**DATA SCIENCE**

(Fake news classifier)

Summer Internship Report Submitted in partial fulfillment

of the requirement for undergraduate degree of

**Bachelor of Technology**

**in**

**Computer Science Engineering**

**by**

**Akshith Reddy Kalvala**

**221710313003**

****

**Department of Computer Science and Engineering**

**GITAM School of Technology**

**GITAM(Deemed to be University)**

**Hyderabad-502329**

**June 2019**

**DECLARATION**

I submit this industrial training work entitled “FAKE NEWS CLASSIFIER” to GITAM(Deemed To Be University), Hyderabad in partial fulfillment of the requirements for the award of the degree of “Bachelor of Technology ” in “ Computer Science Engineering”. I declare that it was carried out independently by me under the guidance of \_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_, GITAM(Deemed to be University),Hyderabad,India. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Place: **Akshith Reddy Kalvala**

**221710313003**

Hyderabad

**Certificate**

**ACKNOWLEDGEMENT**

Apart from my effort, the success of this internship largely depends on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have helped me in the successful competition of this internship.

I would like to thank Dr. N. Siva Prasad, Pro Vice Chancellor, GITAM Hyderabad and Prof. N. Seetha Ramaiah, Principal, GITAM Hyderabad.

I would like to thank respected Prof. S. Phani Kumar, Head of the Department of Computer Science Engineering for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present the internship report. It helped me alot to realize what we study for.

I would like to thank the respected faculties \_\_\_\_\_\_\_\_ who helped me to make this internship a successful accomplishment. I would also like to thank my friends who helped me to make my work more organized and well-stacked till the end.

**Akshith Reddy Kalvala**

**221710313003**

**ABSTRACT**

Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called `fake news' is initially distributed over social media conduits like Facebook and Twitter and later finds its way into mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented.

Machine learning algorithms are used to predict the values from the dataset by splitting the dataset into train and test and building Machine learning algorithms models of higher accuracy to predict the values is the primary task to be performed on news dataset.

Table of Contents:

1.INFORMATION ABOUT DATA SCIENCE

1.1 What is data science

1.2 Need of data science

1.3 Uses of data science

2. INFORMATION ABOUT MACHINE LEARNING :

2.1 Importance of machine learning

2.2Uses of machine learning

2.3 Types of learning algorithms

2.4 Relation between Data mining, Machine learning and Deep learning

3.INFORMATION ABOUT PYTHON:

3.1 Introduction

3.2 History of python

3.3 Features of python

3.4 How to setup python

3.5 Python variable types

3.6 Python functions

3.7 Python using OOPS concepts

4. PROJECT NAME:

4.1 Packages used

4.2 Problem statement

4.3 Data set description

4.4 Objective of the case study

5. DATA PREPROCESSING / FEATURE ENGINEERING & EDA:

5.1 Importing the libraries

5.2Readibg the data set

5.3 Handling missing values

5.4 Categorical data

6.MODEL BUILDING AND EVALUATION

6.1 Gaussian Naive Bayes

6.2 Logistic Regression

6.3KNearest Neighbour

**CHAPTER 1**

**DATA SCIENCE**

1.INFORMATION ABOUT DATA SCIENCE :

1.1 What is Data Science :

Data science is an interdisciplinary field that uses scientific methods,processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, deep learning and big data.

1.2 Need of Data Science :

Industries need data to help them make careful decisions. Data Science churns raw data into meaningful insights. Therefore, industries need data science. A Data Scientist is a wizard who knows how to create magic using data. The model will know how to dig out meaningful information with whatever data he comes across. The company requires strong data-driven decisions. The Data Scientist is an expert in various underlying fields of Statistics and Computer Science.

Companies are applying big data and data science to everyday activities to bring value to consumers. Banking institutions are capitalizing on big data to enhance their fraud detection successes. Asset management firms are using big data to predict the likelihood of a security’s price moving up or down at a stated time.

Companies such as Netflix mine big data to determine what products to deliver to its users. Netflix also uses algorithms to create personalized recommendations for users based on their viewing history. Data science is evolving at a rapid rate, and its applications will continue to change lives into the future.

1.3 Uses of Data Science:

Fraud and Risk Detection:

The earliest applications of data science were in Finance. Companies were fed up with bad debts and losses every year. However, they had a lot of data which used to get collected during the initial paperwork while sanctioning loans. They decided to bring in a data scientist in order to rescue them out of losses.

Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on the customer's purchasing power.

Healthcare:

1.Medical Image Analysis:

Procedures such as detecting tumours, artery stenosis, organ delineation employ various different methods and frameworks like Map Reduce to find optimal parameters for tasks like lung texture classification. It applies machine learning methods, support vector machines (SVM), content-based medical image indexing, and wavelet analysis for solid texture classification.

2.Virtual assistance for patients and customer support:

The AI-powered mobile apps can provide basic healthcare support, usually as chatbots. You simply describe your symptoms, or ask questions, and then receive key information about your medical condition derived from a wide network linking symptoms to causes. Apps can remind you to take your medicine on time, and if necessary, assign an appointment with a doctor.

Targeted Advertising:

Digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user’s past behavior.

This is the reason why you might see ads for Data Science Training Programs while I see an ad for apparels in the same place at the same time.

