# Iris flower classification and Fake News Detection using Machine Learning

A Project report submitted in partial fulfilment of the requirements for the degree of B.Tech in Electrical Engineering By

**Gourab Haldar** (11701617060)

**Srijan Routh** (11701617031)

**Manish Kumar Mandal** (11701617058)

**Suman Kumar** (11701617025)

Under the supervision of

**Mr. Subhasis Bandopadhyay**

Assistant Professor Dept. of Electrical Engineering

Department of Electrical Engineering

**RCC INSTITUTE OF INFORMATION TECHNOLOGY**

CANAL SOUTH ROAD, BELIAGHATA, KOLKATA – 700015, WEST BENGAL

# *CERTIFICATE*

To whom it may concern

This is to certify that the project work entitled is solar and wind hybrid power generation the bona fide work carried out by GOURAB HALDAR (11701617060), SRIJAN ROUTH(11701617031), MANISH KUMAR MANDAL (11701617058), SUMAN KUMAR (11701617025), a student of B.Tech in the Dept. of Electrical Engineering, RCC Institute of Information Technology (RCCIIT), Canal South Road, Beliaghata, Kolkata-700015, affiliated to Maulana Abul Kalam Azad University of Technology (MAKAUT), West Bengal, India, during the academic year 2020-21, in partial fulfillment of the requirements for the degree of Bachelor of Technology in Electrical Engineering and that this project has not submitted previously for the award of any other degree, diploma and fellowship.

HoD, Dept. of EE
RCC Institute of Information Technology
Kolkata-700015

Signature of the HOD

Name: Prof(**Dr.**)**Debasish Mandal**

Designation: HOD

subhasis bandopadhyay

Signature of the Guide

Name: **Subhasis Bandopadhyay**

Designation: Assistant Professor

Signature of the External Examiner

Name:

Designation:

# ACKNOWLEDGMENT

It is my great fortune that I have got opportunity to carry out this project work under the supervision of **Mr. Subhasis Bandopadhyay** assistant professor in the Department of Electrical Engineering, RCC Institute of Information Technology (RCCIIT), Canal South road, Beliaghata, Kolkata-700015, affiliated to Maulana Abul Kalam Azad University of Technology (MAKAUT), West Bengal, India. I express my sincere thanks and deepest sense of gratitude to my guide for his constant support, unparalleled guidance and limitless encouragement.

I wish to convey my gratitude to Prof. (Dr.) Debasish Mondal, HOD, Department of Electrical Engineering, RCCIIT and to the authority of RCCIIT for providing all kinds of infrastructural facility towards the research work.

I would also like to convey my gratitude to all the faculty members and staff of the Department of Electrical Engineering, RCCIIT for their whole hearted cooperation to make this work turn into reality.

Date: 5-Jul-2021

Place: Kolkata

Gourab Haldar.

Suman Kumar

Manish Kumar Mandal

Srijan Routh.

Full Signature of the Student

# CONTENTS

# Abstract

This project is basically used to differentiate between three species of the Iris flower which are setosa, versicolor, and virginica.The application will work on the data given to the machine if the inputs of the flowers such as petals size and sepal size are entered  There is iris data set which contains 3 classes of 50 instances each, where each class refers to a type  of iris plant.The goal here is to design a model that makes useful classifications for new flowers or, in other words, one which exhibits good generalization.From this one will be able to predict the type of iris flower after doing this project correctly. It is one of the basic machine learning applications. Therefore it will give an idea of how ML is implemented and used.

# A study of pattern recognition of Iris flower based on Machine Learning

As we all know from the nature, most of creatures have the ability to recognize the objects in order to identify food or danger. Human beings  can also recognize the types and application of objects. An interesting phenomenon could be that machines could recognize objects just like us someday in the future. This thesis mainly focuses on machine learning in pattern recognition applications.

Machine learning is the core of Artificial Intelligence (AI) and pattern recognition is also an important branch of AI. In this thesis, the conception of machine learning and machine learning algorithms are introduced. Moreover, a typical and simple machine learning algorithm called Kmeans is introduced. A case study about Iris classification is introduced to show how the Kmeans works in pattern recognition.

The aim of the case study is to design and implement a system of pattern recognition for the Iris flower based on Machine Learning. This project shows the workflow of pattern recognition and how to use machine learning approach to achieve this goal. The data set was collected from an open source website of machine learning. The programming language used in this project was Python.

**Keywords:**

Pattern Recognition, Machine Learning, K-means algorithm, Python,dataset,SciKitlearn

# 1. Introduction

Machine learning, as a powerful approach to achieve Artificial Intelligence,  has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

## 1.1 Background

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives, it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition.

In this project, the object is the Iris flower. The data set of Iris contains three different classes: Setosa, Versicolour, and Virginica. The designed recognition system will distinguish these three different classes of Iris.

## 1.2 Objectives

After the project has been settled, the computer should have the ability to aggregate three different classifications of Iris flower to three categories. The whole workflow of machine learning should work smoothly. The users do not need to tell the computer which class the Iris belongs to, the computer can recognize them all by itself.

The final purpose of this project is to let everyone who read this thesis have a basic understanding of machine learning. Even through someone never touched this field, they can realize that the machine learning algorithm will become more popular and useful in the future. Moreover, the case study of Iris recognition will show how to implement machine learning by using Scikit-learn software.

## 1.3 Collecting data set

The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. Each class is linearly separable from the other two classes. The attribute information will include sepal length, sepal width, and petal length and petal width. All of them have the same unit, *cm*.

### 1.4 Using K-means algorithm to achieve clustering

K-means algorithm was used for clustering Iris classes in this project. There are many different kinds of machine learning algorithms applied in different fields. Choosing a proper algorithm is essential for each machine learning project. For pattern recognition, K-means is a classic clustering algorithm. In this project, Kmeans algorithm can be implemented with the Python programming language.

### 1.5 Evaluating result

Evaluation will be the final part of this project. For each scientific project, the final result should be tested and evaluated if that is acceptable. The result will be automatically shown in the end of the program execution. For every machine learning algorithm, exceptions will always exist. In order to find the best result, result analyzing is necessary.

# 2. Literature review

## 2.1 Basic introduction to machine learning

Learning is a very important feature of Artificial Intelligence. Many scientists tried to explain and give a proper definition for learning. However, learning is not that easy to cover with few simple sentences. Many computer scientists, sociologists, logicians and other scientists discussed about this for a long time. Some scientists think learning is an adaptive skill so that the system can perform the similar task better in the next time(Simon 1987). Others claim that learning is a process of collecting knowledge(Feigenbaum 1977). Even though there is no proper definition for learning skill, we still need to give a definition for machine learning. In general, machine learning aims to find out how the computer algorithms can be improved automatically through experience(Mitchell 1997).

Machine learning has an important position in the field of Artificial Intelligence. At the beginning of development of Artificial Intelligence(AI), the AI system does not have a thorough learning ability so the whole system is not perfect. For instance, a computer cannot do self-adjustment when it faces problems. Moreover, the computer cannot automatically collect and discover new knowledge. The inference of the program needs more induction than deduction. Therefore, computer only can figure out already existing truths. It does not have the ability to discover a new logical theory, rules and so on.

### 2.11 Fundamental structure of machine learning system

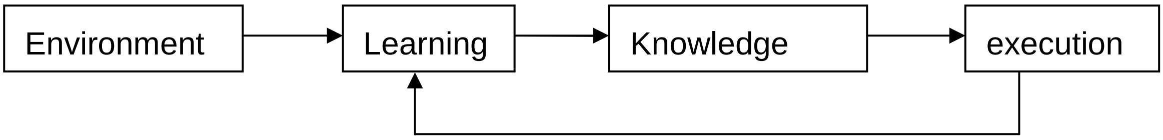


Figure 1. Learning system structure

Figure 1 shows the basic work structure of machine learning. The structure of machine learning system consists of four main parts: Environment, Learning, Knowledge base and Execute.

The environment represents a combination of information from external information source. That would include any information from persons or references materials and so on. It is the learning source for the whole machine learning system. The environment is responsible for transferring data to the system. The quality of the data is very important. In the reality, the data can be complex so it will be difficult for computer to process. In

addition, the data can be incomplete, therefore the illation from the learning system is unauthentic.
Learning is the procedure of transferring the information from the environment to knowledge. The environment will give the computer external information, and then the computer will go through all the information by using analysis, comprehensive induction and analogy to process this information to knowledge. At last, all the knowledge would be imported to the knowledge base.

The knowledge base can be treated as the brain of the whole machine learning system. Different kinds of form and content of knowledge can have different influence on the designing of a machine learning system. Knowledge representation modes are eigenvector, First-order logic statements, production rule, and semantic system. Every mode has its own advantages and disadvantages. Therefore, when users want to design a machine leaning system, a good knowledge representation mode is very important for the whole system.

A proper knowledge representation mode should satisfy four basic requirements:

1. Strong expression
2. Easy theorization
3. Easy to modify the knowledge base
4. Easy to expand the knowledge represenation

Moreover, a machine learning system cannot create new knowledge from nothing. It always needs original knowledge to understand the information from environment. Then the computer can use this information to learn new knowledge step by step. In conclusion, learning process in the whole system is a process of expansion and perfection of the knowledge base.

Execution is the core of the whole machine learning system. Each part of the system aims to make a progress for the execution part. On the other hand, execution also has a connection to each part, especially the learning process. The purpose of a learning process is to make the execution perfect. At last, the complexity, feedback and transparency of execution also has an influence on the learning process.

Complexity

The complexity of knowledge is different depending on the different learning tasks. Some tasks are quite easy, so the system does not need too much information. If the tasks are quite difficult, the system will need more information to learn.

Feedback

After the execution, the execution system can evaluate the leaning task, and then give feedback information to the learning process. The learning process will try to decide whether to collect information from environment to modify or improve the knowledge in knowledge base or not based on the feedback.

Transparency

From the result of execution part, users can easily see the structure of the knowledge base and give the evaluation for it.

### 2.1.2 The applications of Machine Learning.

Machine learning as a very likely approach to achieve human-computer integration and can be applied in many computer fields. Machine learning is not a typical method as it contains many different computer algorithms. Different algorithms aim to solve different machine learning tasks. At last, all the algorithms can help the computer to act more like a human.

Machine learning is already applied in many fields, for instance, pattern recognition, Artificial Intelligence, computer vision, data mining, text categorization and so on. Machine learning gives a new way to develop the intelligence of the machines. It also becomes an easier way to help people to analyse data from huge data sets.

## 2.2 The description of Machine Learning forms

A learning method is a complicated topic which has many different kinds of forms. Everyone has different methods to study, so does the machine. We can categorize various machine learning systems by different conditions. In general, we can separate learning problems in two main categories: supervised learning and unsupervised learning.

### 2.2.1 Supervised learning

Supervised learning is a commonly used machine learning algorithm which appears in many different fields of computer science. In the supervised learning method, the computer can establish a learning model based on the training data set. According to this learning model, a computer can use the algorithm to predict or analyze new information. By using special algorithms, a computer can find the best result and reduce the error rate all by itself. Supervised learning is mainly used for two different patterns: classification and regression. In supervised learning, when a developer gives the computer some samples, each sample is always attached with some classification information. The computer will analyze these samples to get learning experiences so that the error rate would be reduced when a classifier does recognitions for each patterns.

Each classifier has a different machine learning algorithm. For instance, a neutral network algorithm and a decision tree learning algorithm suit to two different classifiers. They have their own advantages and disadvantages so that they can accomplish different learning objectives.

### 2.2.2 Unsupervised learning

Unsupervised learning is also used for classification of original data.

The classifier in the unsupervised learning method aims to find the classification information for unlabeled samples. The objective of unsupervised learning is to let the computer learn it by itself. We do not teach the computer how to do it. The computer is supposed to do analyzing from the given samples.

In unsupervised learning, the computer is not able to find the best result to take and also the computer does not know if the result is correct or not. When the computer receives the original data, it can find the potential regulation within the information automatically and then the computer will adopt this regulation to the new case. That makes the difference between supervised learning and unsupervised learning.

In some cases, this method is more powerful than supervised learning. That is because there is no need to do the classification for samples in advance. Sometimes, our classification method may not be the best one. On the other hand, a computer may find out the best method after it learns it from samples again and again.

## 2.3 Machine Learning in pattern recognition

As mentioned above, the method of machine learning can also be used in pattern recognition. In fact, pattern recognition really needs machine learning to achieve its objective.

Both supervised learning and unsupervised learning are useful for pattern recognition, for example, in this thesis, K-means clustering algorithm in unsupervised learning. The Kmeans clustering algorithm is always used for image segmentation. The image segmentation is so important for image pattern recognition. Because of the technology of image segmentation, it is easier to do the image analyzing so that it will achieve much better results for image pattern recognition.

Moreover, the technology of machine learning has been used in almost every field in pattern recognition. For example, image pattern recognition, voice recognition, fingerprint recognition, character recognition and so on. They all need machine learning algorithms to select features from the objects and to do the analyzing.

### 2.3.1 Basic introduction to pattern recognition

Pattern Recognition is a fundamental human intelligence. In our daily life, we always do ‘pattern recognition’, for instance, we recognize faces and images. Basically, pattern recognition refers to analyzing information and identifying for any kind of forms of visual or phenomenon information. Pattern recognition can describe, recognize, classify and explain the objects or the visual information.

