On –Off Slip Detection

Problem Statement

- Using the data collected by sensors during a drill pipe tripping operation ,we have to detect when the slip is on or off.
- This is a binary Classification challenge, to detect when the slip is on or off.
- If ‘ 1’, Slip is on.
- If ‘0’ , Slip is off.

Dataset Description

- This dataset has 40673 rows and 10 columns.
- The dataset has the following Columns:-

  Features- Unnamed: 0,BDEP,TPO,HL,BHT,RPM,TOR,DEPT,WOB

  Label -Annotation.
- Since the RPM and TOR columns have Zero value throughout the dataset , we will drop them from our analysis.
- Convert the Timestamp Column to Datetime and convert the datetime column further into Hours, Minutes and seconds column.
- Since the time was constant through out the dataset, I have dropped the Hours, Minutes and seconds Column too from the analysis.

Methodology

The flow of the project is as follows:

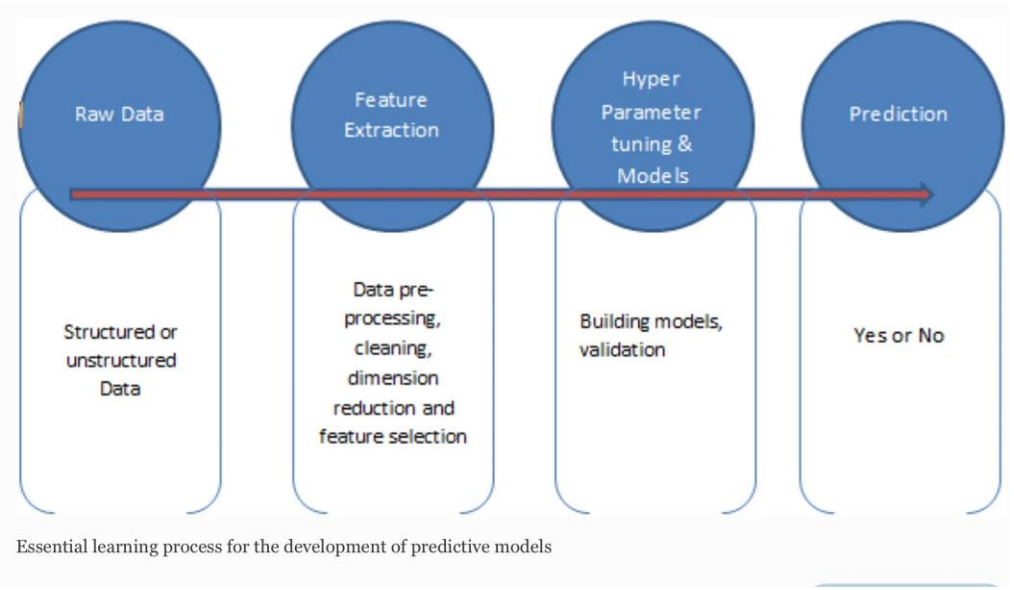
1.Data Collection

2.Data Cleaning and Pre-processing

3. Analysis and Visualization

4.Training and Testing models

5. Result Analysis

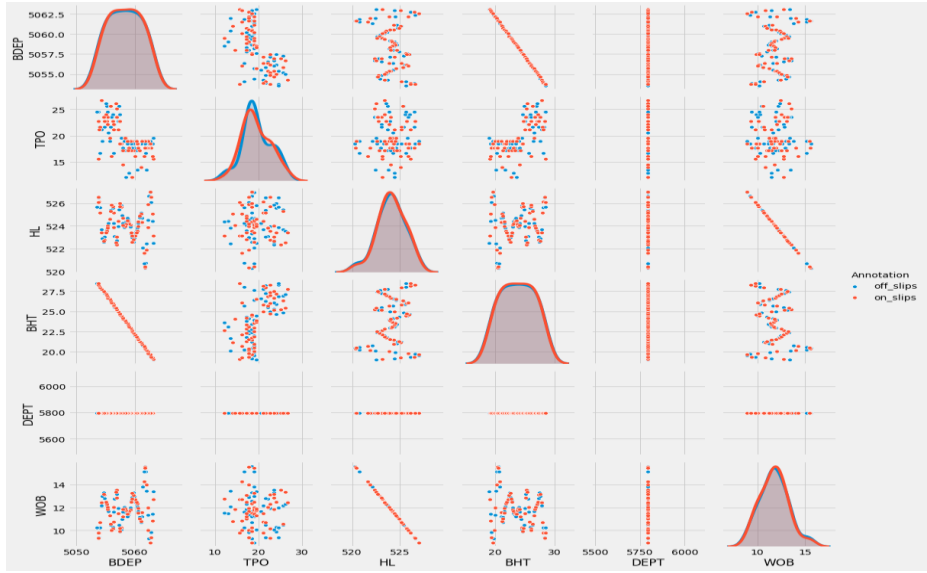


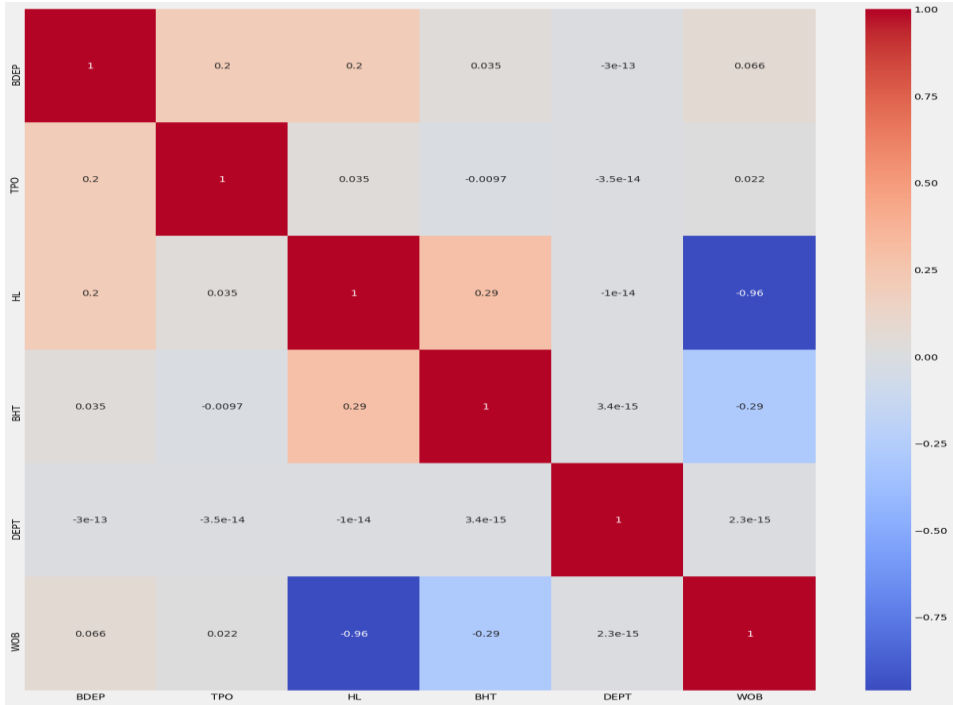
Essential learning process for the development of predictive models