Other uses:

● Speech Recognition

● Airline Route Planning

● Gaming

● Augmented Reality

2. INFORMATION ABOUT MACHINE LEARNING :

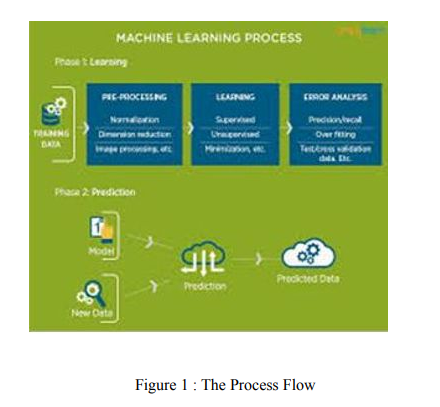
Machine Learning(ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of Artificial Intelligence(AI).

2.1 IMPORTANCE OF MACHINE LEARNING:

Consider some of the instances where machine learning is applied: the self-driving Google car, cyber fraud detection, online recommendation engines—like friend suggestions on Facebook, Netflix showcasing the movies and shows you might like, and “more items to consider” and “get yourself a little something” on Amazon—are all examples of applied machine learning. All these examples echo the vital role machine learning has begun to take in today’s data-rich world.

Machines can aid in filtering useful pieces of information that help in major advancements, and we are already seeing how this technology is being implemented in a wide variety of industries.

With the constant evolution of the field, there has been a subsequent rise in the uses, demands, and importance of machine learning. Big data has become quite a buzzword in the last few years; that’s in part due to increased sophistication of machine learning, which helps analyze those big chunks of big data. Machine learning has also changed the way data extraction, and interpretation is done by involving automatic sets of generic methods that have replaced traditional statistical techniques.



The process flow depicted here represents how machine learning works

2.2 USES OF MACHINE LEARNING:

Earlier in this article, we mentioned some applications of machine learning. To understand the concept of machine learning better, let’s consider some more examples: web search results, real-time ads on web pages and mobile devices, email spam filtering, network intrusion detection, and pattern and image recognition. All these are by-products of applying machine learning to analyze huge volumes of data.

Traditionally, data analysis was always being characterized by trial and error, an approach that becomes impossible when data sets are large and heterogeneous. Machine learning comes as the solution to all this chaos by proposing clever alternatives to analyzing huge volumes of data.

By developing fast and efficient algorithms and data-driven models for real-time processing of data, machine learning can produce accurate results and analysis.

2.3 TYPES OF LEARNING ALGORITHMS:

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

2.3.1 Supervised Learning :

When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of supervised learning.

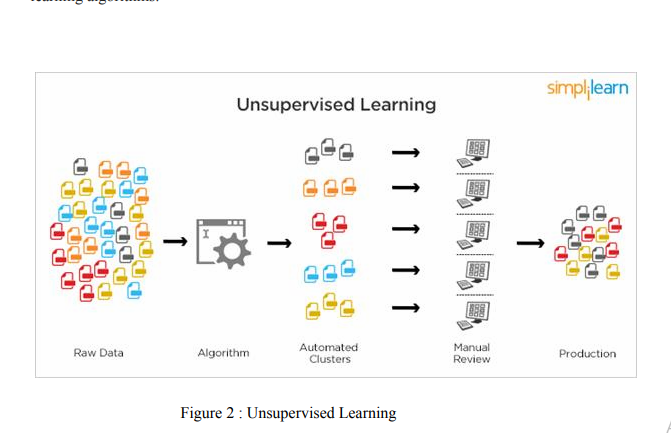
Supervised machine learning algorithms uncover insights, patterns, and relationships from a labelled training dataset – that is, a dataset that already contains a known value for the target variable for each record. Because you provide the machine learning algorithm with the correct answers for a problem during training, it is able to “learn” how the rest of the features relate to the target, enabling you to uncover insights and make predictions about future outcomes based on historical data.

Examples of Supervised Machine Learning Techniques are Regression, in which the algorithm returns a numerical target for each example, such as how much revenue will be generated from a new marketing campaign.

Classification, in which the algorithm attempts to label each example by choosing between two or more different classes. Choosing between two classes is called binary classification, such as determining whether or not someone will default on a loan. Choosing between more than two classes is referred to as multiclass classification.

2.3.2 Unsupervised Learning:

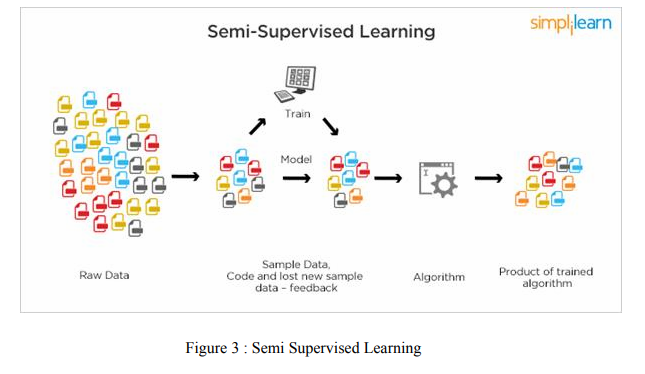
When an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of uncorrelated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.



Popular techniques where unsupervised learning is used also include self-organizing maps, nearest neighbor mapping, singular value decomposition, and k-means clustering. Basically, online recommendations, identification of data outliers, and segment text topics are all examples of unsupervised learning.

2.3.3 Semi Supervised Learning:

As the name suggests, semi-supervised learning is a bit of both supervised and unsupervised learning and uses both labeled and unlabeled data for training. In a typical scenario, the algorithm would use a small amount of labeled data with a large amount of unlabeled data.



2.4 RELATION BETWEEN DATA MINING,MACHINE LEARNING AND DEEP LEARNING:

Machine learning and data mining use the same algorithms and techniques as data mining, except the kinds of predictions vary. While data mining discovers previously unknown patterns and knowledge, machine learning reproduces known patterns and knowledge—and further automatically applies that information to data, decision-making, and actions.

