

FAKE -NEWS DETECTION PROJECT

Submitted by:

HIMANSHU SHARMA

**ACKNOWLEDGMENT**

One of the pleasant aspects of preparing a project report is the opportunity to thank those who have contributed to make this project possible.

We are extremely thankful to Mr. Tushar Saraswat, whose active interest in the project & insight helped us to formulate, redefine implement our approach to the project.

We are also thankful to our institute & Other seen unseen hands which have given us direct & indirect help in completion of this project.

* **Himanshu Sharma**

**INTRODUCTION**

* **Business Problem Framing**

This project is related to 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 given dataset, 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).

Problem in this dataset, build a model to predict whether a news is fake or not fake.

* **Conceptual Background of the Domain Problem**

Conceptual Background in this dataset, i.e.We will use Statistics and machine learning algorithms as well as natural language technic to process the dataset and built the model. Since the target value is classified, we will use the classification model algorithms in it. To visualize the data, we will use matplotlib and seaborn library, which based on python language.

* **Review of Literature**

Review of Literature is basically related to comprehensive summary of dataset as well as descriptions of input variables and output variable.

****

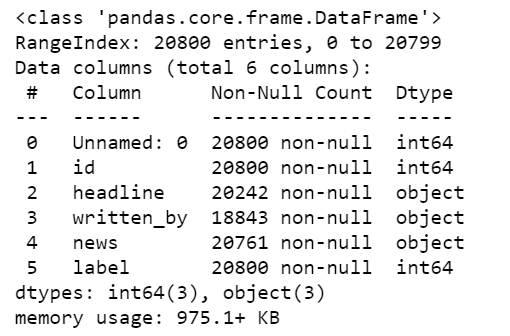
* **Motivation for the Problem Undertaken**

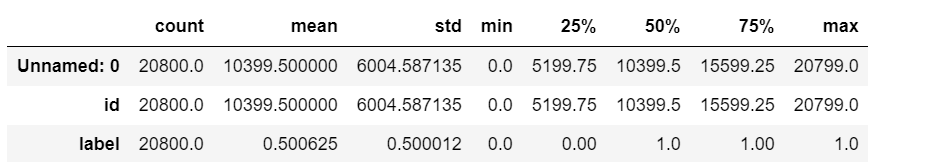
Motivation for the problem undertaken Because this project will help us to know fake news and real news. Which will be helpful in detecting real and fake news to any organization and society.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modelling of the Problem**

In this section we understand and describe the mathematical, statistical and analytics modelling done during this project along with the proper justification. First of all for better understanding about dataset and given attributes information we use **Dataframe.info()** command, which tell me, The total number of attributes, what is name of attributes, datatype of attributes and how many Non-null values are present in dataset.



After understanding the dataset values, we take the statistical descriptions of dataset using **Dataframe.describe()** command in python, which tells the following statistical descriptions:

After know about statistical description we move forward in the way of finding co-relation of variables, then move to data-cleaning, then move to visualized the data set and their relationship. After this we move to data pre-processing and modelling as well as testing the accuracy of model.

* **Data Sources and their formats**

Machine learning algorithms as well as Natural Language processing are almost always optimized for raw, detailed source data. Thus, the data environment must provision large quantities of raw data for discovery-oriented analytics practices such as data exploration, data mining, statistics, and machine learning.

Tabular data for machine learning is typically found is .csv files. Csv files are text-based files containing comma separated values (csv). Csv files are popular for ML as they are easy to view/debug and easy to read/write from programs (no compression/indexing).

* **Data Pre-processing Done**
  1. Read dataset and make it in proper format.
  2. Encode labels
  3. Convert all cases to lower
  4. Remove punctuations
  5. Remove Stopwords
  6. Check stats of messages
  7. Convert all texts into vectors
  8. Import classifier
  9. Train and test
  10. Check the accuracy/confusion matrix.
* **Hardware and Software Requirements and Tools Used**

In this project dataset is too large for processing or modelling, that’s why we use good hardware configuration like as above 4GB RAM, above or equal core i3 processor and also need good storage HDD. In way of software, we use any operating system which support python language for coding.