As machine learning, pattern recognition, can be treated as two different classification methods: supervised classification and unsupervised classification. They are quite similar to supervised learning and unsupervised learning. As supervised classification needs a teacher that gives the category of samples, the unsupervised classification is doing it the other way around.

Pattern recognition is related to statistics, psychology, linguistics, computer science, biology and so on. It plays an important role in Artificial Intelligence and image processing.

### 2.3.2 Machine learning algorithm in pattern recognition.

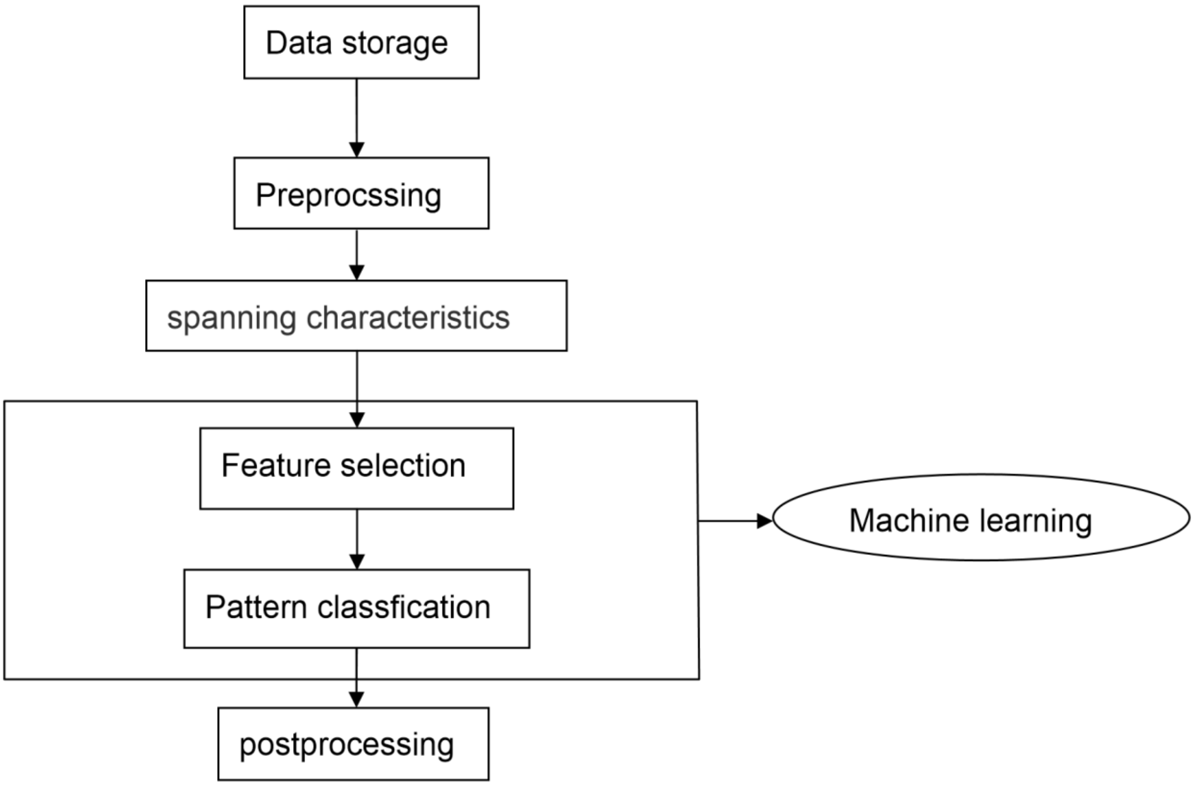


Figure 2. Pattern recognition framework

From Figure 2, we can see that feature selection and pattern classification are the main parts of the whole pattern recognition system.

The core of machine learning is mainly about searching. For different types of patterns, machine learning needs a suitable method to find the proper feature from all information. In order to achieve this, many scientists create many kinds of machine learning algorithms. These algorithms are made for feature selection and pattern classification. For instance, the genetic algorithm, the neural network algorithm, SVM, the K-nearest neighbor algorithm, all support  different types of learning objectives. The process of feature selection is so important that it can have a great effect on the result of pattern recognition. Sometimes, the property of objects is so different. If the selection algorithm is not chosen right, the result of the pattern classification will be different or bad. Bad algorithms could cause plenty of information redundancy. Some useful data are not used. On the contrary, some unuseful data may be used for feature selection. In this case, in the processing of pattern classification, the computer will classify objects in an inappropriate way with many errors. The end, the result would not be acceptable.

# 3. K-means clustering

As mentioned earlier in this thesis, machine learning consists of many kinds of learning algorithms for different learning methods. In this thesis, the classification information is assumed to be unlabeled. In this case, the best choice in unsupervised learning is the Kmeans clustering algorithm.

## 3.1 Introduction to clustering

The K-means clustering algorithm is one of the most popular clustering algorithms in the world. Clustering aims to classify data from the whole data space. The difference between each data object in the same class is similar.

However, the difference between each data objects in different classes is large. Clustering belongs to the unsupervised learning method and it can automatically sort data sets.

Basically, the result of clustering algorithm is to find the same classification of different data in the whole data sets. For example, the data set contains monkey, lion, banana, apple, four different data units. After clustering, these four data will be divided into two main sections. One section includes monkey and lion representing the class of animals. The other section includes apple and banana, this section representing the class of fruits.

A clustering algorithm groups all the same kind of data into one single class. The computer will recognize the specific features of all data so that it can separate data to the proper classes.

## 3.2 K-means algorithm

The K-means algorithm is based on the distance from each data to the initial cluster centers. The distance is the evaluation standard for the similarity of the data. This means that if the distance between two objects is small, then the similar level is high.

In the K-means algorithm, an initial set of k as the clustering point is chosen.

The computer will find all the data which is close to the initial set of k. After that, by using the method of iteration, the computer will update the value of k to get the new cluster for the rest of data. Then, the computer will retrieve the best result after running it again and again.

The formula of K-means algorithm is as following:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad (1)$$

All the results will be related to the initial set of k, therefore the random value of k is very important to the whole algorithm system.

Suppose we have a data set {$X_1$,$X_2$,……$X_n$} consisting of N observations of a random variable x.

1) We choose a number of k cluster centroids from N observations as the initial clustering centroid, $X_1$(1),$X_2$(1),……$X_K$(1), here (1) means the number of times of iteration.

2) Based on the means of each observation, we calculate the distance between each observation to the initial clustering point. According to the rule of minimum distance, we distribute every sample to the one of the k cluster centroids.

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2. \qquad (2)$$

3) We calculate the vector value of each cluster centroid. $X_j$(k+1), j = 1,2,…K. Then we calculate and take the value of sample mean vector as the new cluster centroid.

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}. \qquad (3)$$

4) We loop step 2 and 3 until every cluster does not change anymore.

### 3.2.1 K-means algorithm workflow

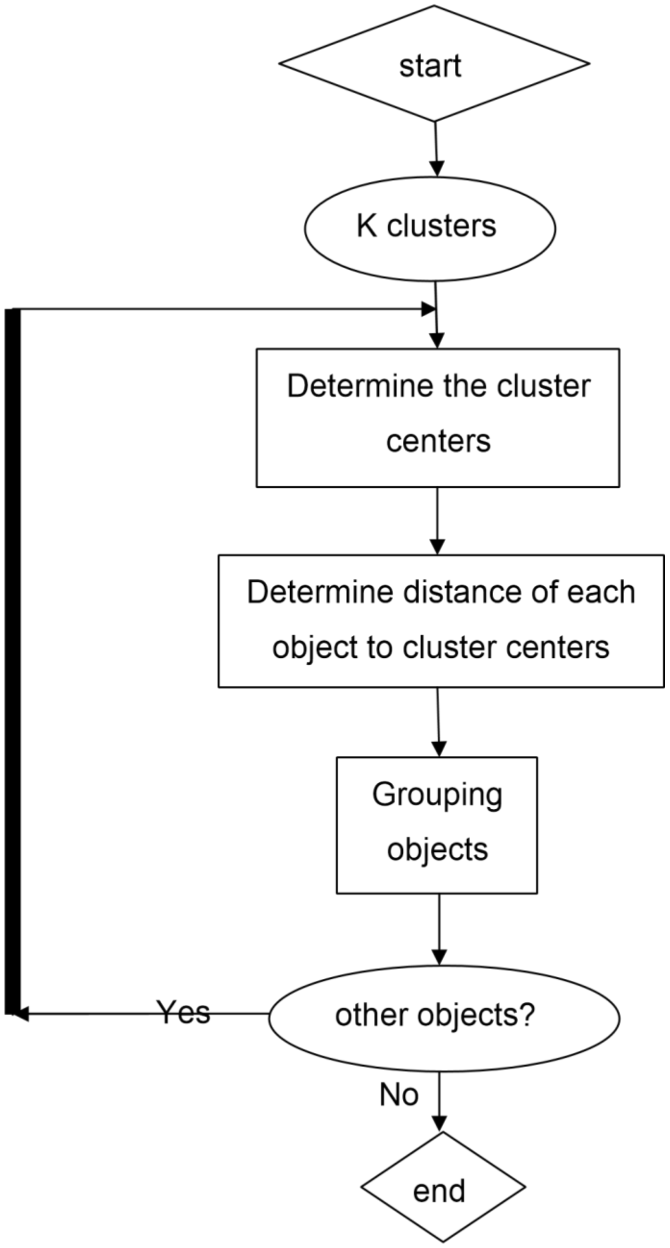

start
K clusters
Determine the cluster centers
Determine distance of each object to cluster centers
Grouping objects
other objects?
Yes
No
end

Figure 3. K-means algorithm workflow

Figure 3 shows the workflow of K-means algorithm. In the very first beginning, the system will choose a number of k clusters from a number of N observations. In the next, for the rest of the objects, the system will distributes these objects to the closest clusters based on the mini distance between objects to the cluster center. Moreover, it will calculate the means of all objects in the same cluster to get the new cluster center. These two steps are repeated until the formula (3) convergences. In general, the equation (3) is based on  mean square deviation theory.

The following tables show a sample of workflow of K-means. The dataset contains 30 samples and the number of clusters is 3.

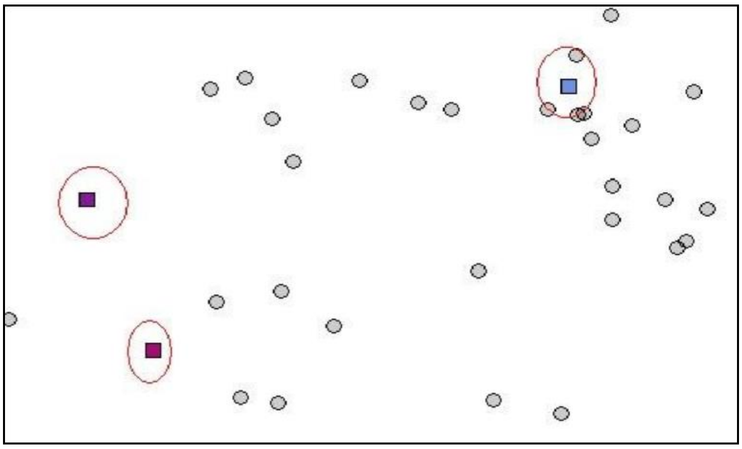


Figure 4. K-means clustering step 1

Now the system generates three cluster points with randomly. There are three different colours: purple(Top left), blue(Top right), pink(Bottom left). These three colours stand for three different clusters.

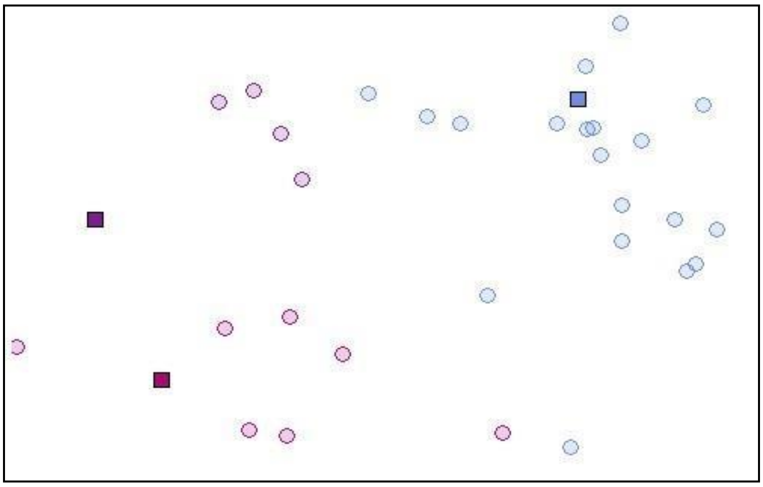


Figure 5. K-means clustering step 2

With the initial the point of k, then the system should calculate the distance of each object to the cluster centers. The new blank box indicates the new cluster center.

