SPAM MAIL DETECTION USING MACHINE LEARNING

- K SUPRIYA

INTRODUCTION

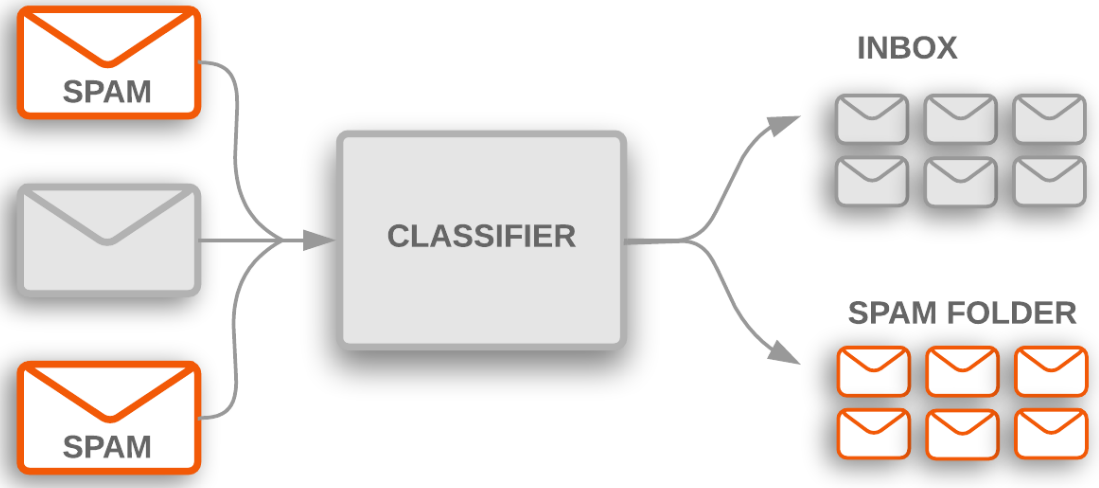
Spam email is unwanted junk email sent out in bulk to a random recipient list. There are different ways that spam can be sent. It could be sent by humans but the most common way is sending them through a network of computers called botnets (spambots). To solve this problem of spotting or detecting a Spam Email, there are various Machine Learning techniques in both Supervised and Unsupervised learning.

WHAT WE AIM TO DO?

Through this project we aim to detect or spot spam emails and classify them using different Machine Learning algorithms. We would also compare the algorithms and check the best algorithm which has a better precision and accuracy value.

For this particular project we would be focusing on using three Supervised Machine Learning techniques:

- KNN or K-nearest neighbors
- SVM or Support vector machine
- Decision Trees



LITERATURE SURVEY

1. ***Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges.***

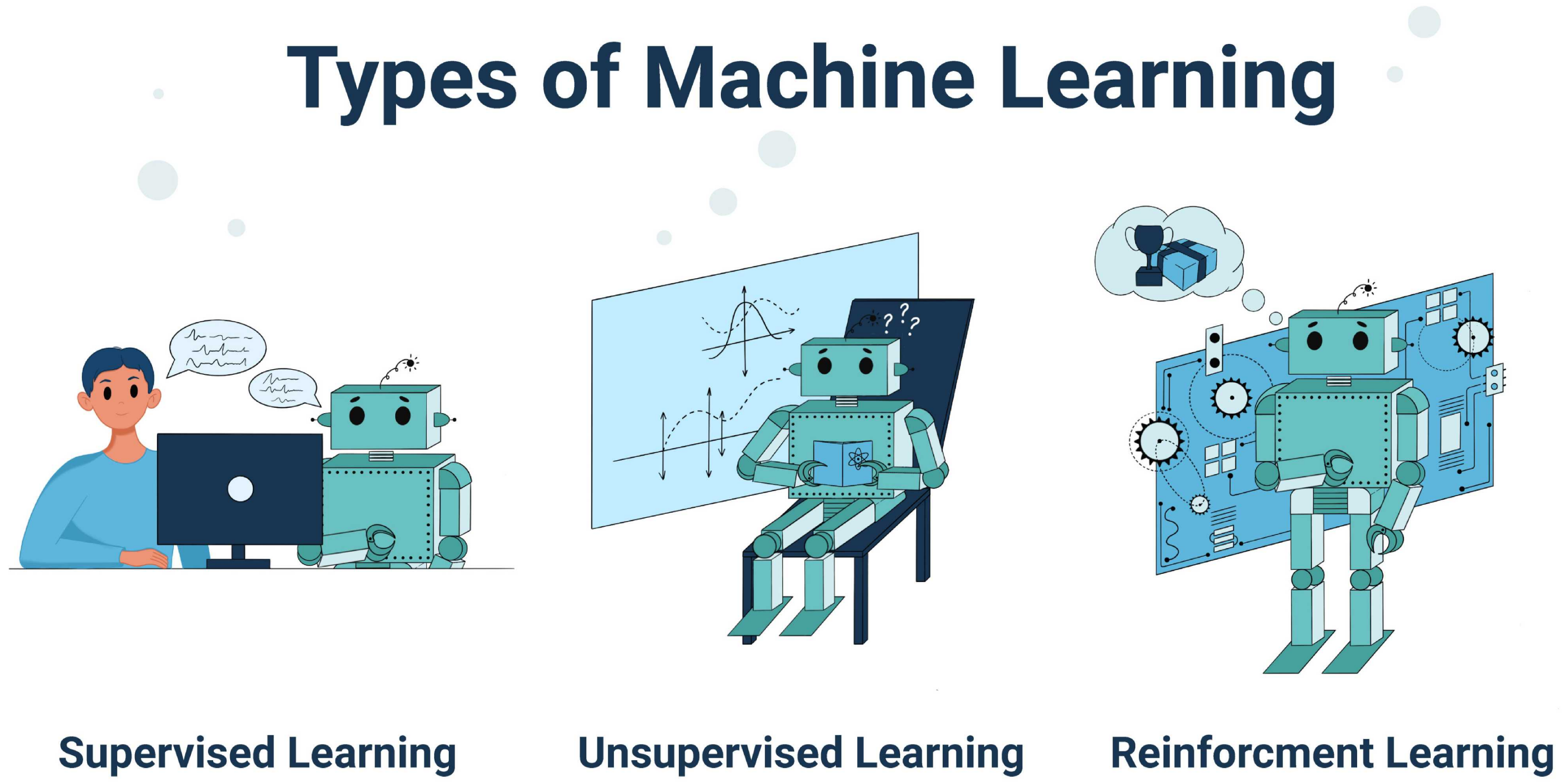
- A detailed comparison including different parameters like accuracy, precision, recall have been discussed in depth. Both Supervised and Unsupervised Machine Learning techniques are used and they are compared based on their accuracy, precision value and other important parameters.
- It also provides a comprehension or understanding of future spam detection or filtering methods that are open to research and provide better security email platforms. Finally the paper gives us research gaps, challenges of spam detection and also future space and area of research.

LITERATURE SURVEY

2. A Comparative Analysis of SMS Spam Detection employing Machine Learning Methods.

- The aim of this study is to detect spam message that are a threat
- In this paper, studies on SMS spam problems to perform a better accuracy using several different techniques such as Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Random Forest, Logistic Regression and some more are performed.
- The result indicated that Support Vector Machine achieved the highest accuracy of 99%, indicating it might be useful as an effective machine learning system for future research.