Deep learning, on the other hand, uses advanced computing power and special

types of neural networks and applies them to large amounts of data to learn, understand, and identify complicated patterns. Automatic language translation and medical diagnoses are examples of deep learning.

**CHAPTER 3**

**PYTHON**

Basic programming language used for machine learning is : PYTHON

3.1 INTRODUCTION TO PYTHON:

● Python is a high-level, interpreted, interactive and object-oriented scripting language.

● Python is a general purpose programming language that is often applied in scripting roles.

● Python is Interpreted: Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is like PERL and PHP.

● Python is Interactive: You can sit at a Python prompt and interact with the interpreter directly to write your programs.

● Python is Object-Oriented: Python supports the Object-Oriented style or technique of programming that encapsulates code within objects.

3.2 HISTORY OF PYTHON:

● Python was developed by GUIDO VAN ROSSUM in early 1990’s

● Its latest version is 3.7 , it is generally called as python3

3.3 FEATURES OF PYTHON:

● Easy-to-learn: Python has few keywords, simple structure, and a clearly defined syntax, This allows the student to pick up the language quickly.

● Easy-to-read: Python code is more clearly defined and visible to the eyes.

● Easy-to-maintain: Python's source code is fairly easy-to-maintaining.

● A broad standard library: Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

● Portable: Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

● Extendable: You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.

● Databases: Python provides interfaces to all major commercial databases.

● GUI Programming: Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.

3.4 HOW TO SETUP PYTHON:

● Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

● The most up-to-date and current source code, binaries, documentation, news, etc., is available on the official website of Python.

3.4.1 Installation(using python IDLE):

● Installing python is generally easy, and nowadays many Linux and Mac OS distributions include a recent python.

● Download python from www.python.org

● When the download is completed, double click the file and follow the instructions to install it.

● When python is installed, a program called IDLE is also installed along with it. It provides a graphical user interface to work with python.



3.4.2 Installation(using Anaconda):

● Python programs are also executed using Anaconda.

● Anaconda is a free open source distribution of python for large scale data processing, predictive analytics and scientific computing.

● Conda is a package manager that quickly installs and manages packages.

● In WINDOWS:

● In windows

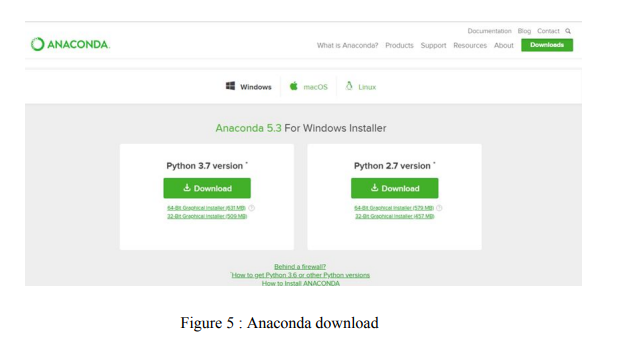
● Step 1: Open Anaconda.com/downloads in a web browser.

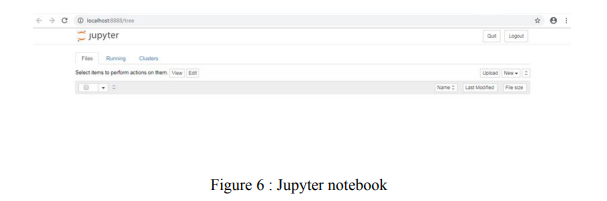
● Step 2: Download python 3.4 version for (32-bits graphic installer/64 -bit graphic installer)

● Step 3: select installation type( all users)

● Step 4: Select path(i.e. add anaconda to path & register anaconda as default python 3.4) next click install and next click finish

● Step 5: Open jupyter notebook ( it opens in default browser)





3.5 PYTHON VARIABLE TYPES:

● Variables are nothing but reserved memory locations to store values. This means that when you create a variable you reserve some space in memory.

● Variables are nothing but reserved memory locations to store values.

● Based on the data type of a variable, the interpreter allocates memory and decides what can be stored in the reserved memory.

● Python variables do not need explicit declaration to reserve memory space. The declaration happens automatically when you assign a value to a variable.

● Python has various standard data types that are used to define the operations possible on them and the storage method for each of them.

● Python has five standard data types –

o Numbers

o Strings

o Lists

o Tuples

o Dictionary

3.5.1 Python Numbers:

● Number data types store numeric values. Number objects are created when you assign a value to them.

● Python supports four different numerical types − int (signed integers) long (long integers, they can also be represented in octal and hexadecimal) float (floating point real values) complex (complex numbers).

3.5.2 Python Strings:

● Strings in Python are identified as a contiguous set of characters represented in the quotation marks.

● Python allows for either pairs of single or double quotes.

● Subsets of strings can be taken using the slice operator ([ ] and [:] ) with indexes starting at 0 in the beginning of the string and working their way from -1 at the end.

● The plus (+) sign is the string concatenation operator and the asterisk (\*) is the repetition operator.

3.5.3 Python Lists:

● Lists are the most versatile of Python's compound data types.

● A list contains items separated by commas and enclosed within square brackets

([]).

● To some extent, lists are similar to arrays in C. One difference between them is that all the items belonging to a list can be of different data types.

● The values stored in a list can be accessed using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the list and working their way to end -1.

● The plus (+) sign is the list concatenation operator, and the asterisk (\*) is the repetition operator.

3.5.4 Python Tuples:

● A tuple is another sequence data type that is similar to the list.