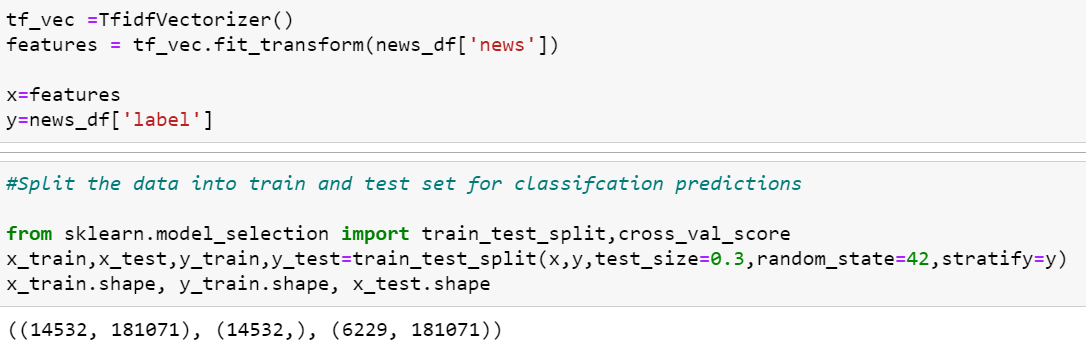
**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

In this dataset we build a model to predict whether a news is fake i.e., “1” or not fake i.e.,”0”.

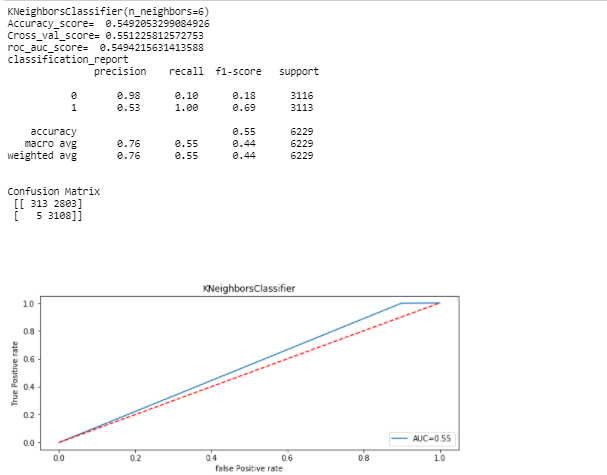
**Testing of Identified Approaches (Algorithms)**