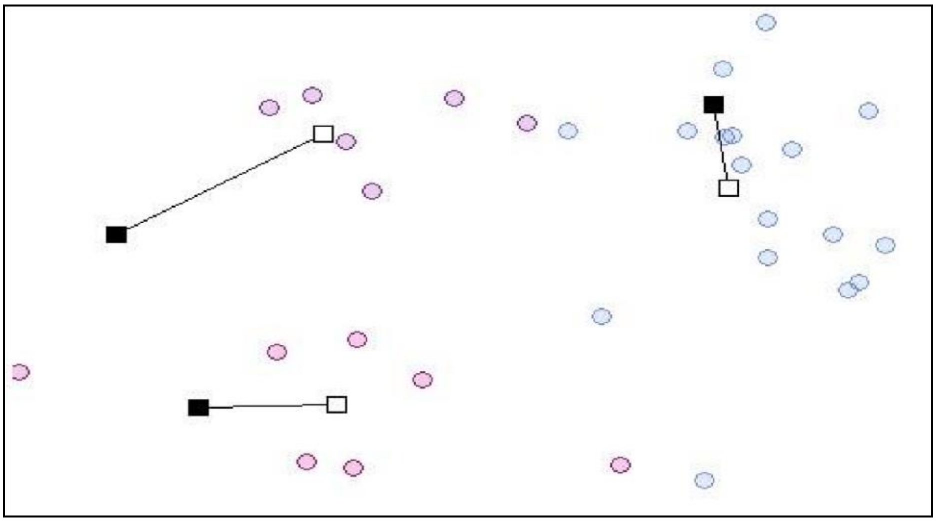


Figure 6. K-means clustering step 3

If there are still some objects missing, then the system will continue to find the new centroid for each cluster until all the samples are grouped. The system will loop equation (2) and (3) until the k cluster centroids will not move any longer. Therefore, In Figure 7, the k cluster centroids move to a new place and the calculation is continued.

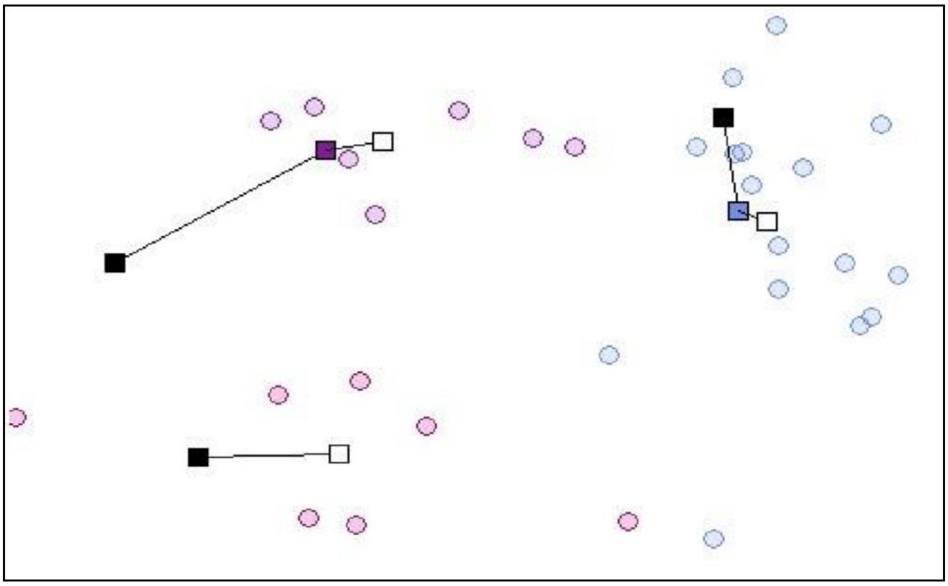


Figure 7. K-means clustering step 4

The next table is the final result. The principle of K-means algorithm is to make all samples in one cluster to be closer to each other, but the distance of each clusters should be larger.

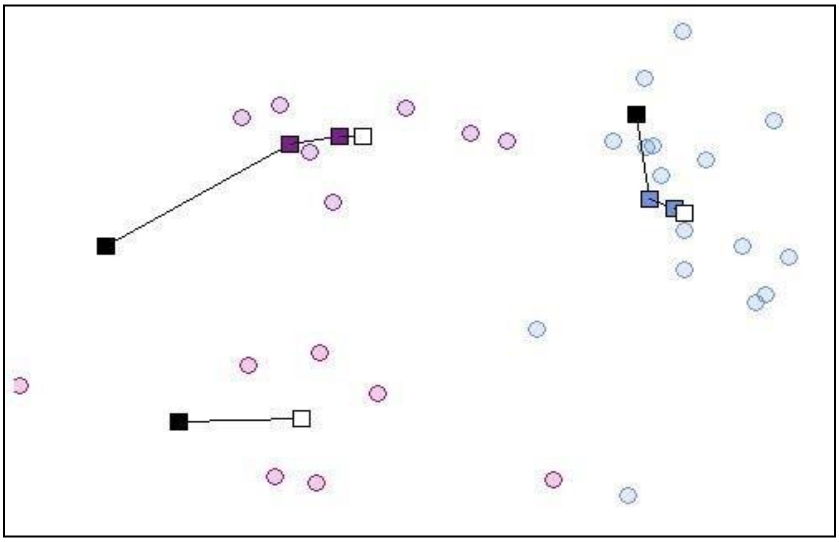


Figure 8. K-means clustering step 5

### 3.2.2 The advantages and disadvantages of K-means

Every machine learning algorithm has advantages and disadvantages. Here are the advantages and disadvantages of K-means.

Advantages of K-means:

If the number of variables is large, K-means computes faster than other clustering algorithms.

K-means can make clusters tighter if the centroid can be found properly.

Disadvantages of K-means:

The value of k is too difficult to choose. Sometimes, the amount of types of dataset is unknown.

Different initialization number affects output of cluster results.

### 3.2.3 Why choosing the K-means clustering algorithm

The K-means clustering algorithm belongs to unsupervised learning. If the classification information is not given, we do not know what kind of types of the object exist in the data set. But we know how many classes exist in the dataset. If we know the number of classes in the dataset, then we know the number of clusters. In this case, we should choose unsupervised learning to find out which sample belongs to which cluster.
The K-means algorithm is the simplest clustering algorithm. For this case, the data set contains 50 samples for each type of Iris flower. That means each type has the same amount

of samples so that the centroid would be easier to calculate. The K-means algorithm is good enough for this case.

# 4. Implementation

## 4.1 Python

Python is a programming language created by Guido van Rossum in 1989. Python is an interpreted, object-oriented, dynamic data type of high-level programming languages.(Python Software Foundation 2013). The programming language style is simple, clear and it also contains powerful different kinds of classes. Moreover, Python can easily combine other programming languages, such as C or C++.

As a successful programming language, it has its own advantages:

**Simple&easy to learn**: The concept of this programming language is as simple as it can be. That makes it easy for everyone to learn and use. It is easy to understand the syntax.

**Open source**: Python is completely free as it is an open source software. Several of open source scientific computing storage has the API for Python. Users can easy to install Python on their own computer and use the standard and extend library.

**Scalability:** Programmers can write their code in C or C++ and run them in Python.

## 4.2 SciKit-learn

Scikit–learn is an open source machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms and is designed to interoperate with the Python numerical libraries NumPy and SciPy (Pedregosa et al. 2011*)*. SciKit-learn contains the Kmeans algorithm based on Python and it helps to figure out how to implement this algorithm in programming.

## 4.3 Numpy, Scipy and Matplotlib

In Python, there is no data type called array. In order to implement the data type of array with python, numpy and scipy are the essential libraries for analyzing and calculating data. They are all open source libraries. Numpy is mainly used for the matrix calculation. scipy is developed based on numpy and it is mainly used for scientific research.

By using them in Python programming, they can be used with two simple commands:

```
>>> import numpy
>>> import scipy
```

Then Python will call the methods from numpy and scipy.

Mathplotlib is a famous library for plotting in Python. It provides a series of API and it is suitable for making interactive mapping. In this case, we need to use it to find the best result visually.

## 4.4 Preparing the Iris flower data set

The data set of Iris flower can be found in UCI Machine Learning Repositor (Bache & Lichman 2013). In this thesis, the famous Fisher's Iris data set will be used.

The data set of Iris flower can be also found in the Scikit-learn library. In sitepackages, there is a folder named sklearn. In this folder, there is a datasets subfolder to contain many kinds of data sets for machine learning study.

The data set can be found in Appendix 1.

In the species of this table, 0 represents setosa, 1 represents versicolor, 2 represents virginica. In the process of preparing a training data set and a testing data set, the greatest problem is how to find the most appropriate way to divide the data set into training data set and testing data set. In some cases, by using sampling theory and estimation theory, we can separate the whole data set into training data set and testing data set. However, sometimes, the method would be changed. The attributes and the property of the data set would be different in various machine learning objects. Thus, in this kind of situation, in order to achieve a better result of machine learning, the data set will be separated according to the property of attributes of the data set.

The K-means algorithm and unsupervised learning does not use a training data set to compute the training sample. Therefore, there is no need to separate the dataset into a training data set and a testing data set. It can simply use this dataset to get the result of clustering.

## 4.5 Machine learning system design

In general, the principles of machine learning system design should follow two basic requirements :

the model selection and creation and the learning algorithm selection and design.

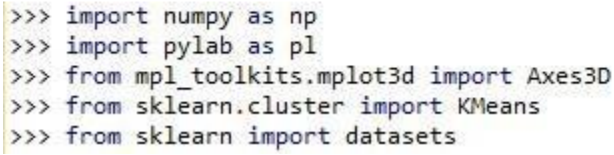
In addition, different models can have different learning systems. On the other hand, the objective function is also different in different learning models. The objective function can help the machine to establish a learning system. Moreover, the accuracy and complexity of different algorithms would be the most important factor of the learning system. If the chosen algorithm is not very adaptive to the learning system, then the efficiency and result of the learning system would be reduced. The selection of training data set can have an influence on learning performance and feature selection.

## 4.6 Using Python to implement the program

For good implementation and good compatibility, Python version 2.7 will be in use. The Integrated Development Environment in this case will be PyScripter.

By using the Scikit-learn software package, there is no need to write a program to implement the K-means algorithm. After the installation has been finished, the K-means algorithm source code can be found in sklearn library. The source code of K-means clustering of Iris recognition can be found in the official website of Scikit-learn.

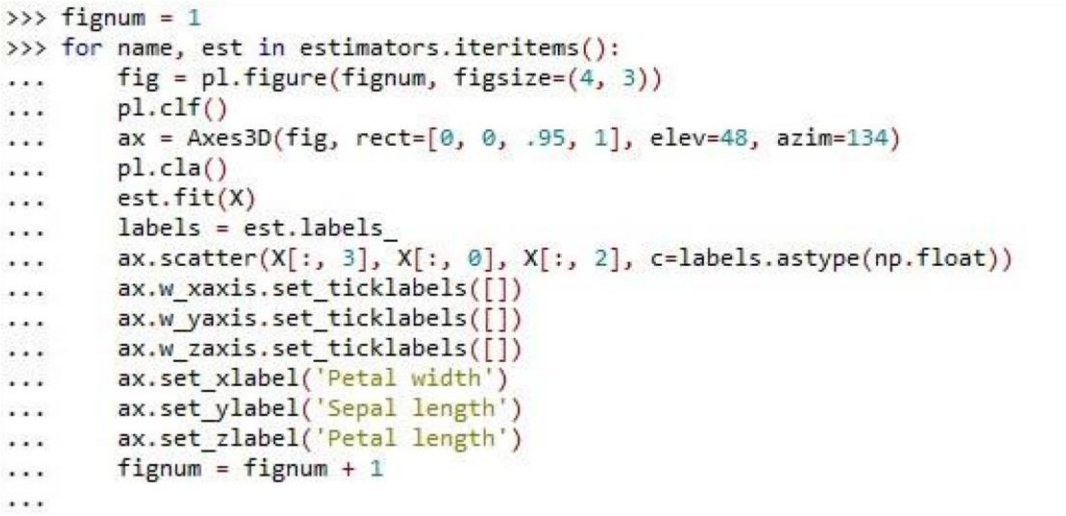
First of all, we need to import the library of numpy, dataset of Iris, K-means and Axes3D into the program. These are needed for this program. Numpy can helps to implement the K-means algorithm, the Iris dataset is the main data to be analyzed, Axes3d can make 3D outputs of this program, and the image will be more visual.



```
>>> import numpy as np
>>> import pylab as pl
>>> from mpl_toolkits.mplot3d import Axes3D
>>> from sklearn.cluster import KMeans
>>> from sklearn import datasets
```

Then, the program loads the Iris dataset and sets the centroid value and the number of clusters.

The result is shown as a table with three feature vectors. The feature vectors consists of petal width, sepal length and petal length. The output table will be three-dimensional.



```
>>> fignum = 1
>>> for name, est in estimators.iteritems():
...     fig = pl.figure(fignum, figsize=(4, 3))
...     pl.clf()
...     ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
...     pl.cla()
...     est.fit(X)
...     labels = est.labels_
...     ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=labels.astype(np.float))
...     ax.w_xaxis.set_ticklabels([])
...     ax.w_yaxis.set_ticklabels([])
...     ax.w_zaxis.set_ticklabels([])
...     ax.set_xlabel('Petal width')
...     ax.set_ylabel('Sepal length')
...     ax.set_zlabel('Petal length')
...     fignum = fignum + 1
...
```

## 5. Evaluating results

The result is shown in four images for the clustering results. Figure 9 will be the result with eight clusters. Figure 10 shows the result with three clusters.