● A tuple consists of a number of values separated by commas. Unlike lists, however, tuples are enclosed within parentheses.

● The main differences between lists and tuples are: Lists are enclosed in brackets ( [ ] ) and their elements and size can be changed, while tuples are enclosed in parentheses ( ( ) ) and cannot be updated.

● Tuples can be thought of as read-only lists.

● For example − Tuples are fixed size in nature whereas lists are dynamic. In other words, a tuple is immutable whereas a list is mutable. You can't add elements to a tuple. Tuples have no append or extend method. You can't remove elements from a tuple. Tuples have no remove or pop method.

3.5.5 Python Dictionary:

● Python's dictionaries are kind of hash table type. They work like associative arrays

or hashes found in Perl and consist of key-value pairs. A dictionary key can be almost any Python type, but are usually numbers or strings. Values, on the other hand, can be any arbitrary Python object.

● Dictionaries are enclosed by curly braces ({ }) and values can be assigned and accessed using square braces ([]).

● You can use numbers to "index" into a list, meaning you can use numbers to find out what's in lists. You should know this about lists by now, but make sure you understand that you can only use numbers to get items out of a list.

● What a dict does is let you use anything, not just numbers. Yes, a dict associates one thing to another, no matter what it is.

3.6 PYTHON FUNCTION:

3.6.1 Defining a Function:

You can define functions to provide the required functionality. Here are simple rules to define a function in Python. Function blocks begin with the keyword def followed by the function name and parentheses (i.e.()).

Any input parameters or arguments should be placed within these parentheses. You can also define parameters inside these parentheses The code block within every function starts with a colon (:) and is indented.

The statement returns [expression] exits a function, optionally passing back an expression to the caller. A return statement with no arguments is the same as return None.

3.6.2 Calling a Function:

Defining a function only gives it a name, specifies the parameters that are to be included in the function and structures the blocks of code. Once the basic structure of a function is finalized, you can execute it by calling it from another function or directly from the Python prompt.

3.7 PYTHON USING OOPs CONCEPTS:

3.7.1 Class:

● Class: A user-defined prototype for an object that defines a set of attributes that characterize any object of the class. The attributes are data members (class variables and instance variables) and methods, accessed via dot notation.

● Class variable: A variable that is shared by all instances of a class. Class variables are defined within a class but outside any of the class's methods. Class variables are not used as frequently as instance variables are.

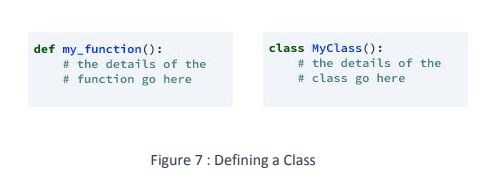
● Data member: A class variable or instance variable that holds data associated with a class and its objects.

● Instance variable: A variable that is defined inside a method and belongs only to the current instance of a class.

● Defining a Class:

o We define a class in a very similar way how we define a function.

o Just like a function ,we use parentheses and a colon after the class name(i.e. ():) when we define a class. Similarly, the body of our class is indented like a functions body is.



3.7.2 \_\_init\_\_ method in Class:

● The init method — also called a constructor — is a special method that runs when an instance is created so we can perform any tasks to set up the instance.

● The init method has a special name that starts and ends with two underscores:\_\_init\_\_().

**CHAPTER 4**

**CASE STUDY**

4.1 PROBLEM STATEMENT:

To predict the amount of purchases in Retail shop using Machine Learning algorithm called MULTIPLE LINEAR REGRESSION.

4.2 DATA SET:

The given data set consists of the following parameters:

A - User\_ID

B - Product\_ID

C - Gender

D - Age

E - Occupation

F - City\_Category

G - Stay\_In\_Current\_City\_Years

H - Marital\_Status I - Product\_Category\_1

J - Product\_Category\_2

K - Product\_Category\_3

4.3 OBJECTIVE OF THE CASE STUDY:

To get a better understanding and chalking out a plan of action for solution of the client, we have adapted the view point of looking at product categories and for further deep understanding of the problem, we have also considered gender age of the customer and reasoned out the various factors of choice of the products and they purchase , and our primary objective of this case study was to look up the factors which were dampening the sale of products and correlate them to product categories and draft out an outcome report to client regarding the various accepts of a product purchases.

**CHAPTER 5**

**MODEL BUILDING**

5.1 PREPROCESSING OF THE DATA:

Preprocessing of the data actually involves the following steps:

5.1.1 GETTING THE DATASET:

We can get the data set from the database or

we can get the data from client.

5.1.2 IMPORTING THE LIBRARIES:

We have to import the libraries as per the requirement of the algorithm.



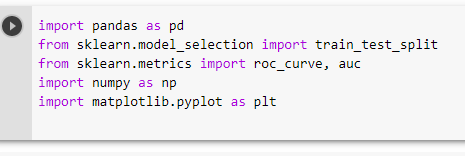


Figure 8 : Importing Libraries

5.1.3 IMPORTING THE DATA-SET:

Pandas in python provide an interesting method read\_csv(). The read\_csv function reads the entire dataset from a comma separated values file and we can assign it to a DataFrame to which all the operations can be performed. It helps us to access each and every row as well as columns and each and every value can be access using the dataframe. Any missing value or NaN value have to be cleaned.