Exploratory Data Analysis

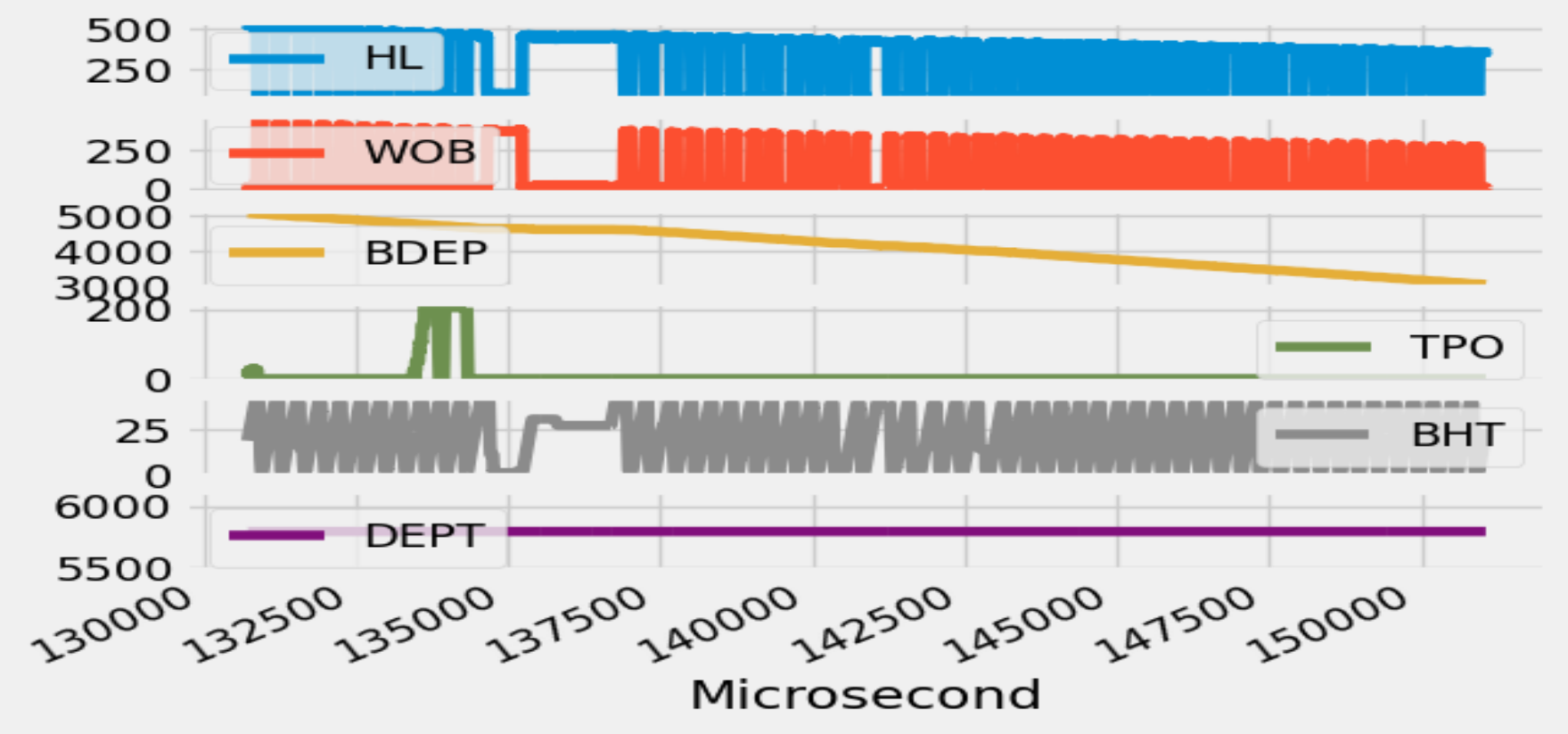
- I have explored the relationships between the variables using Heat Maps and Pair plots.

- **The plots shows that strong negative correlation between:**
    1. BDEP and BHT
    2. HL and WOB
- Hence these will be the most important variables for detecting when the slip is on or off.





Visualizing features with respect to time



Data Pre Processing

Missing Values:

- Most of the columns have a few missing values, which have been replaced by their mean Values.
- The Annotation Column has 99.73% values missing, so I have dropped the column as it does not help in data analysis.

Duplicated Values:

- 403 rows were duplicates, hence we removed the duplicate rows from the dataset.

Outliers/Noise:

- I used box plot to find outliers or noise in the dataset.
- Only TPO and DEPT columns had outliers and since they are normally distributed ,we will used quantiles method to remove the outliers.