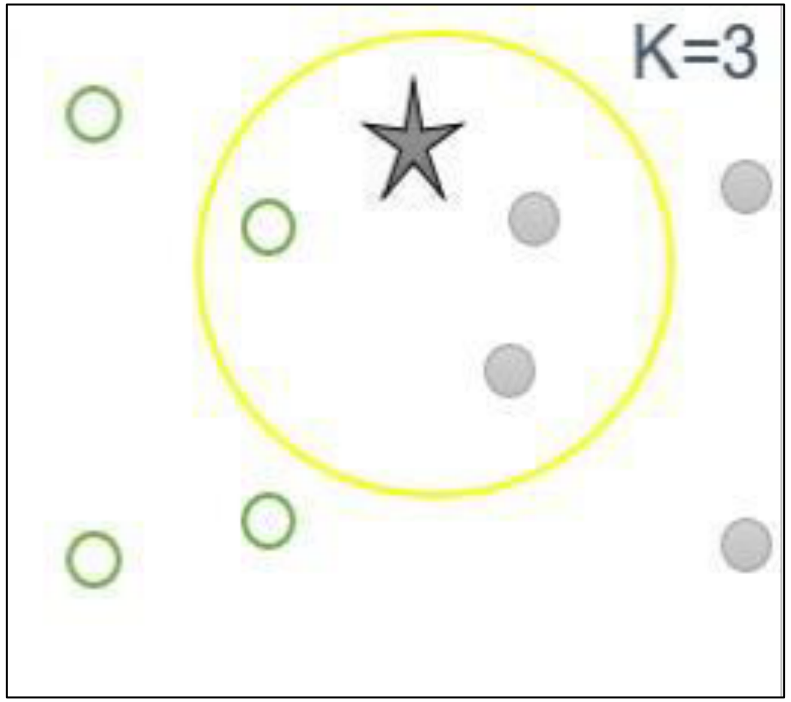
TYPES OF MACHINE LEARNING



MODELS USED

K-NEAREST NEIGHBORS:

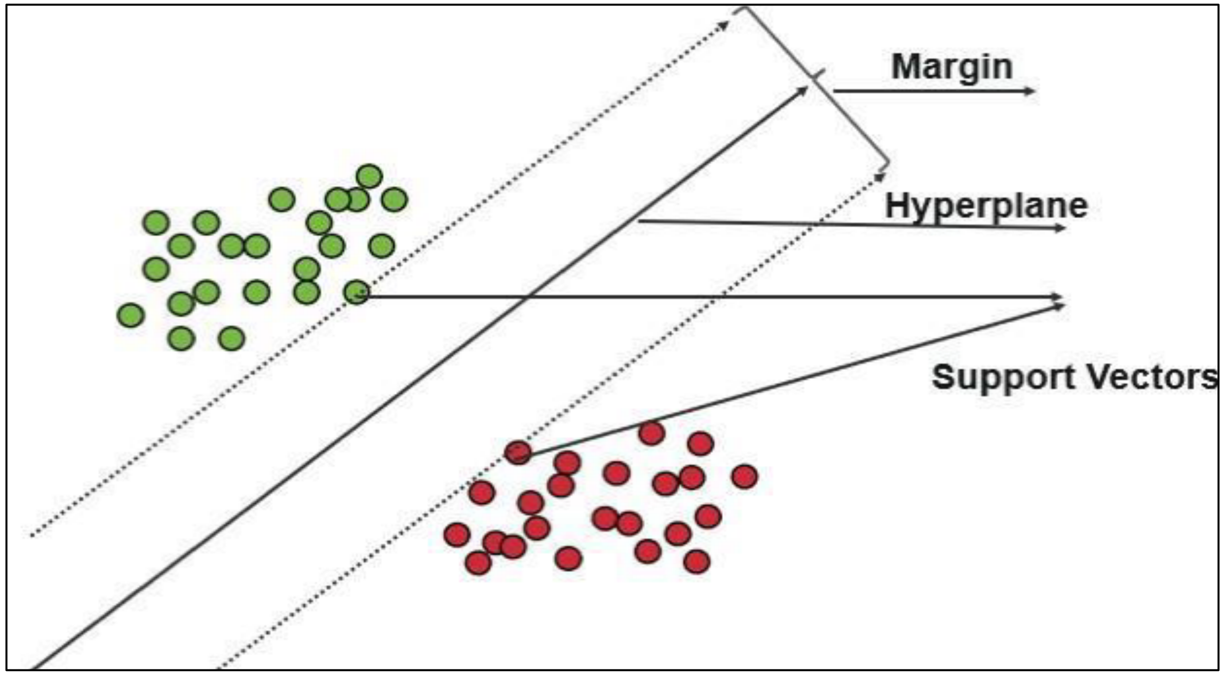
KNN (K-Nearest Neighbors) is one of the simplest supervised learning algorithms. “K” stands for number of data set items that are considered for the classification.