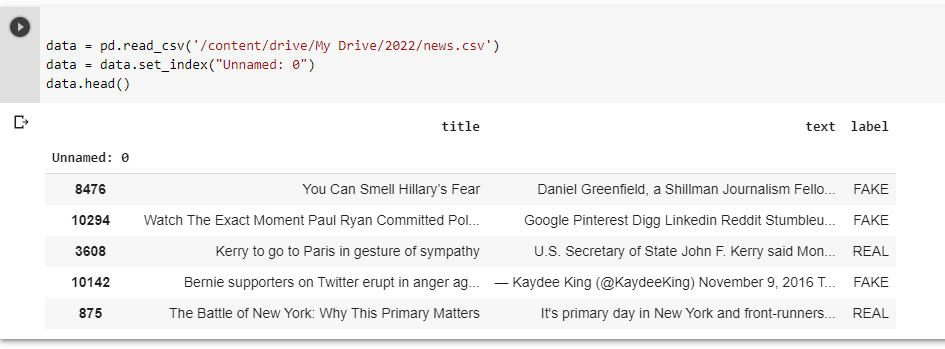


Figure 9 : Reading the dataset

5.1.4 PREPROCESSING OF THE DATA SET:

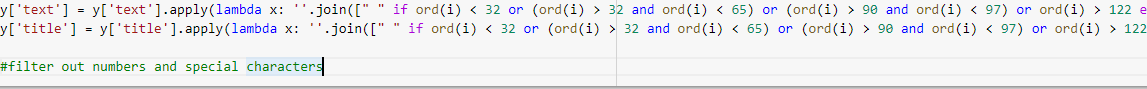
• Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

• Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

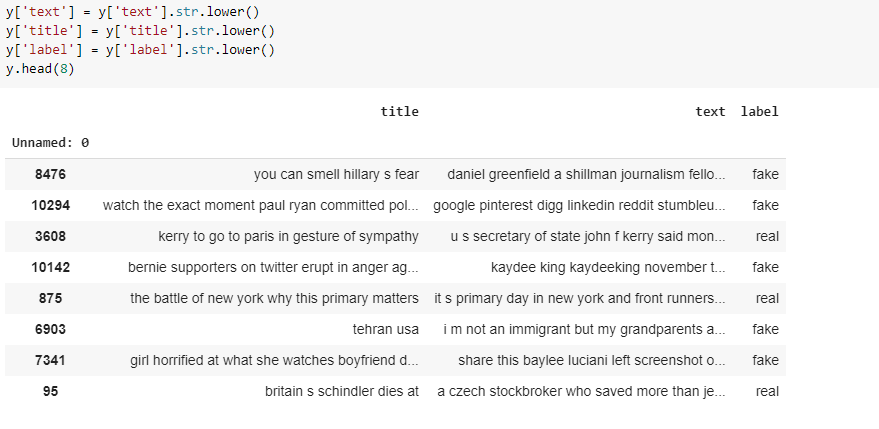
Need of Data Preprocessing

• For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format.

• Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.



To convert data into Lowercase:

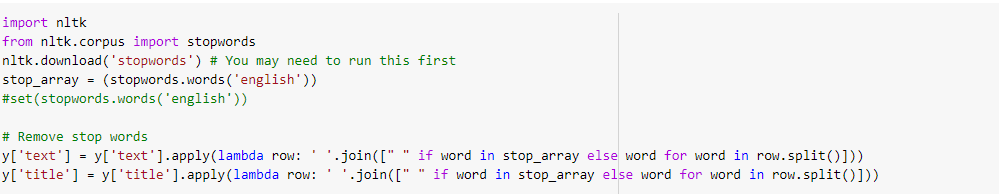


# 5.1.5 Removing stop words with NLTK :

The process of converting data to something a computer can understand is referred to as pre-processing**.** One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

Stop Words**:** A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) .



5.1.6 Splitting data into Test and Train Data Set:

Splitting the data : after the preprocessing is done then the data is split into train

and test sets.

● In Machine Learning in order to access the performance of the classifier. You train the

classifier using 'training set' and then test the performance of your classifier on unseen

'test set'. An important point to note is that during training the classifier only uses the

training set . The test set must not be used during training the classifier. The test set will

only be available during testing the classifier.

● training set - a subset to train a model.(Model learns patterns between Input and Output)

● test set - a subset to test the trained model.(To test whether the model has correctly learnt)

● The amount or percentage of Splitting can be taken as specified (i.e. train data = 75% ,

test data =25% or train data = 80% , test data= 20%) .

● First we need to identify the input and output variables and we need to separate the input set and output set.

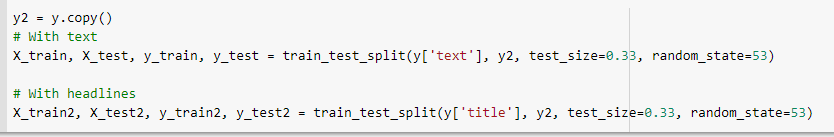
● In scikit learn library we have a package called model\_selection in which train\_test\_splitmethod is available .we need to import this method.

● This method splits the input and output data to train and test based on the percentage

specified by the user and assigns them to four different variables(we need to mention the

variables) .

As we work with datasets, a machine learning algorithm works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python ML.



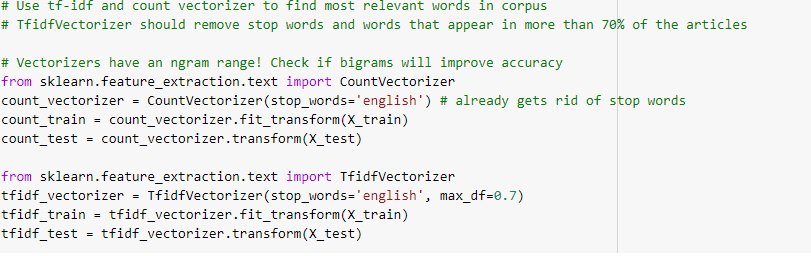
# 5.1.7 Lemmatization**:**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

Applications of lemmatization:

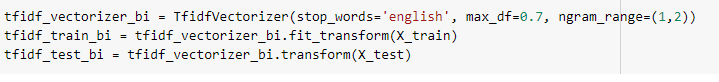
•Used in comprehensive retrieval systems like search engines.