Feature Scaling:

- Since the features have different units , I will use scaling to bring all the feature variables to the same magnitude.

Methodology

- For this project , I have used an Unsupervised Learning Algorithm like K-means clustering to find clusters , as I have dropped the target column . The Optimal clusters were found using the Elbow Method.
- Once I get the clusters , I will then use various Supervised Machine learning algorithms and compare their Performance.
- I have used the below supervised machine learning algorithms for classification because of their superior performance and fast execution:

  1.Support Vector Machines(SVM)

  2.Random Forest(RF)

  3.K-Nearest Neighbour(K-NN)

  4.Decision trees

  5. Artificial Neural Networks

  6. XG Boost
- The dataset was randomly split (70:30) into training and test sets. The models were trained on the training set and its predictive performance on the test set was then evaluated.
- Since the dataset is unbalanced, Accuracy is not a good metric to evaluate the model Performance, so I will evaluate the algorithms on Precision, Recall,F1-score,AUC-ROC curve.
- The best model will be chosen.

K-Means Clustering

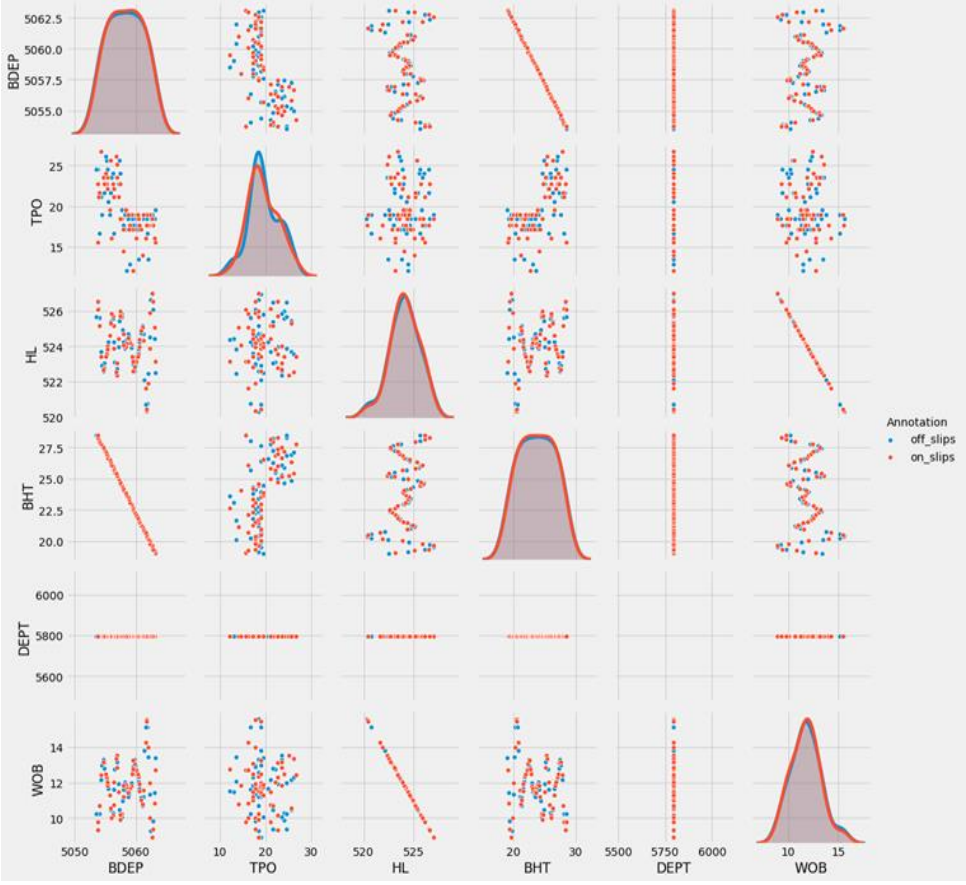
- The K Means algorithm aims to assign the data points in your dataset to K distinct clusters. After running the algorithm each observation (data point) will belong to the cluster whose centre it is closest to.
- I have used K-means++ initialization which generally produces better results then random initialization.
- Using the elbow method the optimal number of clusters were found to be 2.
- Using K-Means clustering I got 2 clusters

