**Fake -News Detection Project**

**Problem statement:**

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

In this project, you are given a dataset in the fake-news\_data.zip folder. The folder contains a CSV files train\_news.csv and you have to use the train\_news.csv data to build a model to predict whether a news is fake or not fake. You have to try out different models on the dataset, evaluate their performance, and finally report the best model you got on the data and its performance.

**Data- Description:**

There are 6 columns in the dataset provided to you. The description of each of the column is given below:

“id”: Unique id of each news article

“headline”: It is the title of the news.

“news”: It contains the full text of the news article

“Unnamed:0”: It is a serial number

“written\_by”: It represents the author of the news article

“label”: It tells whether the news is fake (1) or not fake (0).

**PROCEDURE**

1. The very first step to start the project is to import the data set. we have successfully imported the dataset using pandas library .
2. This dataset is about a news article which has 5 feature id , headline , news , unnamed , written by and label , where label 1 tells that news is fake and 0 tells that it is not fake.
3. This dataset by looking into it reflects that it text dataset and will require NLP approach.
4. This dataset has  2 and 3 feature which belongs to int and categorical respectively

**EXPLORATORY DATA ANALYSIS**

We have performed some data visualization on this dataset in order to be more familiar with the dataset which in turn helps to get the hidden pattern of the dataset.

1] Null Value Analysis – Headline, written\_by , news do contain some Nan values . There are 558, 1957,39 values are null/Missing from the feature headline, written\_by, news respectively

We have tried to replace Nan value with single space . because we want to combine the headline and news feature in order to have all the information in the same column that is on news and also to treat null values for both the feature.

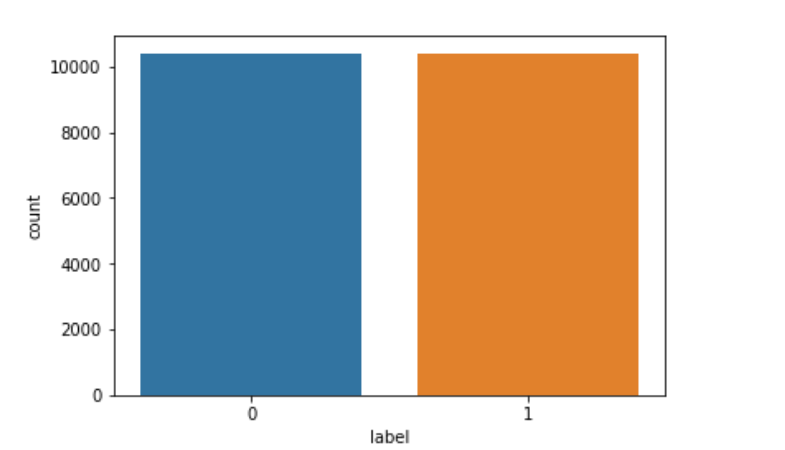
Note :- As we know Nan+ something will result in Nan , therefore Replacing Nan with space in News , headlines feature so that we can be able to combine these two feature .

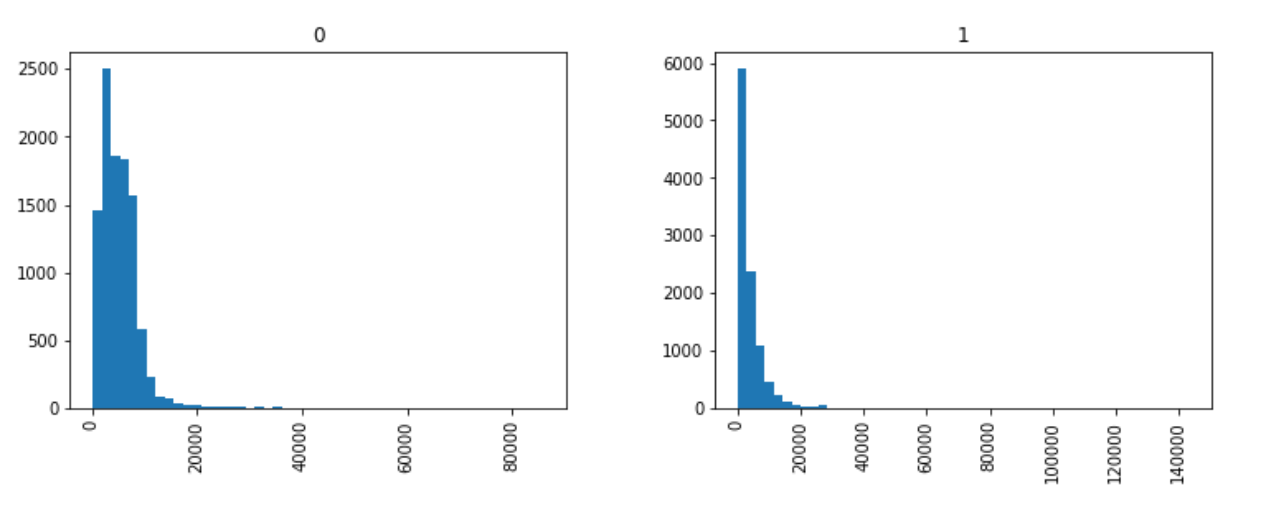
Now after combining we do not find any missing values in news and hence our news feature is now ready for further preprocessing.