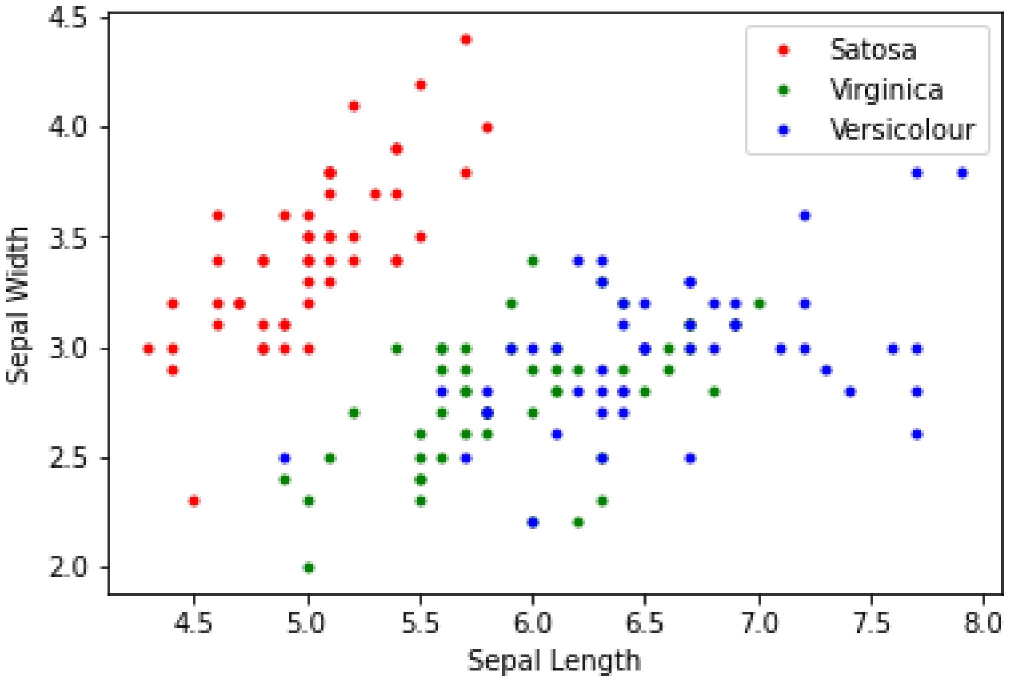


Figure 9. Clustering of Iris dataset with eight clusters

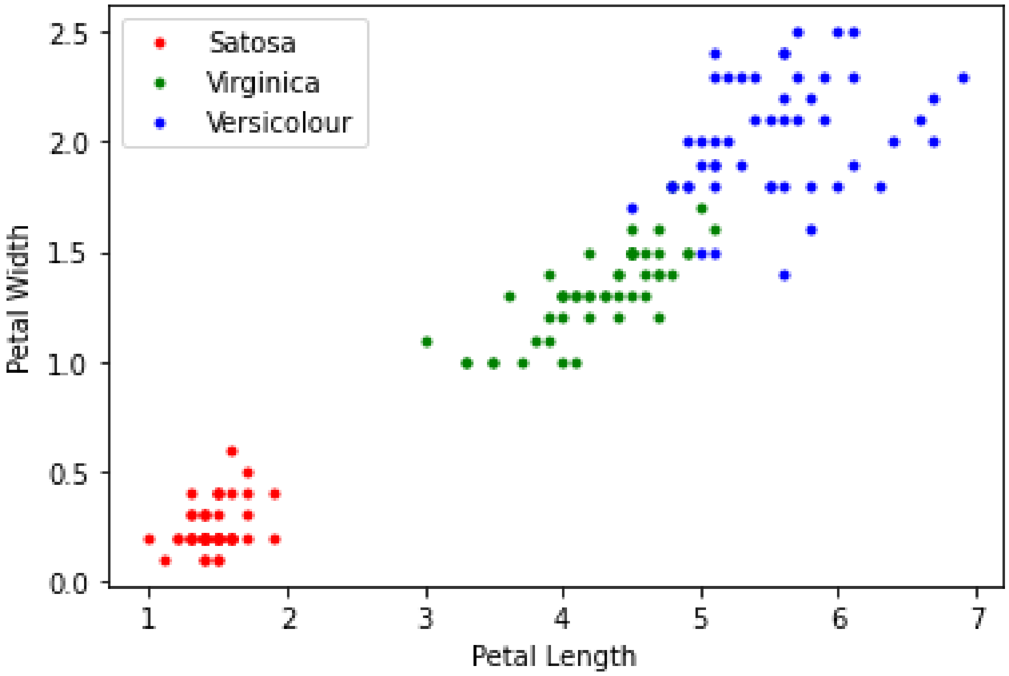


Figure 10. Clustering of Iris dataset with three clusters

As seen in Figure 9 and 10, the whole dataset is separated into eight clusters in

Figure 9 and three clusters are shown in Figure 10 with different colors. In Figure 9, most of the samples stick together, it is really hard to distinguish them very clearly. The differences between each sample is small. In this case, the cluster result is not acceptable. On the other hand, in Figure 10, it can be easily seen that the cluster result is much better than in Figure 9. Even though there are still some overlapping parts between green and purple, but it quite clear to see the difference between these three clusters. This case shows the importance of choosing the number of clusters for K-means algorithm. Sometimes for the real datasets, it is difficult to know how many data sets should be used. Therefore, it is quite hard to choose the number of clusters. One method is to use the ISODATA algorithm, through the merging and division of clusters to obtain a reasonable number of k.

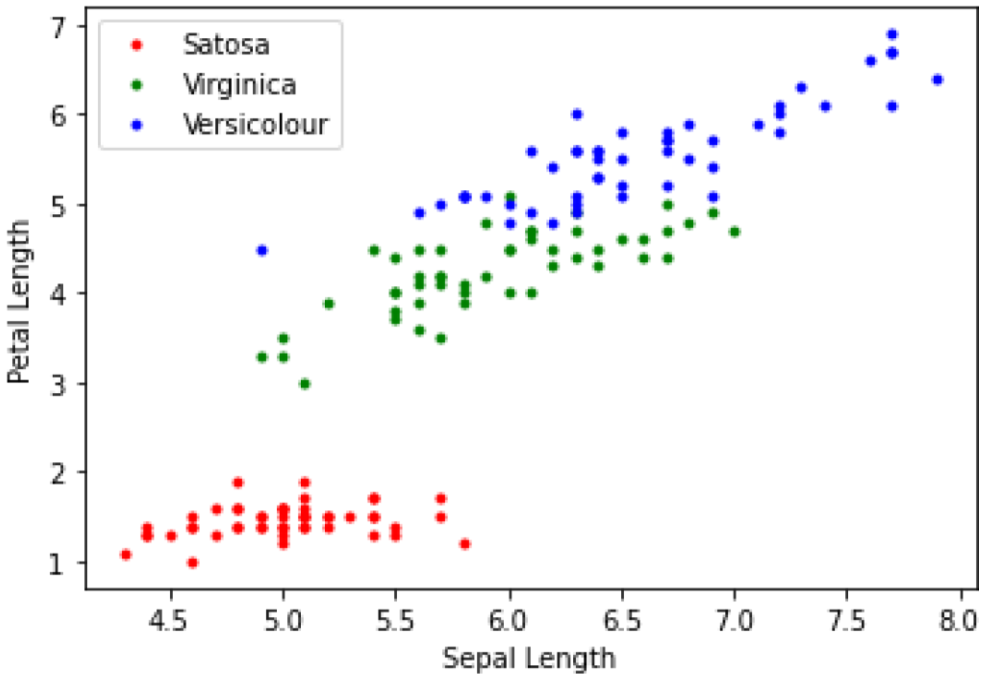


Figure 11. Clustering of Iris dataset with bad initialization

Figure 11 , shows the cluster result with three clusters but bad initialization. We can see that some of the samples change their class compare to the Figure 10. With a random initialization number, the system will obtain different cluster results. Therefore, a random initialization number is very important for a good cluster result. However, we do not know what could be a good initialization number. In this case, in some machine learning systems, the scientists will choose GA(Genetic Algorithm) to have the initialization point.

Figure 12 below illustrates a standard result of K-means clustering of Iris recognition. The term “ground truth” refers to the classification of training datasets in supervised learning. The number of clusters are three and with a good initialization point. This is the best classification of all shown here. The whole dataset has been separated properly and each dataset has good differences. In Figure 10, it shows the stardard result of classification in unsupervised learning. Compare to this figure, Figure 10 still has some small differences but it still works very well. Almost every data belongs to the right place.

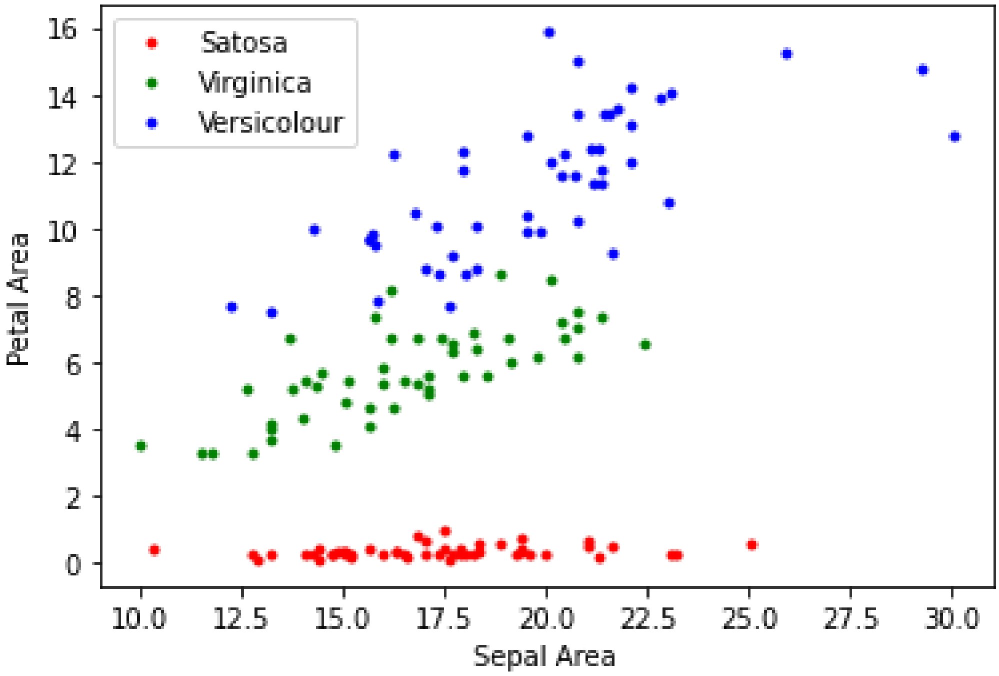


Figure 12. Clustering of Iris dataset in ground truth

These results show the effect that the number of k and the random initialization number have on the clustering result. It is also possible to see the advantages and disadvantages of the K-means clustering algorithm.

# 6. The future prospects

The Iris recognition case study above shows that the Machine Learning algorithm works well in this pattern recognition. The speed of computing is fast and the result is acceptable. However, the K-means clustering algorithm is just one of the clustering algorithm in unsupervised learning. There are more algorithms for different work objectives in different scientific fields.

As it is mentioned above, Machine Learning can be separated into supervised learning and unsupervised learning. However, sometimes, a whole dataset have both labeled data and unlabeled data. In order to process this kind of dataset, a new learning method called Semi-supervised(SSL) Learning has become a research hotspot. Because of this learning method, both machine learning and pattern recognition have a new research direction. It saves a lot of time and human resource to label those large amounts of unlabeled data. The SSL is also significant on  improving learning performance of a computer.

Moreover, a learning system always consists of two parts, learning and environment. The environment gives knowledge to the computer and the computer will transfer this knowledge and store them and select useful information to implements different learning objectives. Therefore, different learning strategies  can also be separated into rote learning, learning from instruction, learning by deduction, learning by analog, explanation-based learning and learning from induction. All of them have different algorithms to process different work objectives.

The implemented case in this thesis is only a simple example of machine learning and pattern recognition. Moreover, the K-means algorithm used in this thesis is a basic algorithm. However, if the data set has many feature dimensions and it is complicated, and if the learning objective is not that simple, the K-means algorithm can not be used.

Nowadays, GA (Genetic Algorithm), Artificial neural network and other machine learning algorithms have become more and more stable and useful. Many scientists are working on improving the performance of machine learning algorithms.The K-means has also its own improved parts. The K-means can also be used along with other algorithms, such as ISODATA ,EM and Kmeans++. A better machine learning algorithm can obtain better results for pattern recognition. As the technology of pattern recognition develops, it requires more professional and more perfect machine learning algorithms. In this case, machine learning has a huge potential for growth.

In general, besides pattern recognition, machine learning can also be widely used in many fields of computer science and Artificial Intelligence. More and more Artificial Intelligence products are coming out on the market. Nowadays, people can use the Artificial Intelligence products every day. For example, people use Google search for seeking information which it is also based on the clustering algorithm of machine learning. All in all, machine learning definitely has a bright prospect.

# 7. Conclusion

With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used. The work also shows how to use SciKit-learn software to learn machine learning.

# References

Google Colab

SKlearn kit

Iris flower data set(wikipedia) www.youtube.com

Bishop, C. 2006. Pattern Recognition and Machine Learning. New York:

Springer, pp.424-428.

Fisher, R.A. 1936. UCI Machine Learning Repository: Iris Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013

Improved Outcomes Software.,2004. K-Means Clustering Overview, Available at: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-

Means_Clustering_Overview.htm. Consulted 22 AUG 2013 Mitchell, T. 1997. Machine learning. McGraw Hill.