* Convert all texts into vectors
* Import classifier
* Train and test

****

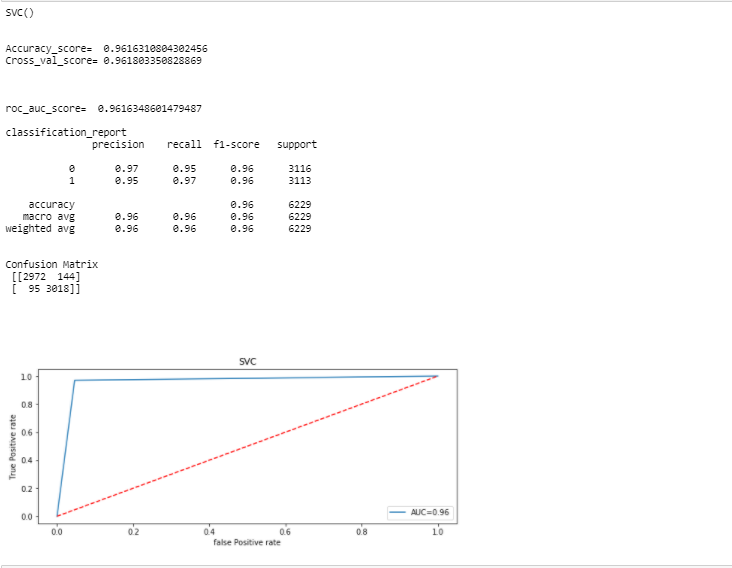
* **Run and Evaluate selected models**

1. **kneighborsclassifier algorithm**



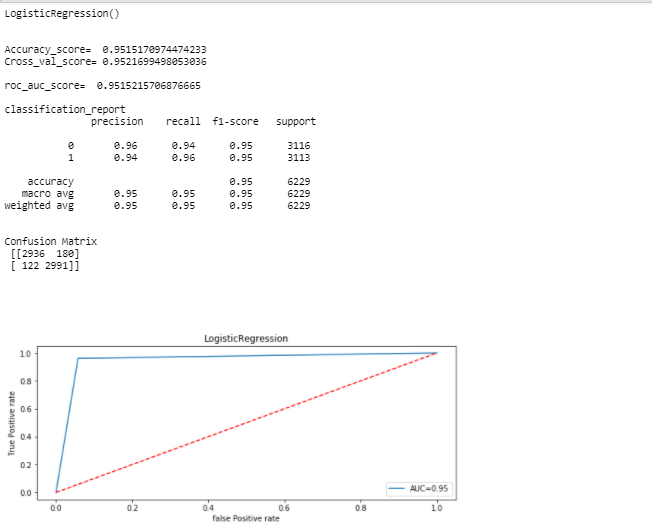
1. **svc algorithm**

****

****

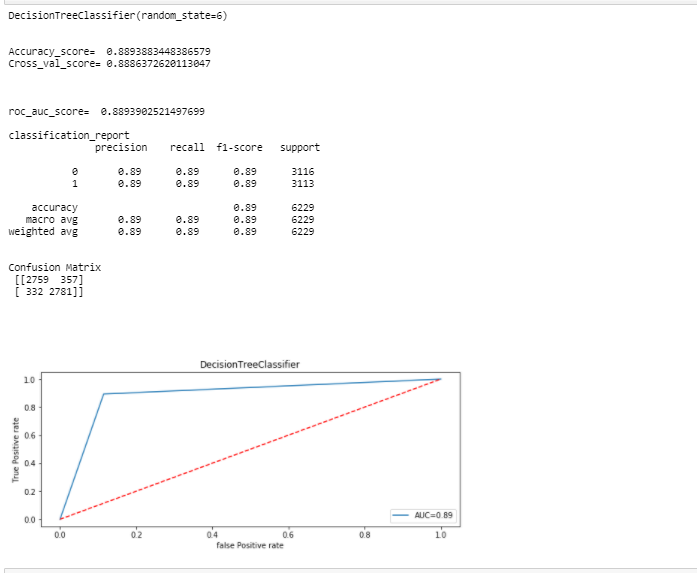
1. **Logistic regression**

****

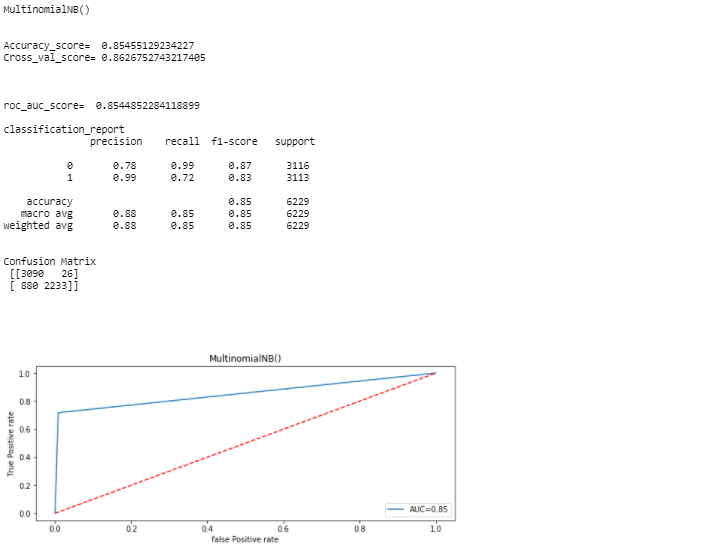
****

1. **Decision tree classifier**

****

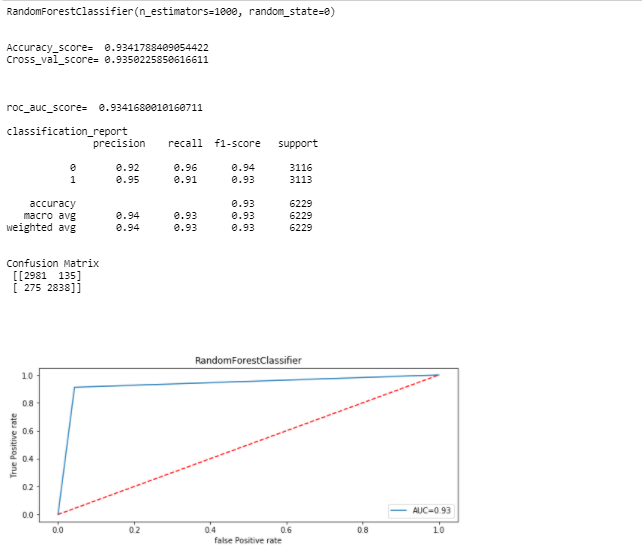