2]Univariate Analysis-

We have generated a new feature called Length of the news which will help us identify that how much length generally a fake news have.

We have performed techniques like count plot, histogram plots to check the trend and pattern shown in the below figure:-





From the above count plot it can be said that or dataset is balanced and from the above histogram plot its said that The count of of fake news of length < 20k is more , which means the fake news article generally tries to be short ,as the frequency is very high at about 6000 for the

fake news.

**PREPROCESSING**

In the Pre-processing phase we did the following in the order below:-

a) Begin by removing the html tags

b) Remove any punctuations or limited set of special characters like , or . or # etc.

c) Check if the word is made up of english letters and is not alpha-numeric

d) Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)

e) Convert the word to lowercase

f) Remove Stopwords

After performing the above our cleaned data that is news is now stored in the variable “**final\_preprocessed\_news”**

**Keypoints :-**

1. Since the news has text, we will be required to perform NLP techniques in order

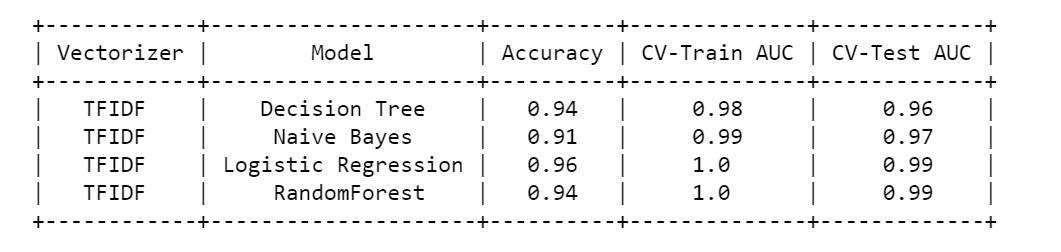
perform Models on it. Therefor we have used Tfidf for this project.

1. We have first split the data using randome state 42 and then performed Tfidf on it in order to avoid data leakage.

**Modelling**

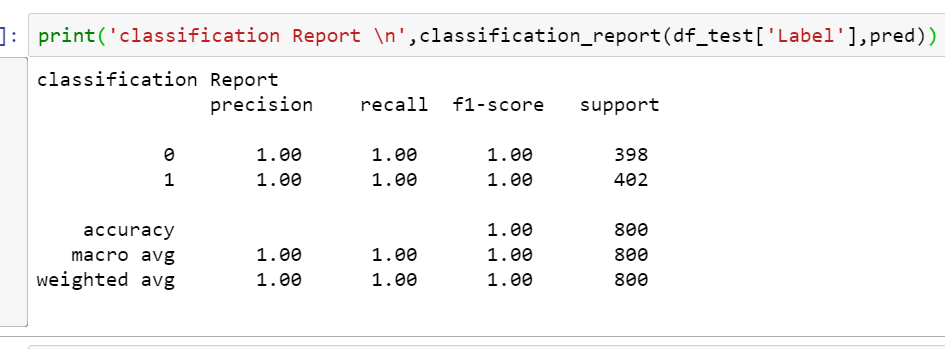
In this section we will discuss the model that we have performed on this dataset , its performance matrix , confusion matrix and at last this section will also show which model is considered best for this model and based on what criteria .Find below table for the each model performance result:-

Note:- We have used grid search cv in each and every model.



After observing the above statistics, Logistic Regression comes out to be our best model as the combination of Accuracy , AUC train , AUC test is convincing

Therefore, we will be taking Logistic Regression as our best model and will use for prediction for the unseen data. Below are the performance matrix statistics:-



Our Final Data Frame for actual and predicted class :-