# APPENDIX

## 1. Fisher’s Iris flower dataset

| Sepal length | Sepal width | Petal length | Petal width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | 0 |
| 4.9 | 3 | 1.4 | 0.2 | 0 |
| 4.7 | 3.2 | 1.3 | 0.2 | 0 |
| 4.6 | 3.1 | 1.5 | 0.2 | 0 |
| 5 | 3.6 | 1.4 | 0.2 | 0 |
| 5.4 | 3.9 | 1.7 | 0.4 | 0 |
| 4.6 | 3.4 | 1.4 | 0.3 | 0 |
| 5 | 3.4 | 1.5 | 0.2 | 0 |
| 4.4 | 2.9 | 1.4 | 0.2 | 0 |
| 4.9 | 3.1 | 1.5 | 0.1 | 0 |
| 5.4 | 3.7 | 1.5 | 0.2 | 0 |
| 4.8 | 3.4 | 1.6 | 0.2 | 0 |
| 4.8 | 3 | 1.4 | 0.1 | 0 |
| 4.3 | 3 | 1.1 | 0.1 | 0 |
| 5.8 | 4 | 1.2 | 0.2 | 0 |
| 5.7 | 4.4 | 1.5 | 0.4 | 0 |
| 5.4 | 3.9 | 1.3 | 0.4 | 0 |
| 5.1 | 3.5 | 1.4 | 0.3 | 0 |
| 5.7 | 3.8 | 1.7 | 0.3 | 0 |
| 5.1 | 3.8 | 1.5 | 0.3 | 0 |
| 5.4 | 3.4 | 1.7 | 0.2 | 0 |
| 5.1 | 3.7 | 1.5 | 0.4 | 0 |
| 4.6 | 3.6 | 1 | 0.2 | 0 |
| 5.1 | 3.3 | 1.7 | 0.5 | 0 |
| 4.8 | 3.4 | 1.9 | 0.2 | 0 |
| 5 | 3 | 1.6 | 0.2 | 0 |
| 5 | 3.4 | 1.6 | 0.4 | 0 |
| 5.2 | 3.5 | 1.5 | 0.2 | 0 |
| 5.2 | 3.4 | 1.4 | 0.2 | 0 |
| 4.7 | 3.2 | 1.6 | 0.2 | 0 |
| 4.8 | 3.1 | 1.6 | 0.2 | 0 |
| 5.4 | 3.4 | 1.5 | 0.4 | 0 |
| 5.2 | 4.1 | 1.5 | 0.1 | 0 |
| 5.5 | 4.2 | 1.4 | 0.2 | 0 |
| 4.9 | 3.1 | 1.5 | 0.1 | 0 |
| 5 | 3.2 | 1.2 | 0.2 | 0 |
| 5.5 | 3.5 | 1.3 | 0.2 | 0 |

| | | | | |
|---:|---:|---:|---:|---:|
| 4.9 | 3.1 | 1.5 | 0.1 | 0 |
| 4.4 | 3 | 1.3 | 0.2 | 0 |
| 5.1 | 3.4 | 1.5 | 0.2 | 0 |
| 5 | 3.5 | 1.3 | 0.3 | 0 |
| 4.5 | 2.3 | 1.3 | 0.3 | 0 |
| 4.4 | 3.2 | 1.3 | 0.2 | 0 |
| 5 | 3.5 | 1.6 | 0.6 | 0 |
| 5.1 | 3.8 | 1.9 | 0.4 | 0 |
| 4.8 | 3 | 1.4 | 0.3 | 0 |
| 5.1 | 3.8 | 1.6 | 0.2 | 0 |
| 4.6 | 3.2 | 1.4 | 0.2 | 0 |
| 5.3 | 3.7 | 1.5 | 0.2 | 0 |
| 5 | 3.3 | 1.4 | 0.2 | 0 |
| 7 | 3.2 | 4.7 | 1.4 | 1 |
| 6.4 | 3.2 | 4.5 | 1.5 | 1 |
| 6.9 | 3.1 | 4.9 | 1.5 | 1 |
| 5.5 | 2.3 | 4 | 1.3 | 1 |
| 6.5 | 2.8 | 4.6 | 1.5 | 1 |
| 5.7 | 2.8 | 4.5 | 1.3 | 1 |
| 6.3 | 3.3 | 4.7 | 1.6 | 1 |
| 4.9 | 2.4 | 3.3 | 1 | 1 |
| 6.6 | 2.9 | 4.6 | 1.3 | 1 |
| 5.2 | 2.7 | 3.9 | 1.4 | 1 |
| 5 | 2 | 3.5 | 1 | 1 |
| 5.9 | 3 | 4.2 | 1.5 | 1 |
| 6 | 2.2 | 4 | 1 | 1 |
| 6.1 | 2.9 | 4.7 | 1.4 | 1 |
| 5.6 | 2.9 | 3.6 | 1.3 | 1 |
| 6.7 | 3.1 | 4.4 | 1.4 | 1 |
| 5.6 | 3 | 4.5 | 1.5 | 1 |
| 5.8 | 2.7 | 4.1 | 1 | 1 |
| 6.2 | 2.2 | 4.5 | 1.5 | 1 |
| 5.6 | 2.5 | 3.9 | 1.1 | 1 |
| 5.9 | 3.2 | 4.8 | 1.8 | 1 |
| 6.1 | 2.8 | 4 | 1.3 | 1 |
| 6.3 | 2.5 | 4.9 | 1.5 | 1 |
| 6.1 | 2.8 | 4.7 | 1.2 | 1 |
| 6.4 | 2.9 | 4.3 | 1.3 | 1 |
| 6.6 | 3 | 4.4 | 1.4 | 1 |
| 6.8 | 2.8 | 4.8 | 1.4 | 1 |
| 6.7 | 3 | 5 | 1.7 | 1 |
| 6 | 2.9 | 4.5 | 1.5 | 1 |
| 5.7 | 2.6 | 3.5 | 1 | 1 |
| 5.5 | 2.4 | 3.8 | 1.1 | 1 |
| 5.5 | 2.4 | 3.7 | 1 | 1 |
| 5.8 | 2.7 | 3.9 | 1.2 | 1 |
| 6 | 2.7 | 5.1 | 1.6 | 1 |

| | | | | |
|---:|---:|---:|---:|---:|
| 5.4 | 3 | 4.5 | 1.5 | 1 |
| 6 | 3.4 | 4.5 | 1.6 | 1 |
| 6.7 | 3.1 | 4.7 | 1.5 | 1 |
| 6.3 | 2.3 | 4.4 | 1.3 | 1 |
| 5.6 | 3 | 4.1 | 1.3 | 1 |
| 5.5 | 2.5 | 4 | 1.3 | 1 |
| 5.5 | 2.6 | 4.4 | 1.2 | 1 |
| 6.1 | 3 | 4.6 | 1.4 | 1 |
| 5.8 | 2.6 | 4 | 1.2 | 1 |
| 5 | 2.3 | 3.3 | 1 | 1 |
| 5.6 | 2.7 | 4.2 | 1.3 | 1 |
| 5.7 | 3 | 4.2 | 1.2 | 1 |
| 5.7 | 2.9 | 4.2 | 1.3 | 1 |
| 6.2 | 2.9 | 4.3 | 1.3 | 1 |
| 5.1 | 2.5 | 3 | 1.1 | 1 |
| 5.7 | 2.8 | 4.1 | 1.3 | 1 |
| 6.3 | 3.3 | 6 | 2.5 | 2 |
| 5.8 | 2.7 | 5.1 | 1.9 | 2 |
| 7.1 | 3 | 5.9 | 2.1 | 2 |
| 6.3 | 2.9 | 5.6 | 1.8 | 2 |
| 6.5 | 3 | 5.8 | 2.2 | 2 |
| 7.6 | 3 | 6.6 | 2.1 | 2 |
| 4.9 | 2.5 | 4.5 | 1.7 | 2 |
| 7.3 | 2.9 | 6.3 | 1.8 | 2 |
| 6.7 | 2.5 | 5.8 | 1.8 | 2 |
| 7.2 | 3.6 | 6.1 | 2.5 | 2 |
| 6.5 | 3.2 | 5.1 | 2 | 2 |
| 6.4 | 2.7 | 5.3 | 1.9 | 2 |
| 6.8 | 3 | 5.5 | 2.1 | 2 |
| 5.7 | 2.5 | 5 | 2 | 2 |
| 5.8 | 2.8 | 5.1 | 2.4 | 2 |
| 6.4 | 3.2 | 5.3 | 2.3 | 2 |
| 6.5 | 3 | 5.5 | 1.8 | 2 |
| 7.7 | 3.8 | 6.7 | 2.2 | 2 |
| 7.7 | 2.6 | 6.9 | 2.3 | 2 |
| 6 | 2.2 | 5 | 1.5 | 2 |
| 6.9 | 3.2 | 5.7 | 2.3 | 2 |
| 5.6 | 2.8 | 4.9 | 2 | 2 |
| 7.7 | 2.8 | 6.7 | 2 | 2 |
| 6.3 | 2.7 | 4.9 | 1.8 | 2 |
| 6.7 | 3.3 | 5.7 | 2.1 | 2 |
| 7.2 | 3.2 | 6 | 1.8 | 2 |
| 6.2 | 2.8 | 4.8 | 1.8 | 2 |
| 6.1 | 3 | 4.9 | 1.8 | 2 |
| 6.4 | 2.8 | 5.6 | 2.1 | 2 |
| 7.2 | 3 | 5.8 | 1.6 | 2 |
| 7.4 | 2.8 | 6.1 | 1.9 | 2 |

|  |  |  |  |  |
|---|---|---|---|---|
| 7.9 | 3.8 | 6.4 | 2 | 2 |
| 6.4 | 2.8 | 5.6 | 2.2 | 2 |
| 6.3 | 2.8 | 5.1 | 1.5 | 2 |
| 6.1 | 2.6 | 5.6 | 1.4 | 2 |
| 7.7 | 3 | 6.1 | 2.3 | 2 |
| 6.3 | 3.4 | 5.6 | 2.4 | 2 |
| 6.4 | 3.1 | 5.5 | 1.8 | 2 |
| 6 | 3 | 4.8 | 1.8 | 2 |
| 6.9 | 3.1 | 5.4 | 2.1 | 2 |
| 6.7 | 3.1 | 5.6 | 2.4 | 2 |
| 6.9 | 3.1 | 5.1 | 2.3 | 2 |
| 5.8 | 2.7 | 5.1 | 1.9 | 2 |
| 6.8 | 3.2 | 5.9 | 2.3 | 2 |
| 6.7 | 3.3 | 5.7 | 2.5 | 2 |
| 6.7 | 3 | 5.2 | 2.3 | 2 |
| 6.3 | 2.5 | 5 | 1.9 | 2 |
| 6.5 | 3 | 5.2 | 2 | 2 |
| 6.2 | 3.4 | 5.4 | 2.3 | 2 |
| 5.9 | 3 | 5.1 | 1.8 | 2 |
| 5.1 | 3.5 | 1.4 | 0.2 | 2 |

## 2. Source code

```
import numpy as np import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier




from sklearn.model_selection import train_test_split

iris = load_iris()

print(iris.DESCR)


print(iris['feature_names'])

print(iris['target'])

X = iris.data
y= iris.target
```

```
X.shape
```

```
plt.plot(X[:, 0][y == 0] * X[:, 1][y == 0], X[:, 2][y == 0] * X[:, 3][y == 0],
'r.', label="Satosa")
plt.plot(X[:, 0][y == 1] * X[:, 1][y == 1], X[:, 2][y == 1] * X[:, 3][y == 1],
'g.', label="Virginica")
plt.plot(X[:, 0][y == 2] * X[:, 1][y == 2], X[:, 2][y == 2] * X[:, 3][y == 2],
'b.', label="Versicolour") plt.legend()
plt.show()


X_train, X_test, y_train, y_test = train_test_split(iris['data'], iris['target'],
random_state=0)
knn = KNeighborsClassifier(n_neighbors=1)

knn.fit(X_train, y_train)

# Imagine that we obtained a new iris

X_new = np.array([[5.7 ,2.5 ,5. , 2.]])
print(X_new.shape)

prediction = knn.predict(X_new)
print(prediction)

print(iris['target_names'][prediction])

print(knn.score(X_test, y_test))
```

# FAKE NEWS DDETECTION USING MACHINE LEARNING

## INTRODUCTION

As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to hunt out

and consume news from social media instead of traditional news organizations.[1] The explanations for this alteration in consumption

behaviors are inherent within the nature of those social media platforms: (i) it's often more timely and fewer expensive to consume news on social media compared with traditional journalism , like newspapers ortelevision; and (ii) it's easier to further share, discuss , and

discuss the news with friends or other readers on social media. For instance, 62 percent of U.S. adults get news on social media in 2016,

while in 2012; only 49 percent reported seeing news on social media [1]. It had been also found that social media now outperforms

television because the major news source. Despite the benefits provided by social media, the standard of stories on social media is less

than traditional news organizations. However, because it's inexpensive to supply news online and far faster and easier to propagate

through social media, large volumes of faux news, i.e., those news articles with intentionally false information, are produced online for a

spread of purposes, like financial and political gain. it had been estimated that over 1 million tweets are associated with fake news

“Pizzagate" by the top of the presidential election. Given the prevalence of this new phenomenon, “Fake news" was even named the word

of the year by the Macquarie dictionary in 2016 [2]. The extensive spread of faux

news can have a significant negative impact on

individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance; it's evident that the

most popular fake news was even more outspread on Facebook than the most accepted genuine mainstream news during the U.S. 2016 presidential election. Second, fake news intentionally persuades consumers to simply accept biased or false beliefs. Fake news is typically manipulated by propagandists to convey political messages or influence for instance, some report shows that Russia has created fake accounts and social bots to spread false stories. Third, fake news changes the way people interpret and answer real news, for instance, some fake news was just created to trigger people's distrust and make them confused; impeding their abilities to differentiate what's true from what's not. To assist mitigate the negative effects caused by fake news (both to profit the general public and therefore the news ecosystem). It's crucial that we build up methods to automatically detect fake news broadcast on social media [3].

