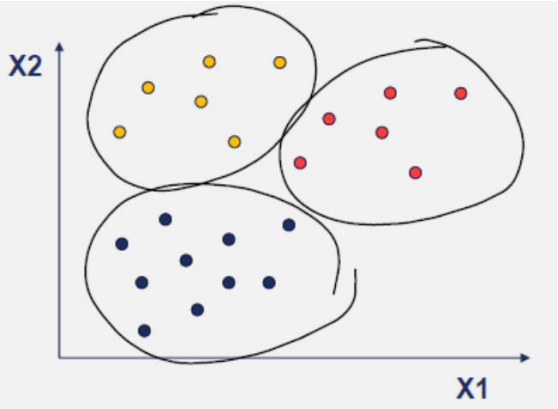
Supervised learning is used to identify the key factors or reasons behind the failure.

Gradient boosting combines a set of weak learners and delivers improved prediction accuracy. The outcomes predicted correctly are given lower weight compared to the miss-classified outcomes. The hyper-parameters of this ensemble model can be divided into 3 categories:

- **Tree-Specific Parameters:** These affect each individual tree in the model.
- **Boosting Parameters:** These affect the boosting operation in the model.
- **Miscellaneous Parameters:** Other parameters for overall functioning.

4. Unsupervised Learning

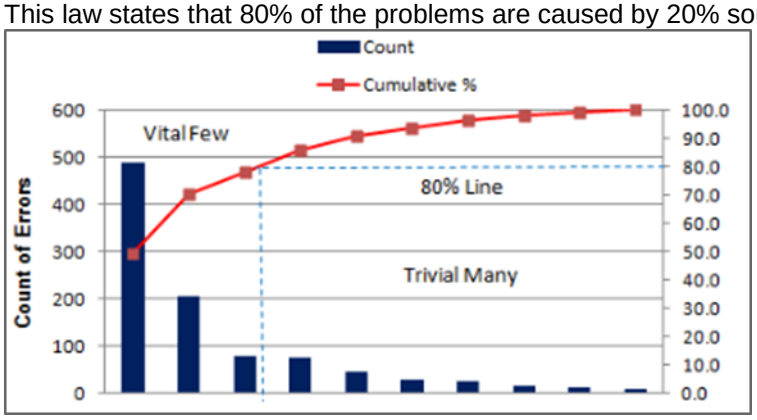
Clustering is defined as the groups of observations in terms of their characteristics. Main task of clustering is exploratory data mining. Unsupervised learning can be used to understand where the errors are clustered.



Gaussian Mixture Model is a probabilistic method for clustering. It is better to use than traditional clustering algorithms, like K-means because the probabilities allow to better evaluate edge cases. The advantages of this approach include no need to standardize data, faster to compute and the cluster sizes do not have specific structures that might or might not apply

5. Pareto's Law

This law states that 80% of the problems are caused by 20% sources.



We can utilize the insights from the above chart and address the top most reasons for failuer, hence avoiding a large numbers of failures in future.