****

1. **MultinomialNB**

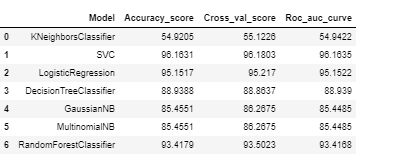
****

1. **RandomForestClassifier**

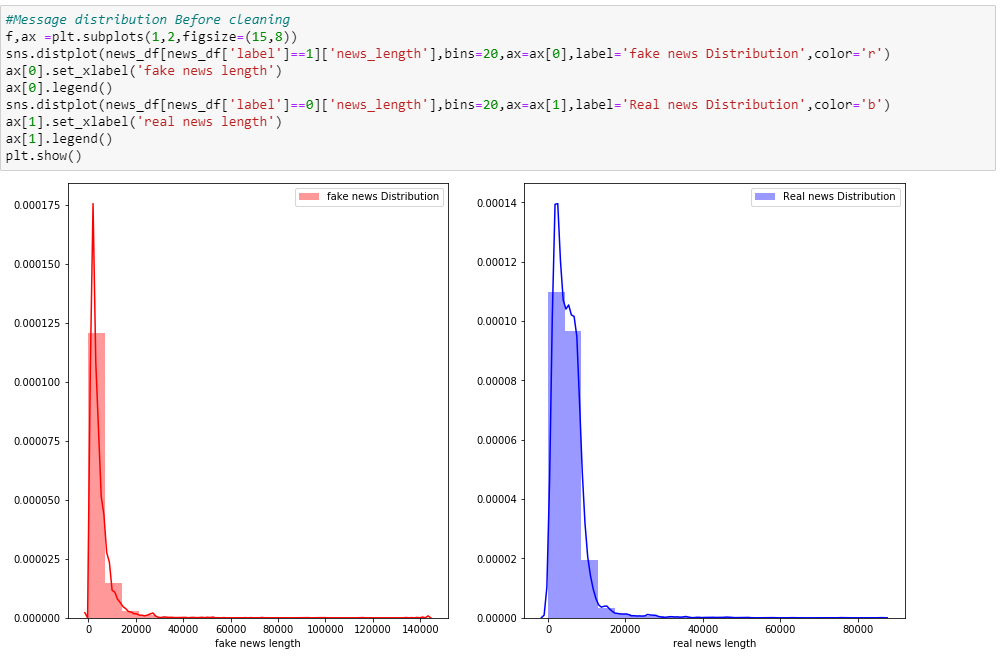
****

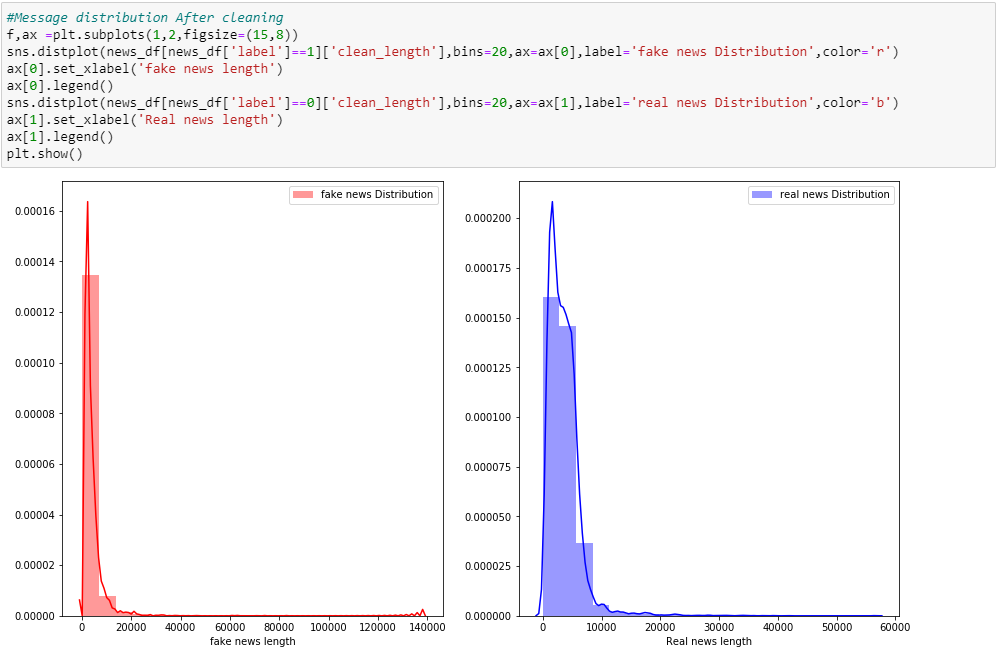
****

* **Final Result of models**

****

* **Visualizations**

**1.**

**2.**

**3.**

**4.**

****

**CONCLUSION**

After analysing data, visualization and modelling, we come to the conclusion that using the Support Vector Classification algorithm is suitable for modelling of label's prediction.