  Cluster 0: 27574,

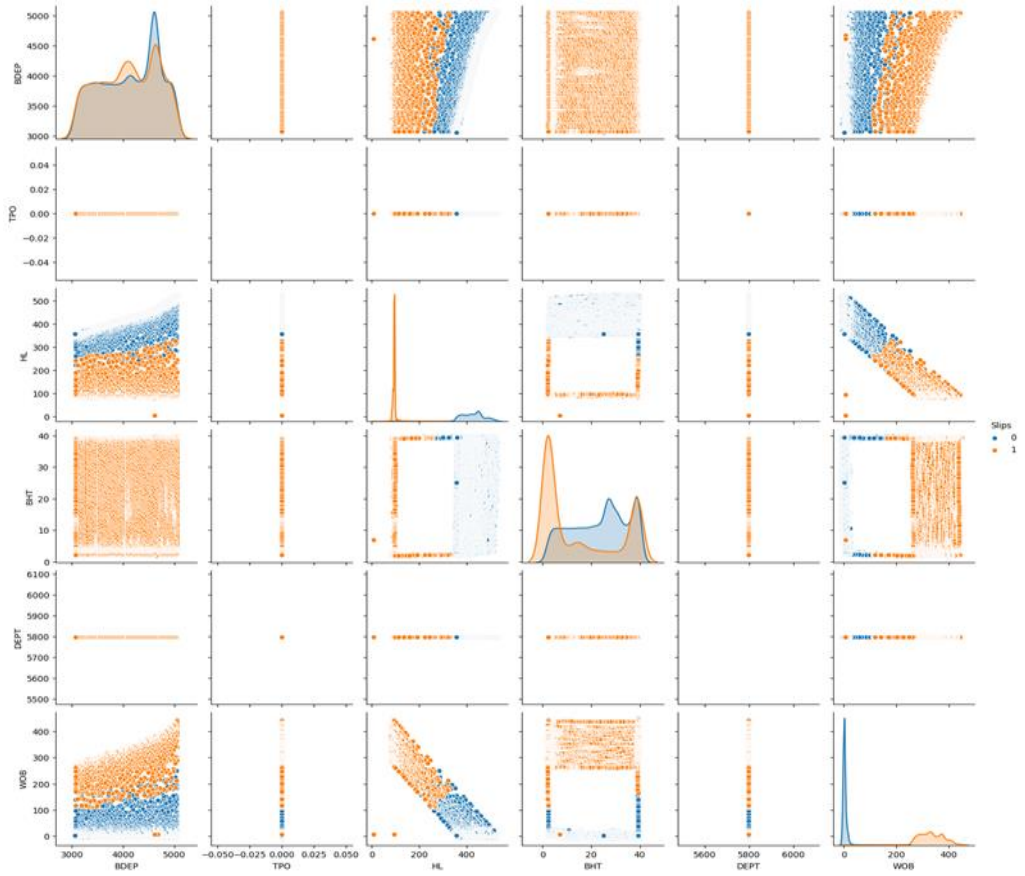
  Cluster1: 12696

K-Means Clustering Visualizations

Before K-Means Clustering



After K-Means Clustering



Supervised Classification Algorithms

- I have now added the output labels of the KMeans clustering to the original Dataset.
- Since now I have the labelled output column , I have now used various Supervised Classification Algorithms to detect when the slip is on or off.
- The Supervised Classification Algorithms used are:
- 1.Support Vector Machines(SVM)
- 2.Random Forest(RF)
- 3.K-Nearest Neighbour(K-NN)
- 4.Decision trees
- 5. Artificial Neural Networks
- 6. XG Boost
- The dataset was randomly split (70:30) into training and test sets. The models were trained on the training set and its predictive performance on the test set was then evaluated.
- Since the dataset is unbalanced, Accuracy is not a good metric to evaluate the model Performance, so I will evaluate the algorithms on Precision, Recall,F1-score,AUC-ROC curve