Internet and social media have made the access to the news information much easier and comfortable [2]. Often Internet users can pursue the events of their concern in online form, and increased number of the mobile devices makes this process even easier. But with great possibilities come great challenges. Mass media have an enormous influence on the society, and because it often happens, there's someone who wants to require advantage of this fact. Sometimes to realize some goals mass-media may manipulate the knowledge in several ways. This result in producing of the news articles that isn’t completely true or maybe completely false. There even exist many websites that produce fake news almost exclusively. They intentionally publish hoaxes, half-truths, propaganda and disinformation asserting to be real news – often using social media to drive web traffic and magnify their effect. The most goals of faux news websites are to affect the general public opinion on certain matters (mostly political). Samples of such websites could also be found in Ukraine, United States of America, Germany, China and much of other countries [4]. Thus, fake news may be a global issue also as a worldwide challenge. Many scientists believe that fake news issue could also be addressed by means of machine learning and AI [5]. There’s a reason for that: recently AI algorithms have begun to work far better on manyclassification problems (image recognition, voice detection

then on) because hardware is cheaper and larger datasets are available. There are several influential articles about automatic deception
detection. In [6] the authors provide a general overview of the available techniques for the matter. In [7] the authors describe their method

for fake news detection supported the feedback for the precise news within the micro blogs. In [8] the authors actually develop two systems for deception detection supported support vector machines and Naive Bayes classifier (this method is employed within the

system described during this paper as well) respectively. They collect the info by means of asking people to directly provide true or false

information on several topics – abortion, execution and friendship. The accuracy of the detection achieved by the system is around 70%.

This text describes an easy fake news detection method supported one among the synthetic intelligence algorithms – naïve Bayes

classifier, Random Forest and Logistic Regression. The goal of the research is to look at how these particular methods work for this

particular problem given a manually labelled news dataset and to support (or not) the thought of using AI for fake news detection. The

difference between these article and articles on the similar topics is that during this paper Logistic Regression was specifically used for

fake news detection; also, the developed system was tested on a comparatively new data set, which gave a chance to gauge its performance on a recent data.

Characteristics of Fake News:

They often have grammatical mistakes. They are often emotionally colored. They often try to affect readers' opinion on some topics.

Their content is not always true. They often use attention seeking words andnews format and click baits. They are too good to be true.

Their sources are not genuine most of the times [9].

## LITERATURE REVIEW

Mykhailo Granik et. al. in their paper [3] shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

Himank Gupta et. al. [10] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of-Words model. 4.

They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%.

Marco L. Della Vedova et. al. [11] first proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing its accuracy up to 78.8%. Second, they implemented their method within a Facebook Messenger Chabot and validate it with a real-worldapplication, obtaining a fake news detection accuracy of 81.7%. Their goal was to classify a news item as reliable or fake; they first described the datasets they used for their test, then presented the content-based approach they implemented and the method they proposed tocombine it with a social-based approach available in the literature. The resulting dataset is composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2, 300, 00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes and 6,577 (42.4%) are non-hoaxes.

Cody Buntain et. al. [12] develops a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowd sourced dataset of accuracy assessments for events in Twitter, and

PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. They apply this method to Twitter

content sourced from BuzzFeed’s fake news dataset. A feature analysis identifies features that are most predictive for crowd sourced and

journalistic accuracy assessments, results of which are consistent with prior work. They rely on identifying highly retweeted threads of

conversation and use the features of these threads to classify stories, limiting this work’s applicability only to the set of popular tweets.

Since the majority of tweets are rarely retweeted, this method therefore is only usable on a minority of Twitter conversation threads.

In his paper, Shivam B. Parikh et. al. [13] aims to present an insight of characterization of news story in the modern diaspora combined

with the differential content types of news story and its impact on readers. Subsequently, we dive into existing fake news detection

approaches that are heavily based on text-based analysis, and also describe popular fake news datasets. We conclude the paper by

identifying 4 key open research challenges that can guide future research. It is a theoretical Approach which gives Illustrations of fake news

detection by analyzing the psychological factors.

## METHODOLOGY

This paper explains the system which is developed in three parts. The first part is static which works on machine learning classifier. We

studied and trained the model with 4 different classifiers and chose the best classifier for final execution. The second part is dynamic which takes the keyword/text from user and searches online for the truth probability of the news. The third part provides the authenticity of the URL input by user.

In this paper, we have used Python and its Sci-kit libraries [14]. Python has a huge set of libraries and extensions, which can be easily

used in Machine Learning. Sci-Kit Learn library is the best source for machine learning algorithms where nearly all types of machine

learning algorithms are readily available for Python, thus easy and quick evaluation of ML algorithms is possible. We have used Django

for the web based deployment of the model, provides client side implementation using HTML, CSS and Javascript. We have also used

Beautiful Soup (bs4), requests for online scrapping.

# IMPLEMENTATION

## DATA COLLECTION AND ANALYSIS

We can get online news from different sources like social media websites, search engine, homepage of news agency websites or the fact_checking websites. On the Internet, there are a few publicly available datasets for Fake news classification like Buzzfeed News, LIAR

[15], BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news. In the

following sections, I have discussed in brief about the sources of the dataset used in this work.

Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites.

However, manually determining the veracity of news is a challenging task, usually requiring annotators with domain expertise who

performs careful analysis of claims and additional evidence, context, and reports from authoritative sources. Generally, news data with

annotations can be gathered in the following ways: Expert journalists, Fact-checking websites, Industry detectors, and Crowd sourced

workers. However, there are no agreed upon benchmark datasets for the fake news detection problem. Data gathered must be pre_processed- that is, cleaned, transformed and integrated before it can undergo training process [16]. The dataset that we used is explained

below:
LIAR: This dataset is collected from fact-checking website PolitiFact through its API [15]. It includes 12,836 human labelled short

statements, which are sampled from various contexts, such as news releases, TV or radio interviews, campaign speeches, etc. The labels

for news truthfulness are fine-grained multiple classes: pants-fire, false, barelytrue, half-true, mostly true, and true.

The data source used for this project is LIAR dataset which contains 3 files with .csv format for test, train and validation. Below is some

description about the data files used for this project.

LIAR: A Benchmark Dataset for Fake News Detection

William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the

55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.

Below are the columns used to create 3 datasets that have been in used in this project-

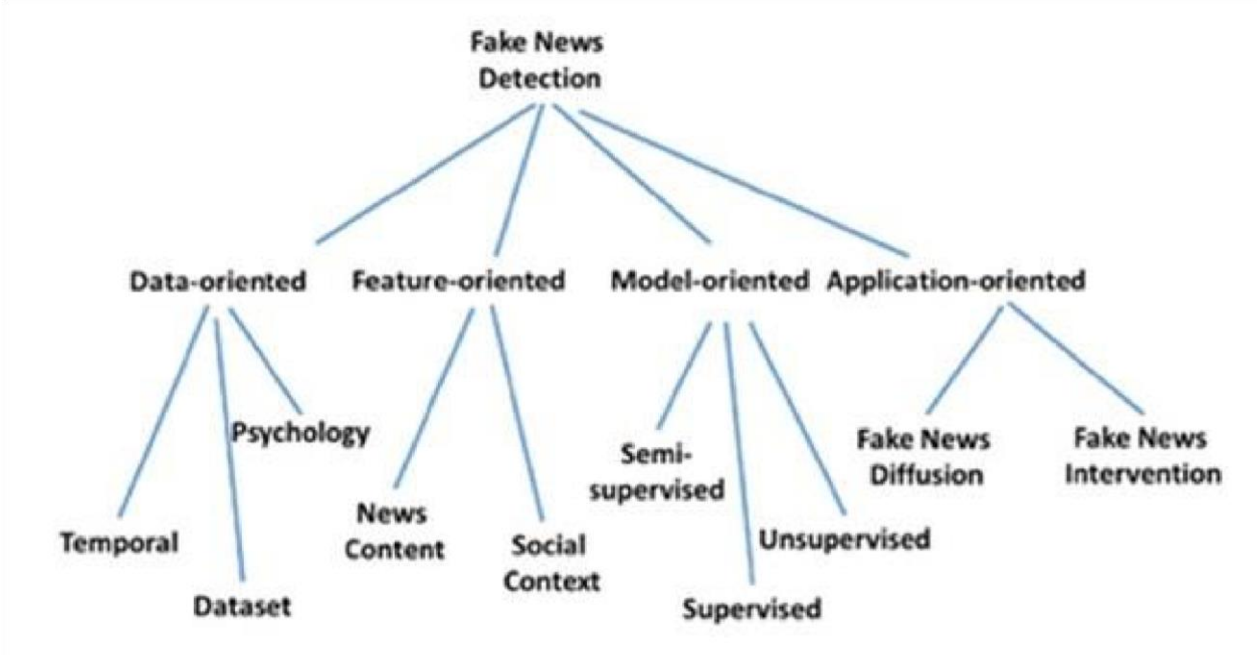
Column1: Statement (News headline or text)

Column2: Label (Label class contains: True, False)

The dataset used for this project were in csv format named train.csv, test.csv and valid.csv.

REAL_OR_FAKE.CSV we used this dataset for passive aggressive classifier. It contains 3 columns viz 1- Text/keyword, 2-Statement,

3-Label (Fake/True)



**DEFINITIONS AND DETAILS**

Pre-processing Data Social media data is highly unstructured – majority of them are informal communication with typos, slangs and bad-grammar etc. [17].

Quest for increased performance and reliability has made it imperative to develop techniques for utilization of resources to make

informed decisions [18]. To achieve better insights, it is necessary to clean the data before it can be used for predictive modelling. For this

purpose, basic pre-processing was done on the News training data. This step was comprised of_Data Cleaning:

While reading data, we get data in the structured or unstructured format. A structured format has a well-defined pattern whereas

unstructured data has no proper structure. In between the 2 structures, we have a semi-structured format which is a comparably better structured than unstructured format.

Cleaning up the text data is necessary to highlight attributes that we're going to want our machine learning system to pick up on. Cleaning

(or pre-processing) the data typically consists of a number of steps:

Remove punctuation

Punctuation can provide grammatical context to a sentence which supports our understanding. But for our vectorizer which counts the number of words and not the context, it does not add value, so we remove all special characters. eg: How are you?->How are you

### Tokenization

Tokenizing separates text into units such as sentences or words. It gives structure to previously unstructured text. eg: Plata o Plomo->

‘Plata’,’o’,’Plomo’.

### Remove stopwords

Stopwords are common words that will likely appear in any text. They don’t tell us much about our data so we remove them. eg: silver or

lead is fine for me-> silver, lead, fine.

### Stemming

Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It removes suffices, like

“ing”, “ly”, “s”, etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected. eg:

Entitling, Entitled -> Entitle. Note: Some search engines treat words with the same stem as synonyms [18].

### Feature Generation

We can use text data to generate a number of features like word count, frequency of large words, frequency of unique words, n-grams etc.

By creating a representation of words that capture their meanings, semantic relationships, and numerous types of context they are used in,

we can enable computer to understand text and perform Clustering, Classification etc [19]. 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.

### Vectorizing Data: Bag-Of-Words

Bag of Words (BoW) or CountVectorizer describes the presence of words within the text data. It gives a result of 1 if present in the

sentence and 0 if not present. It, therefore, creates a bag of words with a document-matrix count in each text document.

### Vectorizing Data: N-Grams

N-grams are simply all combinations of adjacent words or letters of length n that we can find in our source text. Ngrams with n=1 are called unigrams. Similarly, bigrams (n=2), trigrams (n=3) and so on can also be used. Unigrams usually don’t contain much information

as compared to bigrams and trigrams. The basic principle behind n-grams is that they capture the letter or word is likely to follow the

given word. The longer the n-gram (higher n), the more context you have to work with [20].

## Vectorizing Data: TF-IDF

It computes “relative frequency” that a word appears in a document compared to its frequency across all documents TF-IDF weight represents the relative importance of a term in the document and entire corpus [17].

TF stands for Term Frequency: It calculates how frequently a term appears in a document. Since, every document size varies, a term may

appear more in a long sized document that a short one. Thus, the length of the document often divides Term frequency.

Note: Used for search engine scoring, text summarization, document clustering.

$TF(t,d\ )$ =

$Number\ of\ times\ t\ occurs\ in\ document\ d''$

$Total\ word\ count\ of\ document\ d''$

IDF stands for Inverse Document Frequency: A word is not of much use if it is present in all the documents. Certain terms like “a”, “an”,

“the”, “on”, “of” etc. appear many times in a document but are of little importance. IDF weighs down the importance of these terms and

increase the importance of rare ones. The more the value of IDF, the more unique is the word [17].