•Used in compact indexing



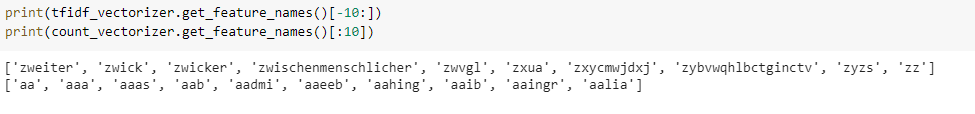
# 5.1.8 Vectorization :

Vectorization is used to speed up the Python code without using loop. Using such a function can help in minimizing the running time of code efficiently. Various operations are being performed over vector such as *dot product of vectors* which is also known as *scalar product* as it produces single output, outer products which results in square matrix of dimension equal to length X length of the vectors, *Element wise multiplication* which products the element of same indexes and dimension of the matrix remain unchanged.



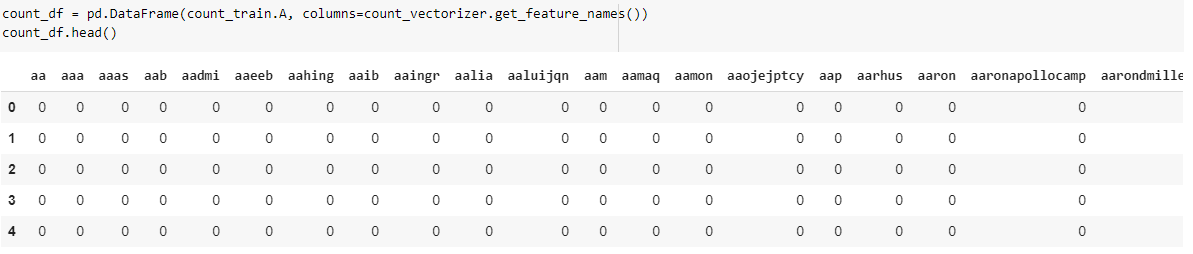
### 5.1.9**Stemming:**

Stemming is the process of getting the root form of a word. Stem or root is the part to which inflectional affixes (-ed, -ize, -de, -s, etc.) are added. The stem of a word is created by removing the prefix or suffix of a word. So, stemming a word may not result in actual words.



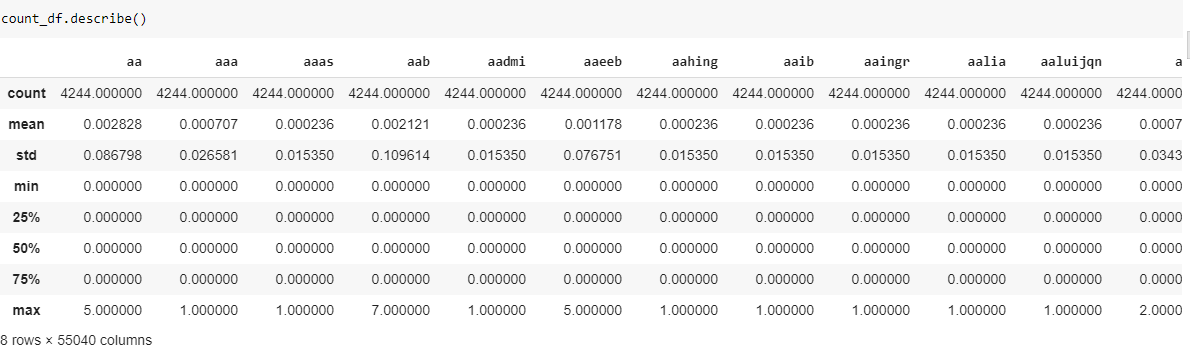
5.1.10Data Frame :

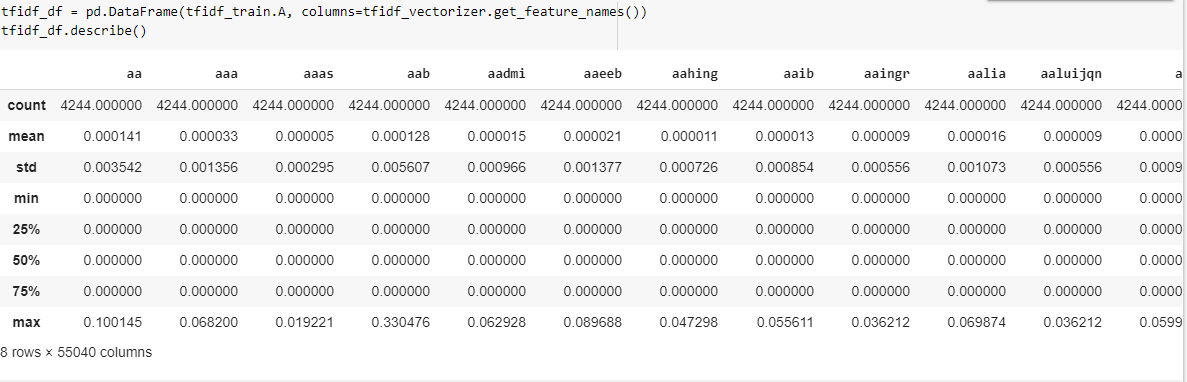
Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns.