MODELS USED

SUPPORT VECTOR MACHINE:

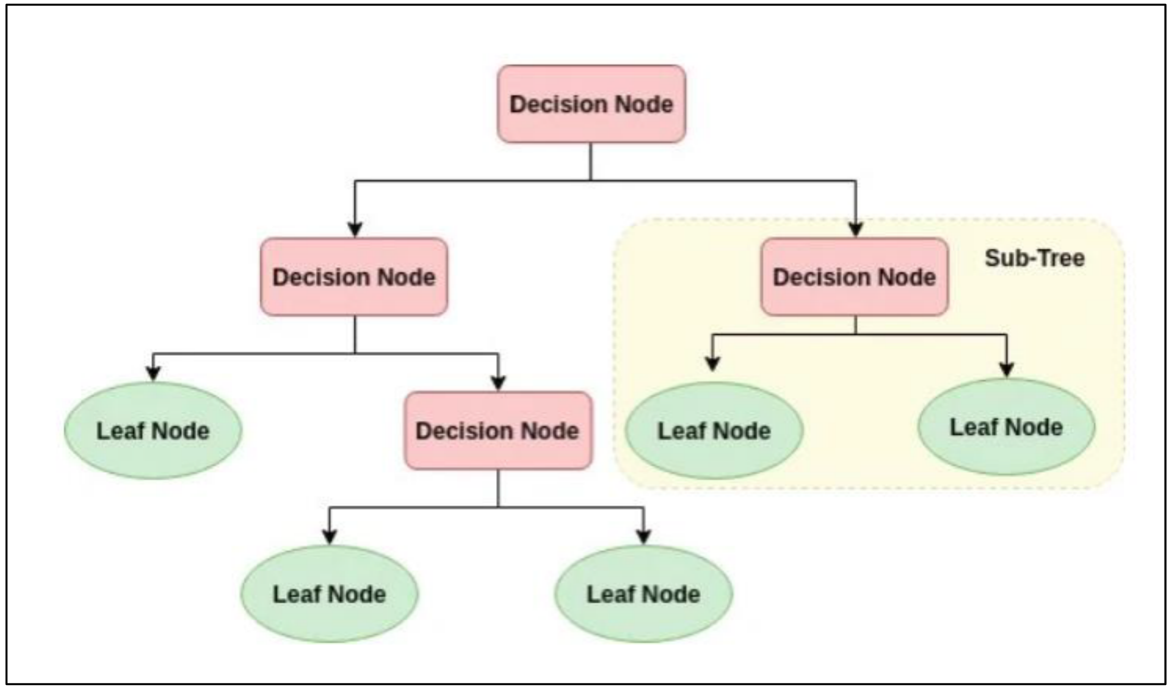
It is a supervised machine learning algorithm that works by finding a hyper plane that classifies the dataset into different classes.



MODELS USED

DECISION TREE:

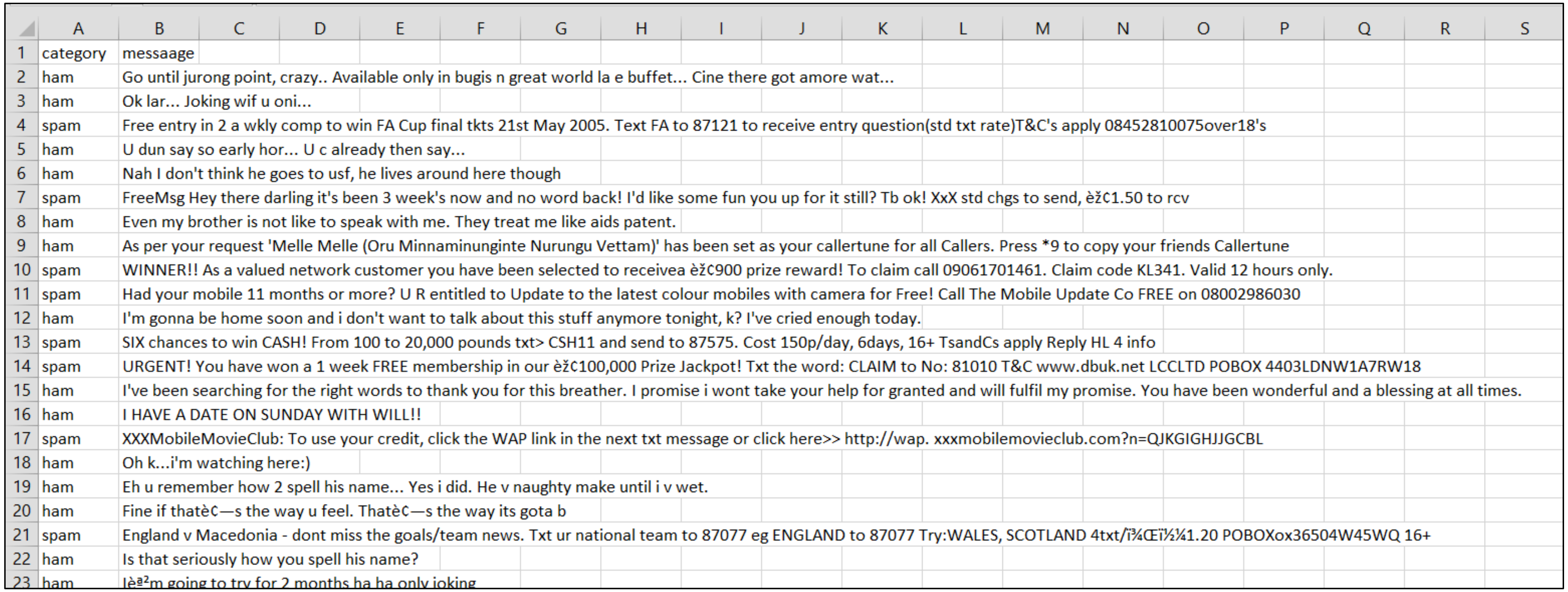
A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered.