$IDF(t, d)$ =

$Total\ number\ of\ documents$

$Number\ of\ documents\ with\ term\ t\ in\ it\ )$

TF-IDF is applied on the body text, so the relative count of each word in the sentences is stored in the document matrix.

$TFIDF(t, d) = TF(t, d) * IDF(t)$

Note: Vectorizers outputs sparse matrices. Sparse Matrix is a matrix in which most entries are 0 [21].

## Algorithms used for Classification

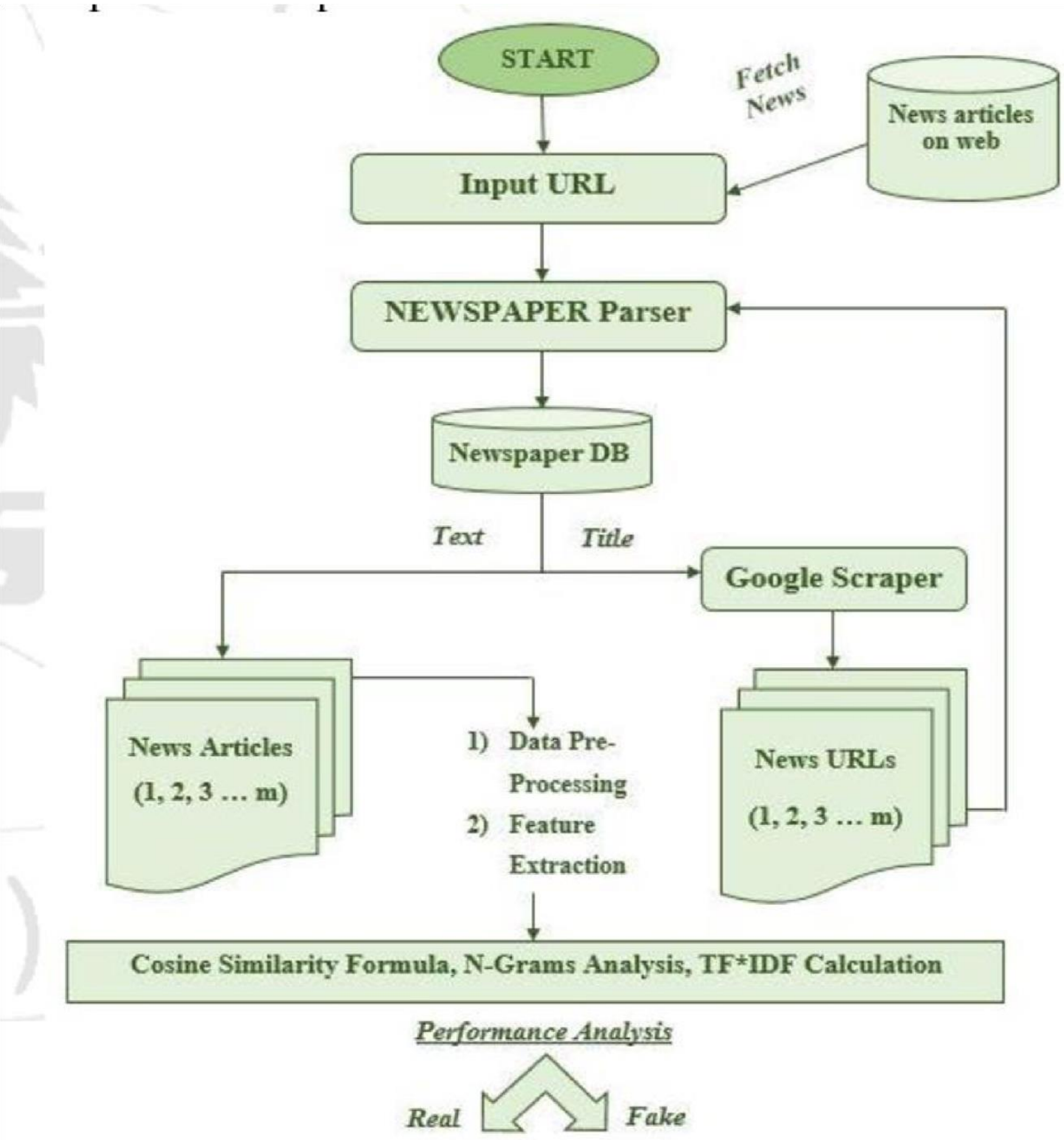


**Figure 1:** Implementation of the proposed model.

This section deals with training the classifier. Different classifiers were investigated to predict the class of the text. We explored

specifically four different machine-learning algorithms – Multinomial Naïve Bayes Passive Aggressive Classifier and Logistic regression.

The implementations of these classifiers were done using Python library SciKit Learn. Brief introduction to the algorithms_1. Naïve Bayes Classifier:

This classification technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. It provides way for calculating the posterior probability. $P(x) =$

$P$

$P$

$(c) * P(c)$

$(x)$

P(c|x)= posterior probability of class given predictor P(c)= prior probability of class

P(x|c)= likelihood (probability of predictor given class)

P(x) = prior probability of predictor Random Forest:

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we've

collection of decision trees (so known as

"Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The

forest chooses the classification having the most votes (over all the trees in the forest). The random forest is a classification algorithm

consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an

uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Random forest, like its name

implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits

out a class prediction and the class with the most votes becomes our model's prediction. The reason that the random forest model works

so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual

constituent models. So how does random forest ensure that the behaviour of each individual tree is not too correlated with the behaviour

of any of the other trees in the model? It uses the following two methods:

Bagging (Bootstrap Aggregation) — Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly

sample from the dataset with replacement, resulting in different trees. This process is known as bagging or bootstrapping.

Feature Randomness — In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random

forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately

results in lower correlation across trees and more diversification [22].

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 variable(s). 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 [23].

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

Passive Aggressive Classifier:

The Passive Aggressive Algorithm is an online algorithm; ideal for classifying massive streams of data (e.g. twitter). It is easy to

implement and very fast. It works by taking an example, learning from it and then throwing it away [24]. Such an algorithm remains

passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most

other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the

weight vector [25].

**IMPLEMENTATION STEPS**

Static Search Implementation_In static part, we have trained and used 3 out of 4 algorithms for classification. They are Naïve Bayes, Random Forest and Logistic Regression.

Step 1: In first step, we have extracted features from the already pre-processed dataset. These features are; Bag-of-words, Tf-Idf Features and N-grams.

Step 2: Here, we have built all the classifiers for predicting the fake news detection. The extracted features are fed into different

classifiers. We have used Naive-bayes, Logistic Regression, and Random forest classifiers from sklearn. Each of the extracted features was used in all of the classifiers.

Step 3: Once fitting the model, we compared the f1 score and checked the confusion matrix.

Step 4: After fitting all the classifiers, 2 best performing models were selected as candidate models for fake news classification.

Step 5: We have performed parameter tuning by implementing GridSearchCV methods on these candidate models and chosen best performing parameters for these classifier.

Step 6: Finally selected model was used for fake news detection with the probability of truth.

Step 7: Our finally selected and best performing classifier was Logistic Regression which was then saved on disk. It will be used to

classify the fake news. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of

truth.

Dynamic Search Implementation_Our dynamic implementation contains 3 search fields which are_1) Search by article content.

Search using key terms.

Search for website in database.

In the first search field we have used Natural Language

Processing for the first search field to come up with a proper solution for the problem, and hence we have attempted to create a model

which can classify fake news according to the terms used in the newspaper articles. Our application uses NLP techniques like

CountVectorization and TF-IDF Vectorization before passing it through a Passive Aggressive Classifier to output the authenticity as a percentage probability of an article.

The second search field of the site asks for specific keywords to be searched on the net upon which it provides a suitable output for the

percentage probability of that term actually being present in an article or a similar article with those keyword references in it.

The third search field of the site accepts a specific website domain name upon which the implementation looks for the site in our true sites

database or the blacklisted sites database. The true sites database holds the domain names which regularly provide proper and authentic

news and vice versa. If the site isn't found in either of the databases then the implementation doesn't classify the domain it simply states that the news aggregator does not exist.

Working_The problem can be broken down into 3 statements_1) Use NLP to check the authenticity of a news article.

If the user has a query about the authenticity of a search query then we he/she can directly search on our platform and using our custom algorithm we output a confidence score.

Check the authenticity of a news source.

These sections have been produced as search fields to take inputs in 3 different forms in our implementation of the problem statement.

## EVALUATION MATRICES

Evaluate the performance of algorithms for fake news detection problem; various evaluation metrics have been used. In this subsection,

we review the most widely used metrics for fake news detection. Most existing approaches consider the fake news problem as a classification problem that predicts whether a news article is fake or not:

True Positive (TP): when predicted fake news pieces are actually classified as fake news; True Negative (TN): when predicted true news pieces are actually classified as true news; False Negative (FN): when predicted true news pieces are actually classified as fake news; False Positive (FP): when predicted fake news pieces are actually classified as true news. Confusion 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

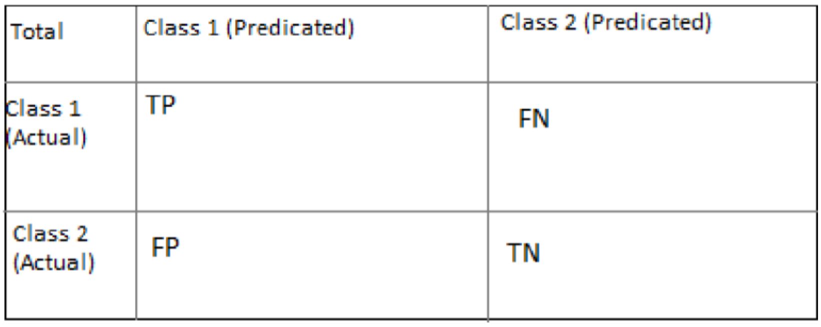
for which the true values are known. It allows the visualization of the performance of an algorithm. A confusion matrix is a summary of

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 [26].

Table 1 : Confusion Matrix



| Total | Class 1 (Predicated) | Class 2 (Predicated) |
|---|---|---|
| Class 1 (Actual) | TP | FN |
| Class 2 (Actual) | FP | TN |

By formulating this as a classification problem, we can define following metrics_1. Precision = $|T\ P|$

$T|\ P|+|F\ P|$

2. Recall = $|T\ P|$

$T|\ P|+|F\ N|$

3. F1 Score = 2 * $Precisionn*\ Recall$

$Precision\ +Recall$

4. Accuracy =

$T|\ P|+|T\ N|$

$T|\ P|+|TN|+|F\ P|+|F\ N|$

These metrics are commonly used in the machine learning community and enable us to evaluate the performance of a classifier from different perspectives. Specifically, accuracy measures the similarity between predicted fake news and real fake news.

## RESULT

Implementation was done using the above algorithms with Vector features- Count Vectors and Tf-Idf vectors at Word level and Ngram_level. Accuracy was noted for all models. We used K-fold cross validation technique to improve the effectiveness of the models.

### Dataset split using K-fold cross validation

This cross-validation technique was used for splitting the dataset randomly into k-folds. (k-1) folds were used for building the model

while kth fold was used to check the effectiveness of the model. This was repeated until each of the k-folds served as the test set. I used 3-

fold cross validation for this experiment where 67% of the data is used for training the model and remaining 33% for testing.