# Python Dataframe.describe() method:

Pandas **describe()** is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. When this method is applied to a series of string, it returns a different output which is shown in the examples below.





6. MODEL BUILDING AND EVALUATION:

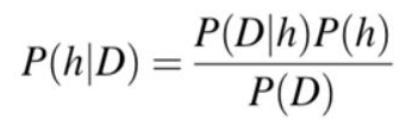
6.1 Naive Bayes

Naive Bayes is the most straightforward and fast classification algorithm, which is

suitable for a large chunk of data. Naive Bayes classifier is successfully used in various

applications such as spam filtering, text classification, sentiment analysis, and recommender systems. It uses Bayes theorem of probability for prediction of unknown class.

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

****

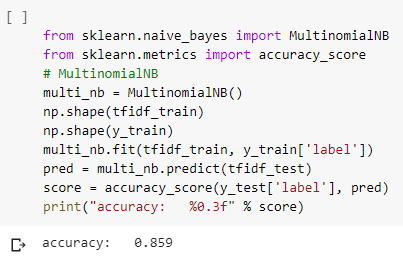
● P(h): the probability of hypothesis h being true (regardless of the data). This is known as

the prior probability of h.● P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

● P(h|D): the probability of hypothesis h given the data D. This is known as posterior

Probability.

● P(D|h): the probability of data d given that the hypothesis h was true. This .



6.1.1. ROC Curve

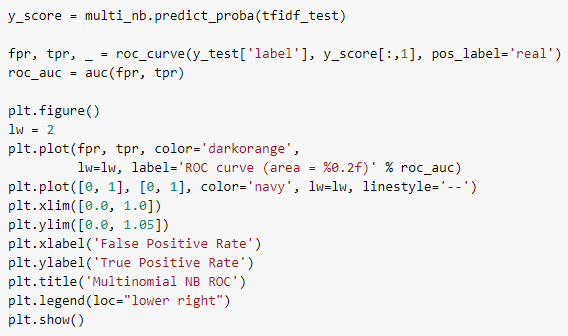
The ROC curve is a useful tool for a few reasons:

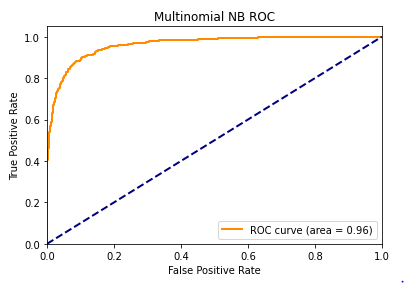
* The curves of different models can be compared directly in general or for different thresholds.
* The area under the curve (AUC) can be used as a summary of the model skill.

The shape of the curve contains a lot of information, including what we might care about most for a problem, the expected false positive rate, and the false negative rate.

To make this clear:

* Smaller values on the x-axis of the plot indicate lower false positives and higher true negatives.
* Larger values on the y-axis of the plot indicate higher true positives and lower false negatives.





6.1.2. Confusion Matrix in Machine Learning:

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

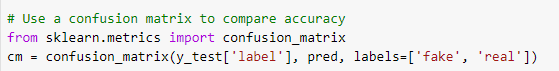
It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix.

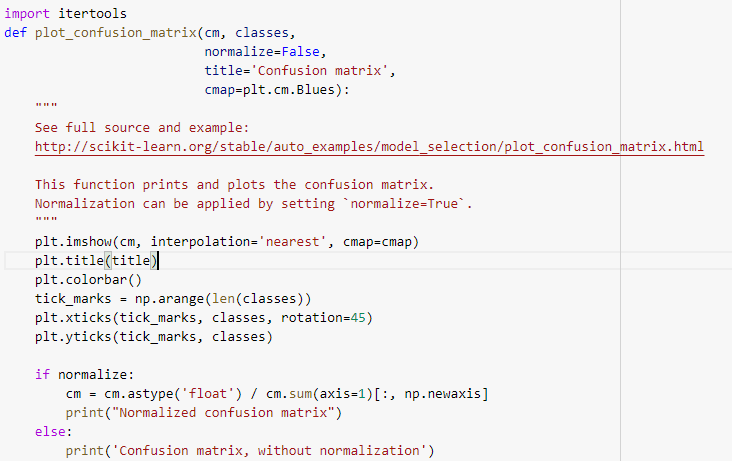
**This article aims at:**

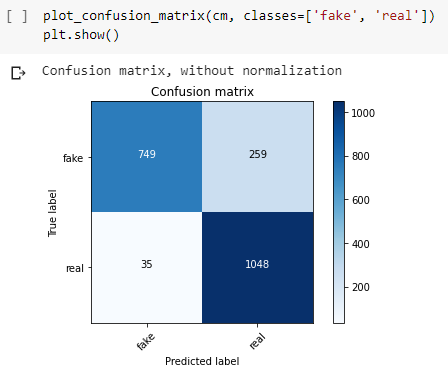
1. What the confusion matrix is and why you need to use it.
2. How to calculate a confusion matrix for a 2-class classification problem from scratch.
3. How to create a confusion matrix in Python.

**Confusion Matrix:**

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.





