A Project Report

on

Fake News Detection Using Machine Learning

Submitted in partial fulfillment of the
requirement for the award of the degree of

Bachelor Of Technology

GALGOTIAS UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of

Dr. S. Rakesh Kumar
Assistant Professor

Submitted By

S. No	Enrollment Number	Admission Number	Student Name	Degree / Branch	Sem
1	19021011496	19SCSE1010310	ANJALI KUMARI	B-Tech / CSE	V
2	19021011438	19SCSE1010249	UTKARSH RAJ	B-Tech / CSE	V

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
December, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"FAKE NEWS DETECTION USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the Bachelor of Technology-submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of July, 2021 to December 2021, under the supervision of Mr. S. Rakesh Kumar Assistant Professor, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Anjali Kumari (19SCSE1010310)

Utkarsh Raj (19SCSE1010249)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. S. Rakesh Kumar

(Assistant Professor)

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Anjali Kumari (19SCS1010310) and Utkarsh Raj (19SCSE1010249) has been held on _________________ and his/her work is recommended for the award of Bachelor of Technology~~.~~

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

Abstract

The fake news on social media and other various platform is widely spreading rumours and is a matter of serious concern as it is having ability to cause major problems for social and national damage with destructive impacts. A lot of research is going on just to focus on detecting it. This project makes an analysis of research related to fake news detection and can explore the tradition of machine learning models to choose the best in the order to create a model of the product with supervised machine learning algorithm, that can classify or detect fake news whether its true or false, by using tools like python scikit-learn, NPL for textual analysis. This process will result in for the extraction of the features and vectorization; we are supposed to use python scikit-learn library for performing tokenization and extraction for the feature of text data, as this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Then we will perform feature selection methods for experiment and choose the best fil features to obtain the highest precision, according to confusion matrix results.

TABLE OF CONTENTS

Topics **Page number:**

Declaration ii

certificate iii

Abstract iv

Table of content v

Chapter 1: Introduction about Project 1

1.1 Introduction

1.2 Literature Review

1.3 Problem Formulation

Chapter 2: System Analysis, Required tools and Libraries 4

2.1 System Analysis

2.2 Required tools and libraries

Chapter 3: Proposed system & Methodology 5

3.1 Proposed System

3.2 Methodology

Chapter 4: Design Architecture 8

Chapter 5: Implementation 9

5.1 Tools and libraries for implementation

5.2 Algorithms used for classification

Chapter 6: Future scope and Conclusion **14**

6.1 Future scope

6.2 Conclusion

References

List Of Figures

Figure No.	Title	Page No.
1.	Proposed system methodology	6
2.	Classification Algorithm	7
3.	Architecture Diagram	8

Chapter: 1

Introduction about Project

1.1) INTRODUCTION

The appearance of the World wide web and the expenditious adoption and uses of social media platforms (such as Fakebook, Twitter, Instagram and so on) cover the way for information spreading that has never been observed in the history of human before. Besides other use cases, new outlets gain from the universal use of social media platform by providing updates on the news for the real time to its subscribers. The news that are developed from newspapers, tabloids, and magazines to a digital world such as online news platforms, blogs, social media feeds and many other digital platforms or formats. Now-a-days it is becoming more and more easier for the consumers to obtain the latest news on their fingertips. On Facebook there are total 70% of news webpages or websites. These kinds of social media platform in their current condition are exceedingly powerful and are very useful for their capacity to allow the users to discuss about the ideas and to share the ideas ant to debate the issues such as Democracy, Education, Health and many more serious issues. However, these kinds of platforms are also been used with negative viewpoints by certain entities for commonly financial gain.

And as in other cases for creating biased judgements, manipulating mindsets, and spreading irony or ridiculousness. This phenomenon is known as fake news.

This paper put forward a methodology just to create a model that will determine or detect whether a given article is genuine or is fake based on its words, title, paragraph, phrases and sources, simply by applying machine learning algorithms on an explained dataset, which are manually classified and guaranteed. Then, feature selection methods are applied on the experiment to choose the best fit features to obtain the highest accuracy, according to confusion matrix results. We suggested to create a model by using different classification algorithms. Then the product model later on will be test the unseen data, the result will be plotted, and then accordingly the product will be model that will detect and classify fake articles and can be used and unified with any system for the future uses.