IMPLEMENTATION

DATASET AND SOFTWARE:

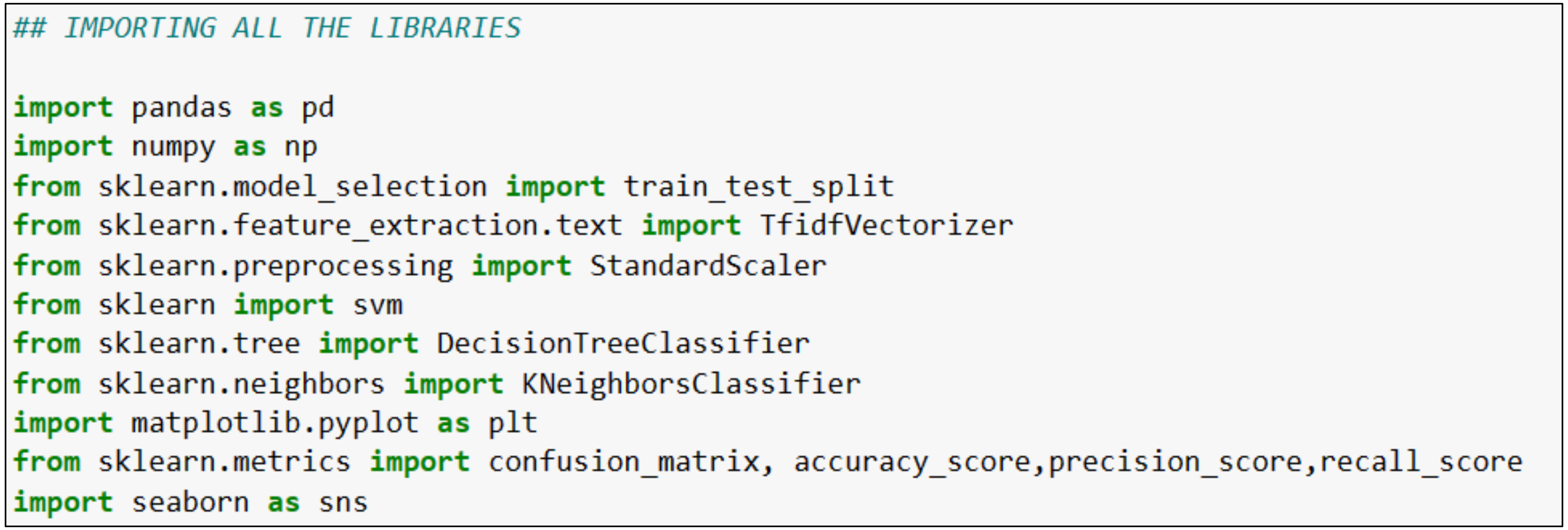
We have used ANACONDA and Jupyter Notebook to run and simulate the code and produce the output and graphs.