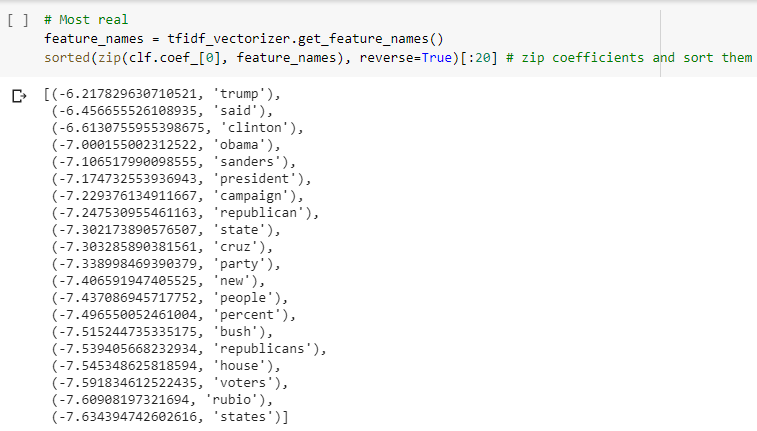


**6.2PASSIVE AGGRESSIVE CLASSIFIER:**

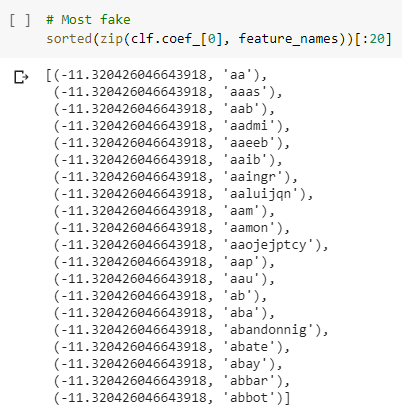
Passive-Aggressive Classifier belong to the family of online learning algorithms for binary classification. It is similar to support vector machine classifier and can be considered as the online version of SVM. It finds a hyperplane to separate the instance into two halves. The margin of an example is proportional to the distance of the hyperplane. The errors in predicting examples can be corrected by using margin of thehyperplane. The update of the classifier follows the constraints passive update and aggressive update. The advantage is that it follows an online learning pattern and update the separating hyperplane for the next example and ensure the performance of the algorithm[13]. Because of the theoretical loss bound, the performance of the classifier can be easily predicted.



Some of the real data from the given data set is as follows:



Some of the fake data from the given data set is as follows:



**6.3 Logistic regression :**

Logistic Regression is used when the dependent variable(target) is

Categorical.

Logistic Regression is the appropriate regression analysis to conduct when the dependent

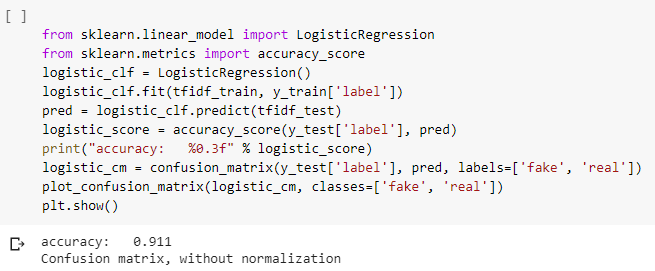
variable is dichotomous (binary). Like all regression analyses, the logistic regression is a

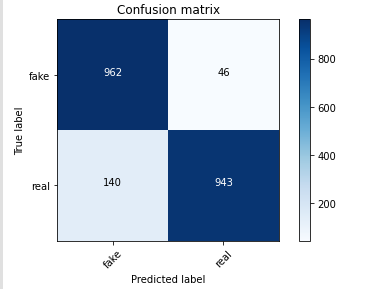
predictive analysis and it predicts the probability

Example: Yes or No, get a disease or not, pass or fail, defective or non-defective, etc.,

Also called a classification algorithm, because we are classifying the data. It predicts the

probability associated with each dependent variable category.

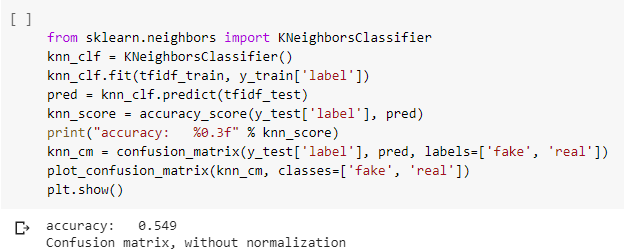


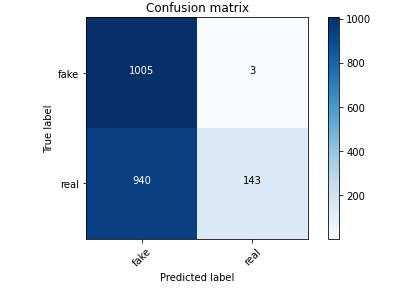


**6.4 K Nearest Neighbour**

K nearest neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It should also be noted that all three distance measures are only valid for continuous variables.

This works based on minimum distance from the query instance to the training samples to determine the k-nearest neighbors. After we gather these k-nearest neighbors, we take the simple majority of these k nearest neighbors to be the prediction of the query instance.





**CONCLUSION:**

In this project, I used Machine Learning to predict whether a news is fake or real. After importing the data, I analysed it using plots. Then, I generated dummy variables for categorical features and scaled other features. I then applied four Machine Learning Algorithms, K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier andRandom Forest Classifier.I varied parameters across each model to improve their scores.

1.K Neighbors Classifier: 54.9%

2.Naives Bayes:85%

3.PassiveAggressive Classifier: 93.8%

4LogisticRegression:91.1%

PassiveAggressive Classifier scored the best score of 93.8%

**REFERENCES:**

1.<https://en.wikipedia.org/wiki/Machine_learning>

2.[https://www.edureka.co/blog/what-is-data-science](https://www.edureka.co/blog/what-is-data-science/)

3..[https://www.edureka.co/blog/data-science-applications](https://www.edureka.co/blog/data-science-applications/)