1.2) LITERATURE SURVEY

In this paper, we intends to introduce an understanding of portrayal of report in the advanced diaspora joined with the differential substance kinds of report and its effect on performance of users. Therefore, we plunge into existing phony news location moves toward that are vigorously founded on text-based investigation, and furthermore depict famous phony news datasets. We finish up the paper by recognizing 4 key open exploration challenges that can direct future examination. It is a hypothetical Approach which gives Illustrations of phony news location by dissecting the mental factors. This approach was executed as a product framework and tried against an informational index of Facebook news posts. They were gathered from three huge Facebook pages each from the right and from the left, just as three enormous standard political news pages. find caustic remark than faux News. Rubin et al used a little dataset with solely 290 coaching articles and ninety take a look at articles. As a classifier they use a Support Vector Machine that is like minded for binary classification however not for multiclass classification.

1.3) Problem Formulation

The multiplication of phone news via web-based media and web is misleading people to a degree that has tobe halted. The principle objective is to get a model which will segregate between "counterfeit" or " valid " news stories whenever it's prepared with certain datasets. Fake news (or data) can pose many dangers to our world. Imagine what happens if due to some false information you are given the wrong medicine. Luckily, this problem can be addressed using machine learning. We can develop a machine learning model in python which can detect whether the news is fake or not.

CHAPTER: 2

System Analysis & Required tools and Libraries

2.1) System Analysis

- **Hardware Requirements**
  - Processor – i3
  - Hard Disk – 5 GB
  - Memory – 1GB RAM

- **Software Requirements**
  - Windows Xp,
  - Windows 7(ultimate, enterprise)

    Visual studio 2010

2.2) Required Tools and Libraries

- Jupiter Notebook
- Pandas – 1.2.4
- Scikit-learn – 0.24.1
- NLTK – 3.6.2

To install above modules please run the following command:
pip install pandas scikit-learn nltk

CHAPTER: 3

Proposed System & Methodology

3.1) Proposed system

In this PROJECT, we will fundamentally zero in on the text-based news and attempt to assemble a model that will assist us with deciding whether a piece of the given news is genuine or not. To manage the location of the news whether or not it is genuine, we will foster an undertaking in the python with the utilization of sk-learn, and we will be likewise going to utilize TfidfVectorizer in our all dataset (Set of information as text-based news). Where that will be accumulated from online media or website. At the point when our initial step is done, we will again introduce the classifier, change and pick best fit model. Eventually, when each progression has been done, we will work out the general presentation of the model utilizing legitimate execution grid or frameworks. Finally, we will actually want to see that our general model is turned out great.

3.2) Methodology

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by pre-processing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. Figure describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named “Random forest” other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision.