	A	B
1	category	messaage
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, èž¢1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea èž¢900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our èž¢100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if thatè¢—s the way u feel. Thatè¢—s the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ì¼Œì½¼1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	Ièª²m going to try for 2 months ha ha only joking

IMPLEMENTATION

LIBRARIES/ MODULES USED:



```
## IMPORTING ALL THE LIBRARIES

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, accuracy_score,precision_score,recall_score
import seaborn as sns
```

IMPLEMENTATION

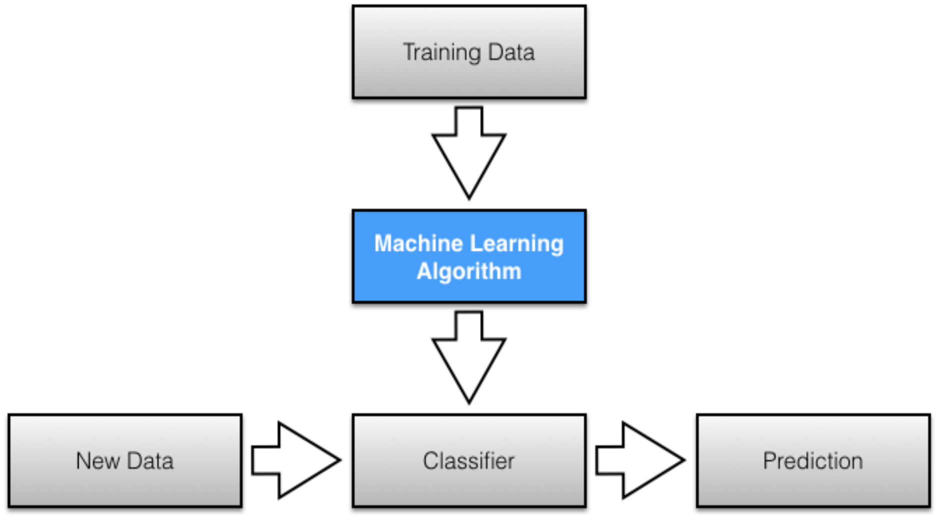
DATA PREPROCESSING:

- The original dataset when imported consists of 5 columns. Out of the five columns, three columns are unnamed or not named and they had null values in them. These three columns don’t have any use so we have removed them.

TRAINING AND TESTING:

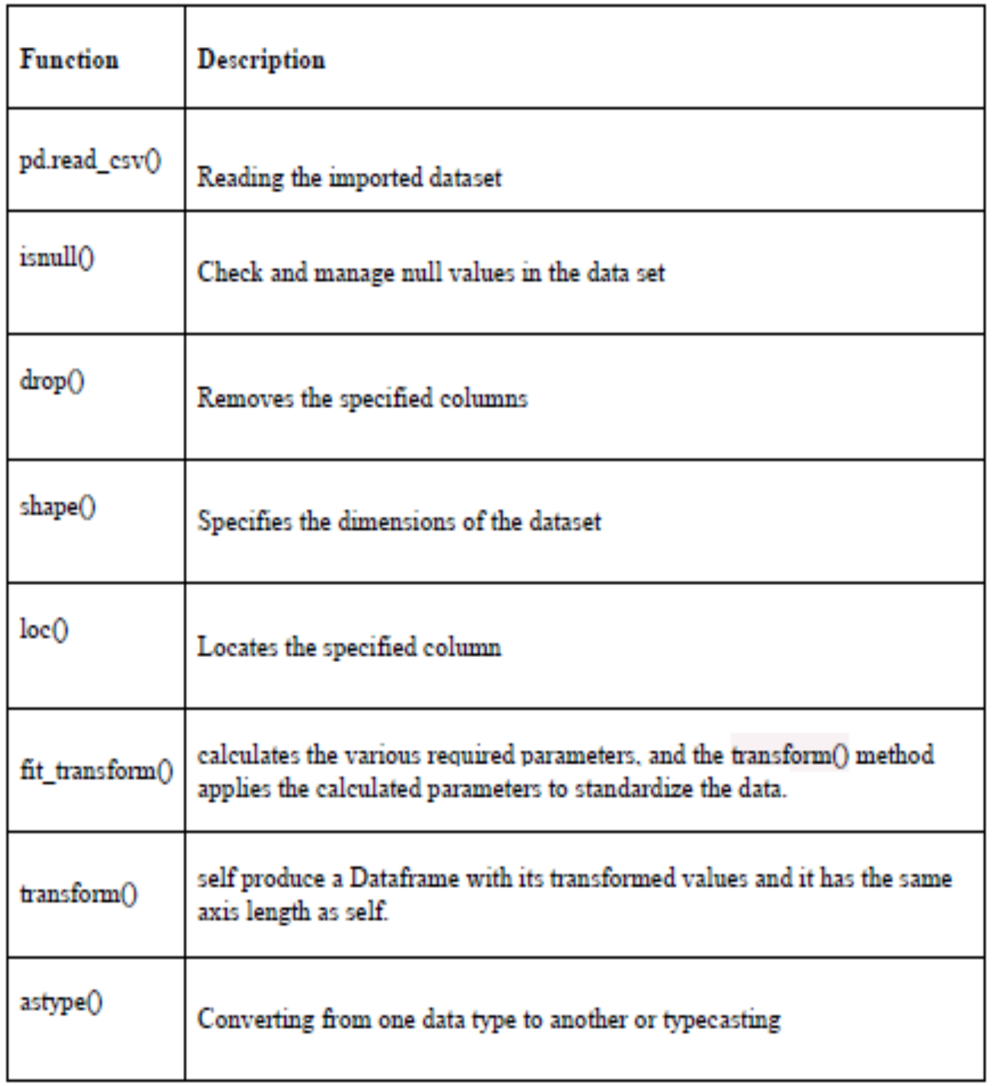
- The ‘train_test_split’ function is used and