Conclusion

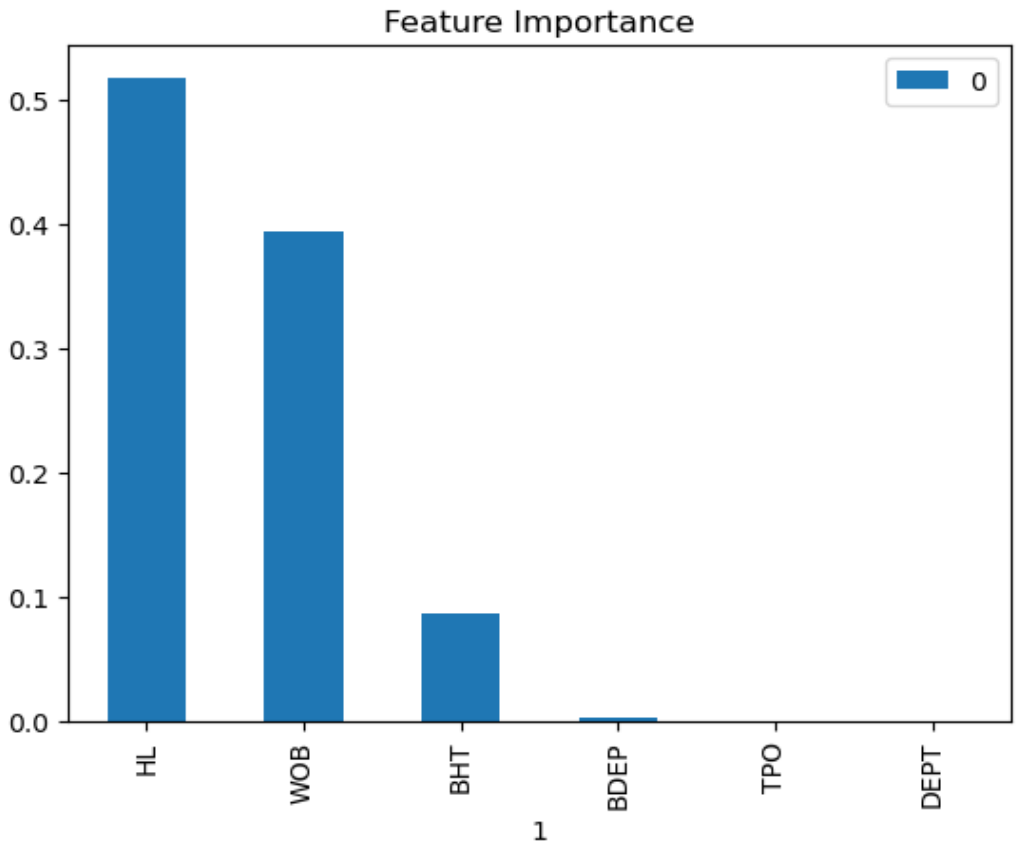
Classifier	Precision	Recall	F1-score	Misclassification Rate
Decision Tree	0.99	0.99	0.99	17
Random Forest	0.99	0.99	0.99	3
XG Boost	0.99	0.99	0.99	4
Support Vector Machines	0.99	1	0.99	3
Artificial Neural Network	0.99	1	0.99	25
LSTM	0.99	0.99	0.99	25

Conclusion

- Results of all models with their optimal parameters are provided in the table:
- The models were evaluated using Precision,Recall,F1_Score,performance evaluation metrics to determine their efficiency and quality.
- All the algorithms performed extremely well, but I found that the best performance was achieved Random Forest Classifiers as the misclassification rate was lowest .

Random Forest Feature Importance

- Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction.
- Fig below shows that HL,WOB,BHT are the most important features in detecting when the slip is on or off.