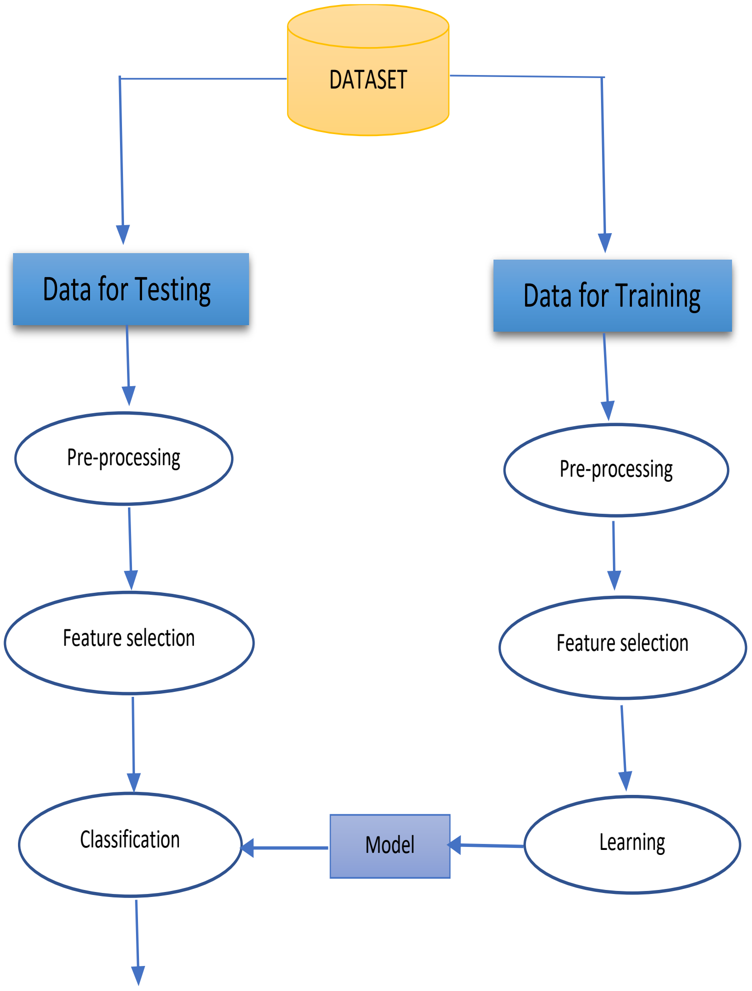


Figure 1: Proposed system methodology

The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data. Also, appropriate re-factorings have been performed on the Python code to produce an optimized code. The classification algorithms applied .in this model are Logistic Regression, Gradient Boosting, Decision Tree, Random Forests and. All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them. As shown in the figure, the dataset is applied to different algorithms in order to detect a fake news. The accuracy of the results obtained are analysed to conclude the final result.

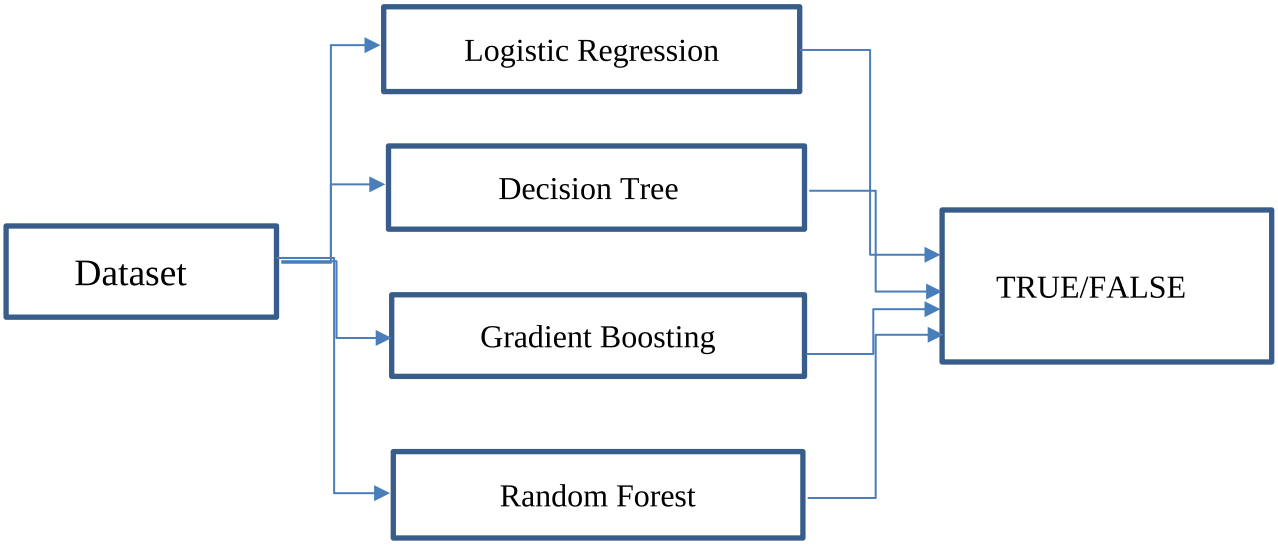


Figure 2: Classification Algorithm

CHAPTER 4:

Design Architecture

The architecture of Static part of fake news detection system is quite simple and is done keeping in mind the basic machine learning process flow. The system design is shown below and self- explanatory. The main processes in the design are:

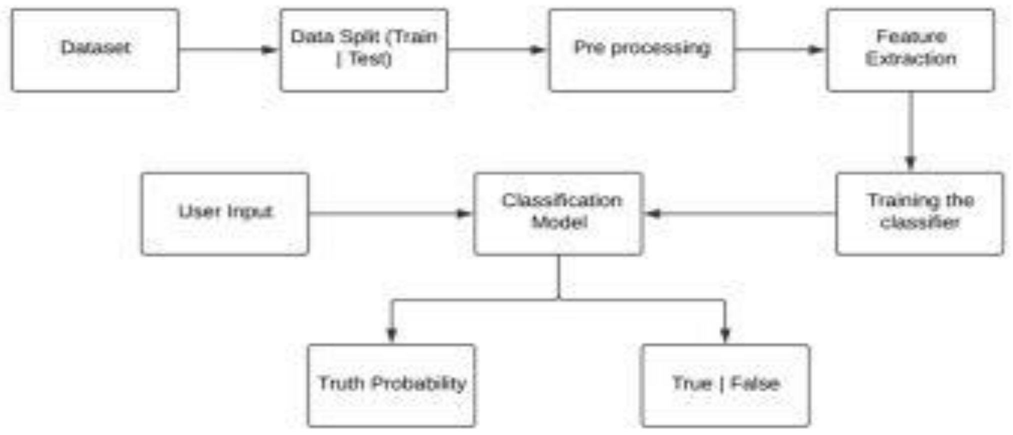


Figure 3: Architecture Diagram

CHAPTER: 5

Implementation

5.1) Tools And Libraries for Implementationn

In this project, the libraries that are going to be used are listed below:

- Python-3
- Pandas-1.2.4
- Scikit-learn(sk-learn)-0.24
- NLKT-3.6.2

Importing libraries:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

Fake News Dataset

The dataset for this python project contains two directories inside it. One contains the true (correct) news and the other contains the fake news. Summing both of them up, the total dataset comprises 44,898 instances. The real news and the fake ones are reported into two csvs.

Vectorizing Data: Vectorizing is the process of encoding text as integers i.e. numeric form to create feature vectors so that machine learning algorithms can understand our data.

Tfidf Vectorizer

We spilt the term Tfidf into two parts as TF and IDF.

TF (Term Frequency): The number of times a word seems in an exceedingly document is its Term Frequency. the next worth means that a term seems additional usually than others, and so, the document may be a smart match once the term is a component of the search terms.

IDF (Inverse Document Frequency): Words that occur repeatedly in a document, however additionally occur repeatedly in several others, however vital a term is within the entire corpus.

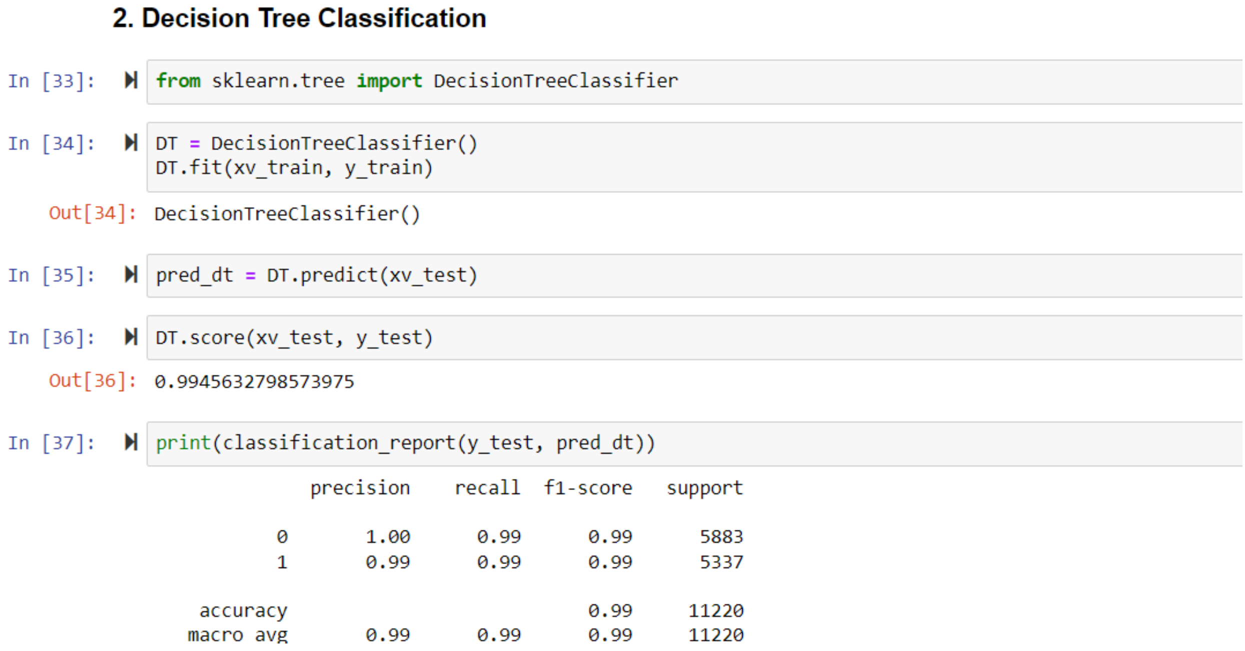
The TfidfVectorizer converts a group of raw documents(data) into a matrix of TF-IDF options.