the train:test ratio is taken to be **8:2**



IMPLEMENTATION

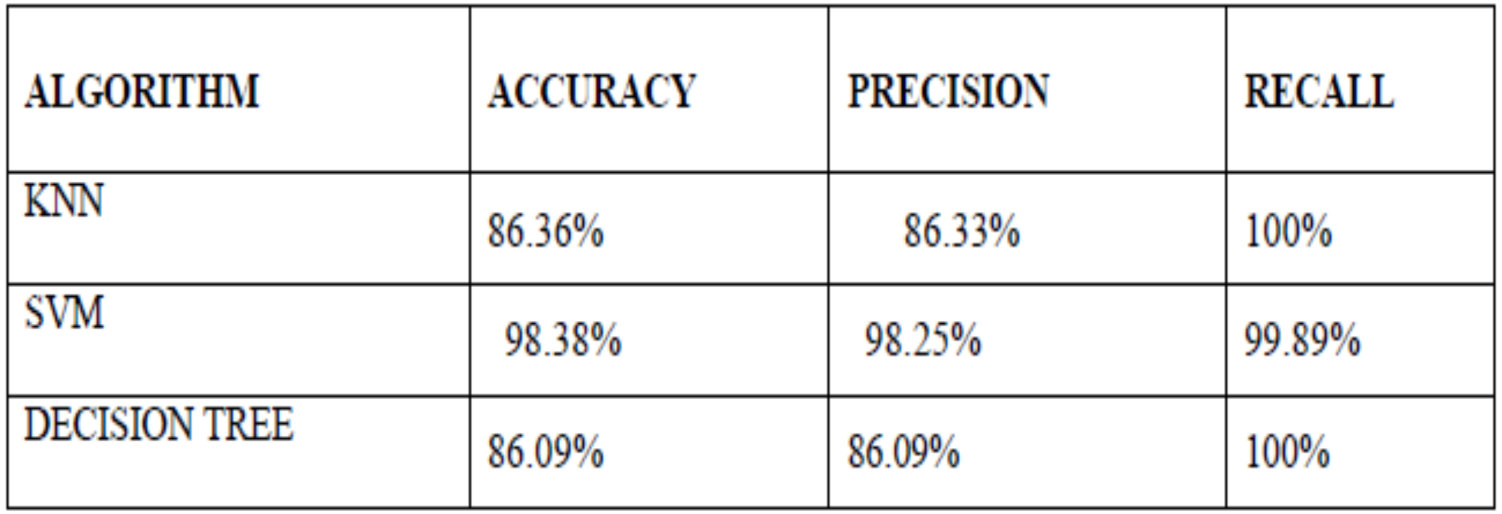
FUNCTIONS USED AND THEIR DESCRIPTION:



Function	Description
pd.read_csv()	Reading the imported dataset
isnull()	Check and manage null values in the data set
drop()	Removes the specified columns
shape()	Specifies the dimensions of the dataset
loc()	Locates the specified column
fit_transform()	calculates the various required parameters, and the transform() method applies the calculated parameters to standardize the data.
transform()	self produce a Dataframe with its transformed values and it has the same axis length as self.
astype()	Converting from one data type to another or typecasting