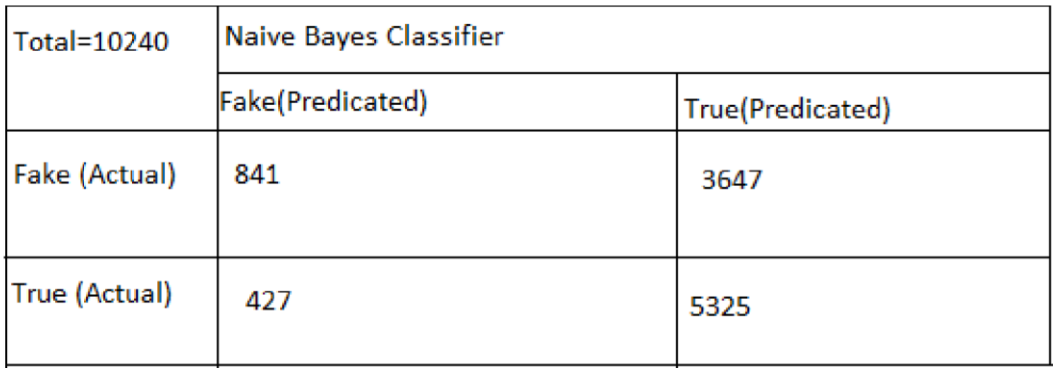
### Confusion Matrices for Static System

After applying various extracted features (Bag-of-words, Tf-Idf. N-grams) on threedifferent classifiers (Naïve bayes, Logistic Regression

and Random Forest), their confusion matrix showing actual set and predicted sets are mentioned below:

Table 2: Confusion Matrix for Navies Bayes classifier

using tf-idf features-



| Total=10240 | Naive Bayes Classifier | |
|---|---|---|
| | Fake(Predicated) | True(Predicated) |
| Fake (Actual) | 841 | 3647 |
| True (Actual) | 427 | 5325 |

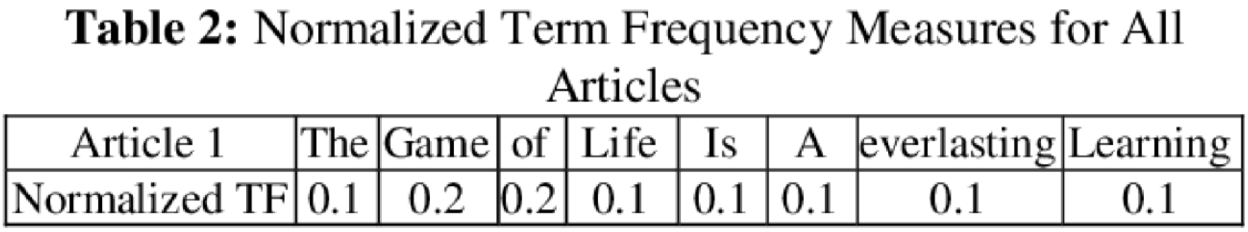
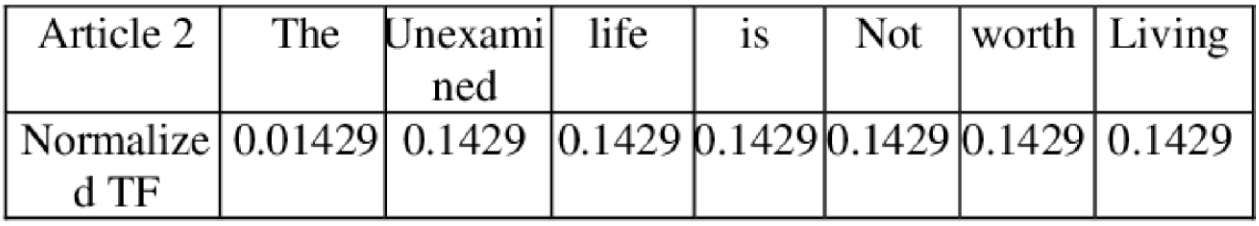
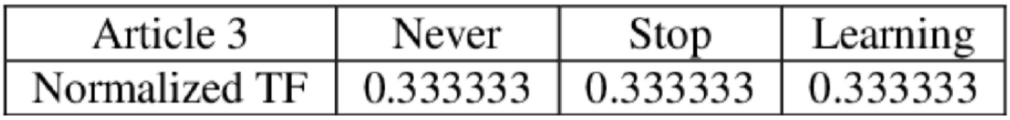


**Table 2:** Normalized Term Frequency Measures for All Articles

| Article 1 | The | Game | of | Life | Is | A | everlasting | Learning |
|---|---|---|---|---|---|---|---|---|
| Normalized TF | 0.1 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |



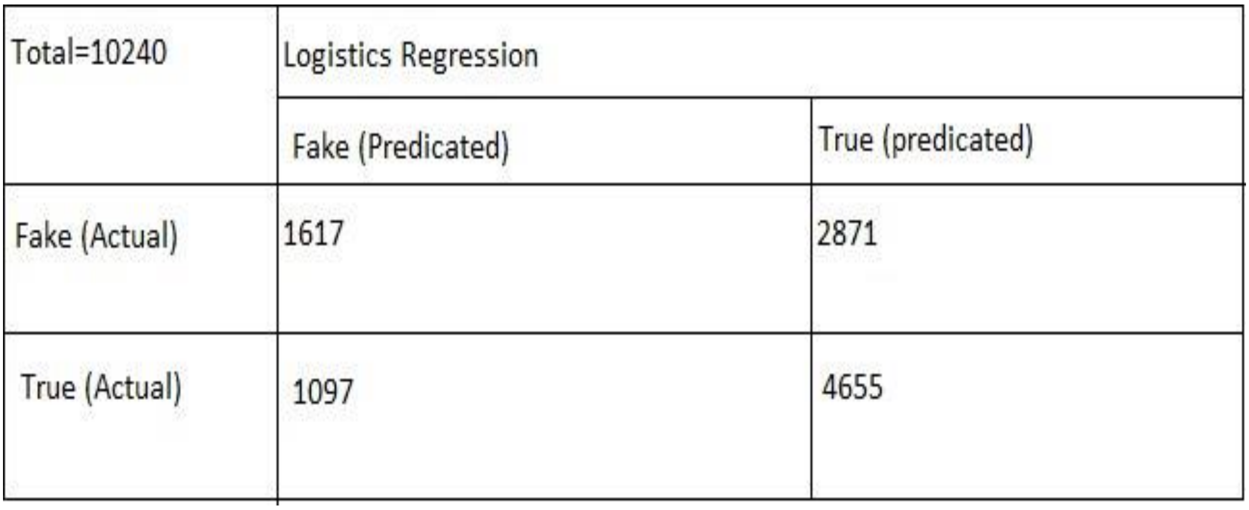
| Article 2 | The | Unexamined | life | is | Not | worth | Living |
|---|---|---|---|---|---|---|---|
| Normalized TF | 0.01429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 |



| Article 3 | Never | Stop | Learning |
|---|---|---|---|
| Normalized TF | 0.333333 | 0.333333 | 0.333333 |

Table 3: Confusion Matrix for Logistics Regresssion using

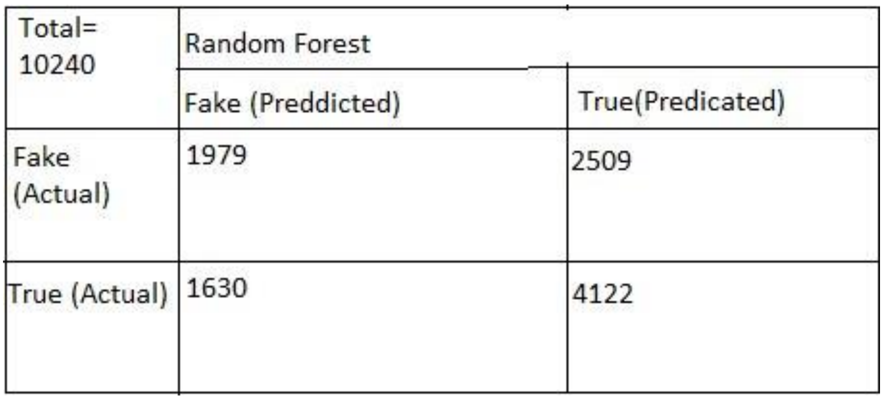
Tf-Idf feautures-



| Total=10240 | Logistics Regression | |
|---|---|---|
| | Fake (Predicated) | True (predicated) |
| Fake (Actual) | 1617 | 2871 |
| True (Actual) | 1097 | 4655 |

Table 4: Confusion Matrix for Random Forest Classifier using

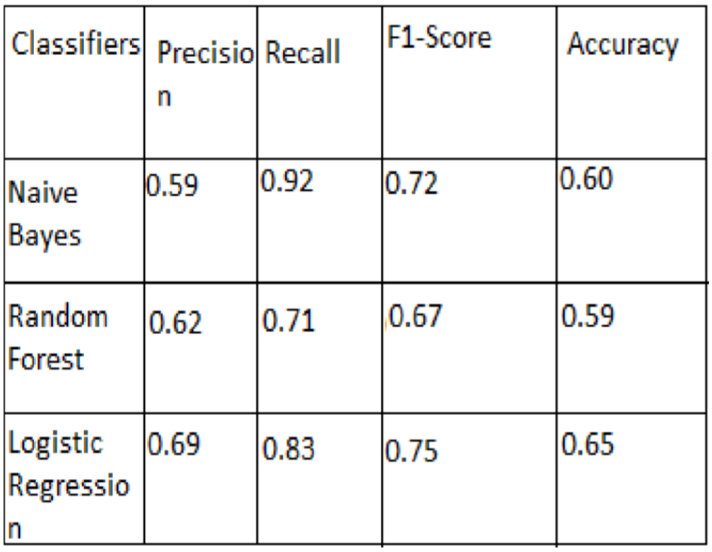
Tf-Idf features



| Total= 10240 | Random Forest | |
|---|---|---|
| | Fake (Preddicted) | True(Predicated) |
| Fake (Actual) | 1979 | 2509 |
| True (Actual) | 1630 | 4122 |

Table 5:Comparsion of Precision,Recall,F1-scores and

Accuracy for all three classifiers-



| Classifiers | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Naive Bayes | 0.59 | 0.92 | 0.72 | 0.60 |
| Random Forest | 0.62 | 0.71 | 0.67 | 0.59 |
| Logistic Regression | 0.69 | 0.83 | 0.75 | 0.65 |

As evident above our best model came out to be Logistic Regression with an accuracy of 65%. Hence we then used grid search parameter

optimization to increase the performance of logistic regression which then gave us the accuracy of 80%.

Hence we can say that if a user feed a particular news article or its headline in our model, there are 80% chances that it will be classified

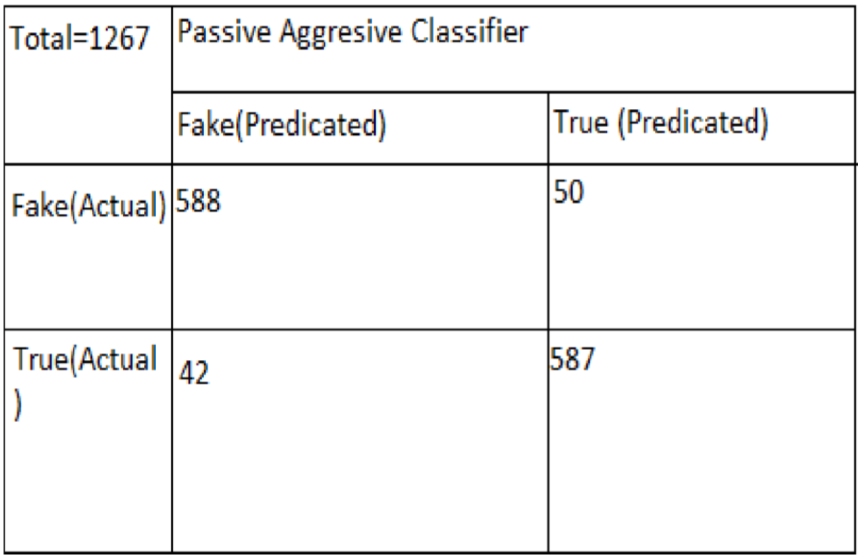
to its true nature.

Confusion Matrix for Dynamic System

We used real_or_fake.csv with passive aggressive classifier and obtained the following confusion matrix-

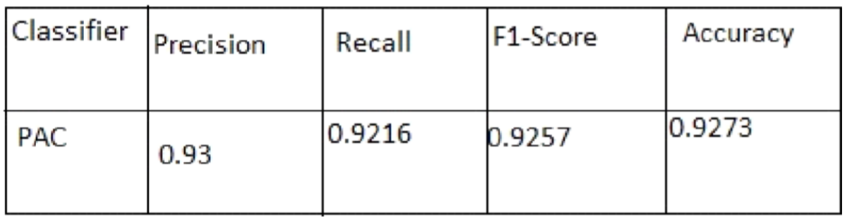
Table 6: Confusion Matrix for passive aggressive

Classifier-



| Total=1267 | Passive Aggresive Classifier | |
|---|---|---|
| | Fake(Predicated) | True (Predicated) |
| Fake(Actual) | 588 | 50 |
| True(Actual<br>) | 42 | 587 |

Table 7: Performance measures



| Classifier | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| PAC | 0.93 | 0.9216 | 0.9257 | 0.9273 |

**CONCULUSION**

In the 21st century, the majority of the tasks are done online. Newspapers that were earlier preferred as hard-copies 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. As evident

above for static search, our best model came out to be Logistic Regression with an accuracy of 65%. Hence we then used grid search

parameter optimization to increase the performance of logistic regression whichthen gave us the accuracy of 75%. Hence we can say that

if a user feed a particular news article or its headline in our model, there are 75% chances that it will be classified to its true nature.

The user can check the news article or keywords online; he can also check the authenticity of the website. The accuracy for dynamic system is 93% and it increases with every iteration.

We intend 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.

## REFERANCE

**Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake NewsDetection on Social Media: A Data Mining Perspective” arXiv:1708.01967v3 [cs.SI], 3 Sep 2017**

**Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake News Detection on Social Media: A Data Mining Perspective” arXiv:1708.01967v3 [cs.SI], 3 Sep 2017**

**M. Granik and V. Mesyura, "Fake news detection using naive Bayesclassifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.**

**Fake news websites. (n.d.) Wikipedia. [Online]. Available: [https://en.wikipedia.org/wiki/Fake_news_website](https://en.wikipedia.org/wiki/Fake_news_website). Accessed Feb. 6, 2017**

**Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news.**

**Conroy, N., Rubin, V. and Chen, Y. (2015). "Automatic deception detection: Methods for finding fake news" at Proceedings of the**

**Association for Information Science and Technology, 52(1), pp.1-4.**

**Markines, B., Cattuto, C., & Menczer, F. (2009, April). "Social spam detection". In Proceedings of the 5th International Workshop on**

**Adversarial Information Retrieval on the Web (pp. 41-48)**

**Rada Mihalcea , Carlo Strapparava, The lie detector: explorations in theautomatic recognition of deceptive language, Proceedings of the ACL-IJCNLP**

**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 Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6,**

**March 2019**

**H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th**