Passive Aggressive Classifier

Passive Aggressive algorithms area unit on-line learning algorithms. Such an algorithmic program remains passive for an accurate classification outcome, and turns aggressive within the event of a misestimation, change and adjusting. In contrast of most alternative algorithms, it doesn't converge. Its purpose is to form updates that correct the loss, inflicting little or no modification within the norm of the burden vector.

5.2) Algorithms used for Classification

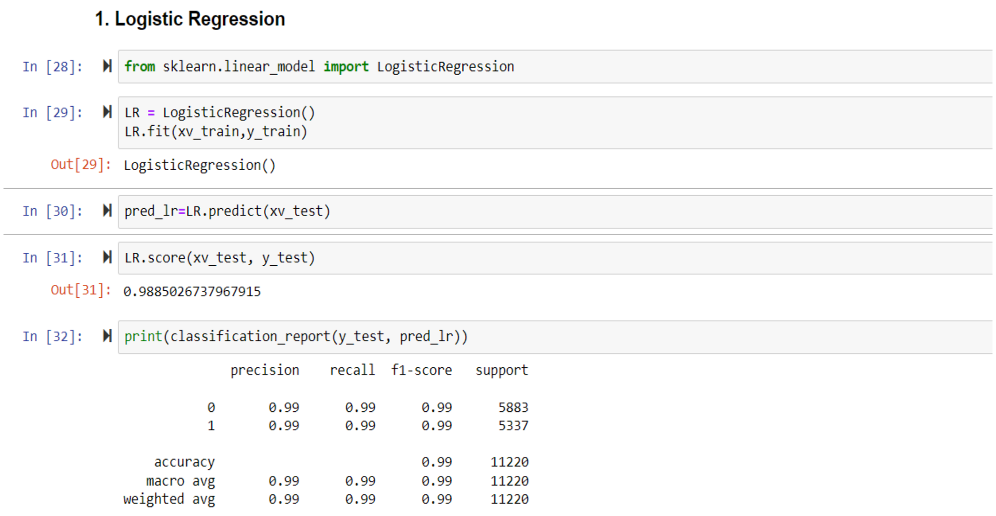
Decision Tree: The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a “test” on an attribute and the branching is done on the basis of the test conditions and result. Finally, the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation between the variables quite aptly. They are significant in creating new variables and features which is useful for data exploration and predicts the target variable quite efficiently.



2. Decision Tree Classification

```
In [33]:  from sklearn.tree import DecisionTreeClassifier

In [34]:  DT = DecisionTreeClassifier()
          DT.fit(xv_train, y_train)

Out[34]:  DecisionTreeClassifier()

In [35]:  pred_dt = DT.predict(xv_test)

In [36]:  DT.score(xv_test, y_test)

Out[36]:  0.9945632798573975

In [37]:  print(classification_report(y_test, pred_dt))

                        precision    recall  f1-score   support

                   0       1.00      0.99      0.99      5883
                   1       0.99      0.99      0.99      5337

            accuracy                           0.99     11220
           macro avg       0.99      0.99      0.99     11220
```