Strengths and Weaknesses of the models

Decision Tree

Advantages: Interpretability, no need for feature scaling, works on both linear / non – linear problems.
Disadvantages: Poor results on very small datasets, overfitting can easily occur.

Random Forest

Advantages: Powerful and accurate, good performance on many problems, including non – linear.
Disadvantages: No interpretability, overfitting can easily occur, need to choose the number of trees manually.

Support Vector Machines

Advantages: SVM's can model non-linear decision boundaries, and there are many kernels to choose from. They are also fairly robust against overfitting, especially in high-dimensional space.
Diadvantage: SVM's are memory intensive, trickier to tune due to the importance of picking the right kernel, and don't scale well to larger datasets.

KNN

Advantage: Simple to understand, fast and efficient.
Disadvantage: Need to manually choose the number of neighbours ‘k’.

Artificial Neural networks

Advantages: perform very well on image, audio, and text data, and they can be easily updated with new data using batch propagation. Their architectures (i.e. number and structure of layers) can be adapted to many types of problems, and their hidden layers reduce the need for feature engineering.
Disadvantages: Deep learning algorithms are usually not suitable as general-purpose algorithms because they require a very large amount of data. In fact, they are usually outperformed by tree ensembles for classical machine learning problems. In addition, they are computationally intensive to train, and they require much more expertise to tune (i.e. set the architecture and hyperparameters).

Appendix

Brief description of the Supervised machine learning algorithms used for classification:

1.Support Vector Machines(SVM)- We find a hyperplane that discriminates between two classes by maximising the margin between the classes. The instances that are the closest to the hyperplane are called support vectors. In this project I have used SVM with a linear kernel.

2.**Random Forest(RF)** method constructs each tree from a bootstrap sample drawn with replacement from the original dataset. In this project , I have used criterion as Gini, maximum depth=6 and random state=23.

3.K-Nearest Neighbour(K-NN) is a non-parametric**,** lazy learning algorithm. It classifies new cases based on a similarity measure (i.e., distance functions).In this project, I have used K=14.

4.Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node. In this project , I have used criterion as entropy, maximum depth=4 and maximum leaf nodes=10.

5.**Artificial neural network (ANN**) is a computational model that consists of several processing elements that receive inputs and deliver outputs based on their predefined activation functions.

6.**XGBoost-** is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

Appendix

Evaluation Metrics

Confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It is a table with four different combinations of predicted and actual values in the case for a binary classifier.

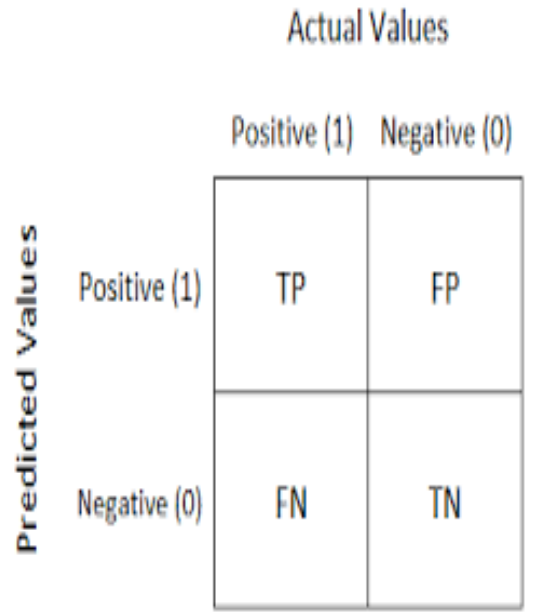
- Accuracy measures the fraction of correct predictions made. It is the ratio of the number of correct predictions over the total number of predictions.
- Precision measures a classifier's ability to not falsely label a positive document as negative. It is the ratio of true positives over all positive predictions.
- Recall is the measure of the success in retrieving all positive samples. It is the ratio of the number of correct predictions made over the total number of positively labelled documents .

Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$



Thank You!