OUTPUT/RESULT

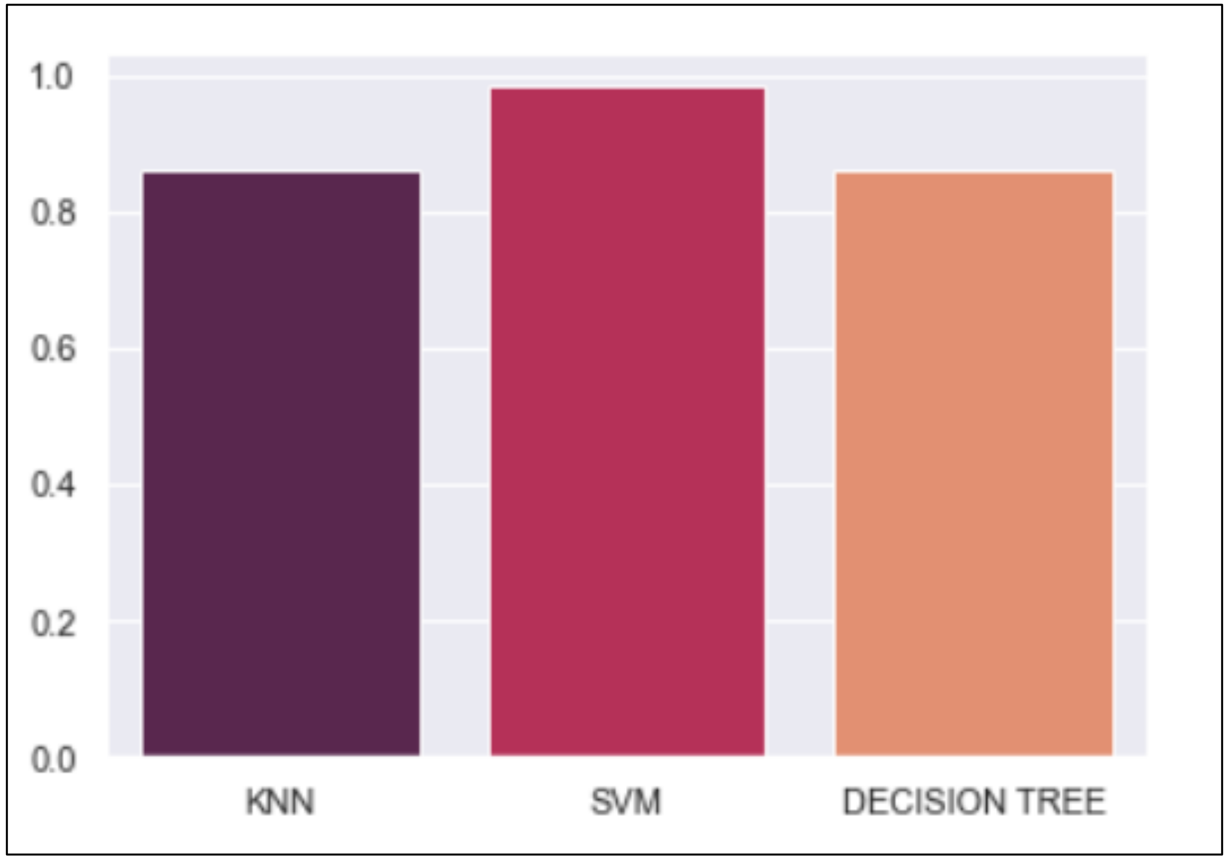
COMPARISION OF THE DIFFERENT MODELS:



ALGORITHM	ACCURACY	PRECISION	RECALL
KNN	86.36%	86.33%	100%
SVM	98.38%	98.25%	99.89%
DECISION TREE	86.09%	86.09%	100%

OUTPUT/RESULT

COMPARISION OF THE DIFFERENT MODELS:



CONCLUSION

We can conclude or infer from the above table that the most accurate model here is SVM.SVM also has the best precision value and both KNN and decision tree have the best recall value. Through this project a comprehensive analysis of various classifiers was implemented on a common dataset. The results were compared based accuracy, precision, and recall score. Models like SVM are a good example with high accuracy.