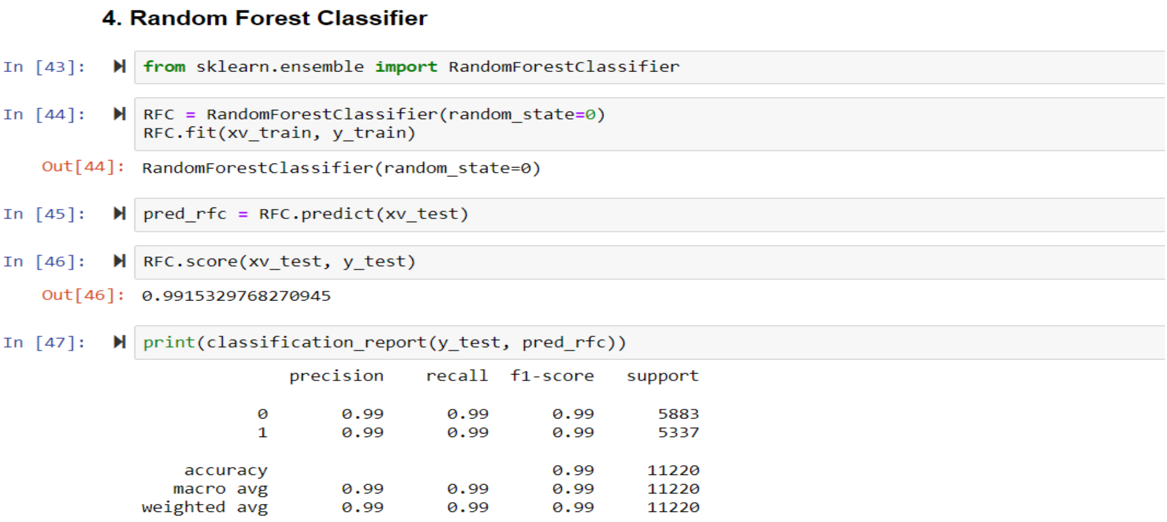
Logistic Regression: It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variables). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected). Mathematically, the log odds of the outcome are modelled as a linear combination of the predictor variables. Odds = p/(1-p) = probability of event occurrence / probability of not event occurrence ln(odds) = ln(p/(1-p)) logit(p)=ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk



1. Logistic Regression

```
In [28]:  from sklearn.linear_model import LogisticRegression

In [29]:  LR = LogisticRegression()
          LR.fit(xv_train,y_train)

Out[29]:  LogisticRegression()

In [30]:  pred_lr=LR.predict(xv_test)

In [31]:  LR.score(xv_test, y_test)

Out[31]:  0.9885026737967915

In [32]:  print(classification_report(y_test, pred_lr))

                        precision    recall  f1-score   support

                   0         0.99      0.99      0.99      5883
                   1         0.99      0.99      0.99      5337

            accuracy                             0.99     11220
           macro avg         0.99      0.99      0.99     11220
        weighted avg         0.99      0.99      0.99     11220
```

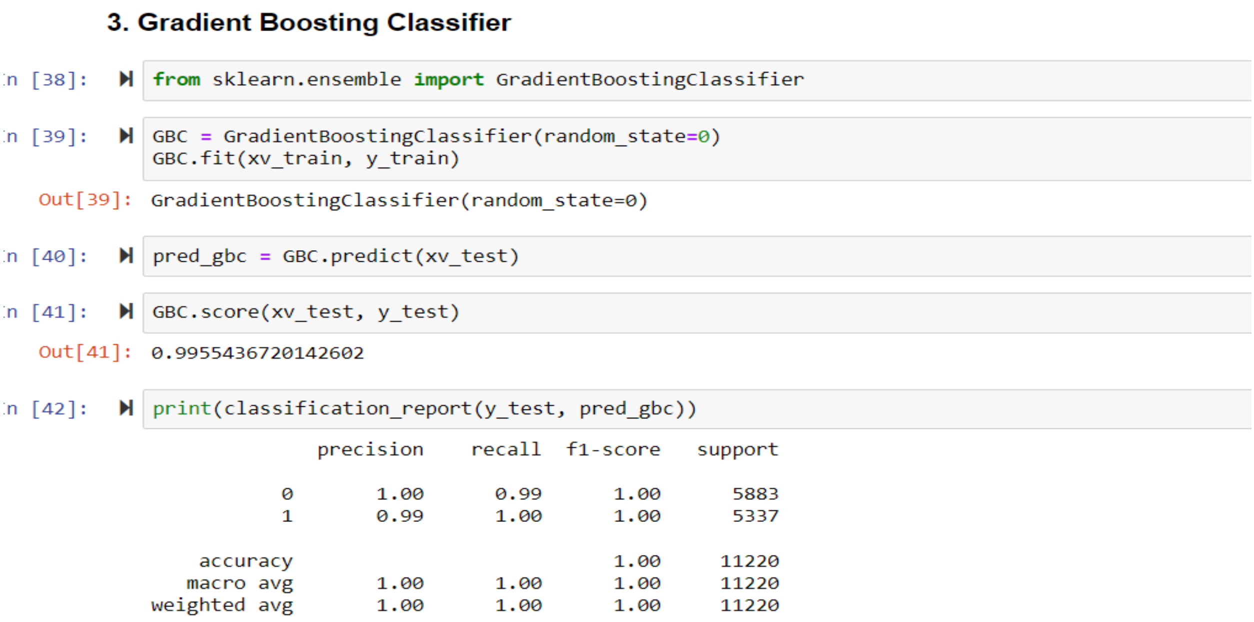
Random Forest: are built on the concept of building many decision tree algorithms, after which the decision trees get a separate result. The results, which are predicted by large number of decision tree, are taken up by the random forest. To ensure a variation of the decision trees, the random forest randomly selects a subcategory of properties from each group [16][17] The applicability of Random forest is best when used on uncorrelated decision trees. If applied on similar trees, the overall result will be more or less similar to a single decision tree. Uncorrelated decision trees can be obtained by bootstrapping and feature randomness.



4. Random Forest Classifier

```
In [43]:  from sklearn.ensemble import RandomForestClassifier

In [44]:  RFC = RandomForestClassifier(random_state=0)
          RFC.fit(xv_train, y_train)

Out[44]:  RandomForestClassifier(random_state=0)

In [45]:  pred_rfc = RFC.predict(xv_test)

In [46]:  RFC.score(xv_test, y_test)

Out[46]:  0.9915329768270945

In [47]:  print(classification_report(y_test, pred_rfc))

                        precision    recall  f1-score   support

                   0         0.99      0.99      0.99      5883
                   1         0.99      0.99      0.99      5337

            accuracy                             0.99     11220
           macro avg         0.99      0.99      0.99     11220
        weighted avg         0.99      0.99      0.99     11220
```

Gradient Boosting: are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many Kaggle data science competitions.



3. Gradient Boosting Classifier

```
In [38]:  from sklearn.ensemble import GradientBoostingClassifier

In [39]:  GBC = GradientBoostingClassifier(random_state=0)
          GBC.fit(xv_train, y_train)

Out[39]:  GradientBoostingClassifier(random_state=0)

In [40]:  pred_gbc = GBC.predict(xv_test)

In [41]:  GBC.score(xv_test, y_test)

Out[41]:  0.9955436720142602

In [42]:  print(classification_report(y_test, pred_gbc))

                        precision    recall  f1-score   support

                     0       1.00      0.99      1.00      5883
                     1       0.99      1.00      1.00      5337

              accuracy                           1.00     11220
             macro avg       1.00      1.00      1.00     11220
          weighted avg       1.00      1.00      1.00     11220
```

CHAPTER 6:
Future Scope and Conclusion

6.1) Future Scope

More complex and efficient methods could be surely applied to this dataset, for example using the entire text or extracting other features.

6.2) CONCLUSION

In the 21st century, the majority of the tasks are done online. Newspapers that were earlier preferred as hardcopies are now being substituted by applications like Facebook, Twitter, and news articles to be read online. Whatsapp‘s forwards are also a major source. The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, we have developed our Fake news Detection system that takes input from the user and classify it to be true or fake. To implement this, various NLP and Machine Learning Techniques have to be used. The model is trained using an appropriate dataset and performance evaluation is also done using various performance measures. The best model, i.e. the model with highest accuracy is used to classify the news headlines or articles. We are in the fantastic era of Deep Learning. One of the great thing about it is that while it is extremely difficult to train a state of art neural network, it is way easier and faster to use a pretrained neural network, fine tune it and obtain state of art results on your dataset. If you have something else you would like to add hit me at. We intent to build our own dataset which will be kept up to date according to the latest news.

All the live news and latest data will be kept in a database using Web Crawler and online database.

REFRENCES

1.https://projectgurukul.org/fake-news-detection-project-python-machine-learning/

2.https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/

3. https://en.wikipedia.org/wiki/Fake_news

4.https://downloads.hindawi.com/journals/complexity/2020/8885861.pdf

5. https://www.sciencedirect.com/topics/computer-science/logistic-regression

6.https://towardsdatascience.com/pre-processing-in-natural-language-machine-learning-898a84b8bd47#:~:text=their%20thy%20your-,Stemming,models%2C%20including%20Porter%20and%20Snowball

7. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, ―Fake News Detection using Machine Learning and Natural Language Processing,‖ International Journal of Recent Technology and International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NTASU - 2020 Conference Proceedings Volume 9, Issue 3 Special Issue - 2021 517 Engineering (IJRTE) I