VISVESVARAYA TECHNOLOGICALUNIVERSITY

"Jnana Sangama", Belgaum 590014, KARNATAKA, INDIA

Project Report

On

"CLASSIFICATION OF UTERINE FIBROID MEDICINES REVIEWS USING MACHINE LEARNING MODEL"

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

Of

Bachelor of Engineering
In
Computer Science & Engineering

Of Visvesvaraya Technological University, Belgaum.

Submitted by:

VIJAY K: USN(1AM15CS212)

SYED ASHFAQ AHMED: USN(1AM15CS194)

SACHIN KUMAR S: USN(1AM16CS405)

JUNAID: USN(1AM15CS223)

Under the Guidance of:

Mrs. VINEETA

Assistant Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AMC ENGINEERING COLLEGE
(NAAC & NBA Accredited, Approved by AICTE, New Delhi & Affiliated to VTU, Belagavi)
18th K.M. Bannerghatta Main Road, Bengaluru – 560083

Year 2019-20

(Accredited by NAAC & NBA, MINISTRY OF HRD, NEW DELHI& Affiliated to VTU Belagavi)

AMC ENGINEERING COLLEGE

(NAAC & NBA Accredited, Approved by AICTE, New Delhi & Affiliated to VTU, Belagavi)

Bengaluru– **560083**

Department of Computer Science & Engineering

CERTIFICATE

Certified that the Project work entitled "CLASSIFICATION OF UTERINE FIBROID MEDICINES REVIEWS USING MACHINE LEARNING MODEL" carried outby bonafide students VIJAY K(1AM15CS212), SYED ASHFAQ AHMED(1AM15CS194), SACHIN KUMAR S (1AM16CS405) , JUNAID AHMED(1AM15CS223) of AMC Engineering College, in partial fulfillment for the award of Bachelor of Engineering in Computer Science & Engineering of Visvesvaraya Technological University, Belgaum during the year 2018-2019. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of Project Work prescribed for the said Bachelor of Engineering degree.

______________________	______________________	____________________
Project Guide	**HOD**	**Principal**
Mrs. Vineeta	**Dr. Latha C A**	**Dr. A G Nataraj .**
Prof. ,Dept. of CSE	**Prof., & Head,Dept. of CSE**	AMCEC
AMCEC	**AMCEC**	

External Viva

External Name	**Signature with Date**
1.________________	**1.**________________
2.________________	**2.**________________

DECLARATION

We the undersigned students of 8th semester Department of Computer Science & Engineering, AMC Engineering College, declare that our project work entitled “” is a bonafide work of ours. Our project is neither a copy nor by means a modification of any other engineering project.

We also declare that this project was not entitled for submission to any other university in the past and shall remain the only submission made and will not be submitted by us to any other university in the future.

Name	USN	Signature
Vijay K	1AM15CS212	________________
Syed Ashfaq Ahmed	1AM15CS194	________________
Sachin Kumar S	1AM16CS405	________________
Junaid Ahmed	1AM15CS223	________________

Table of Contents

CHAPTER 1

INTRODUCTION

The rise of social media such as blogs and social networks has fueled to contribute their contents to the internet. Users can share their experience about a particular product via these blogs and social networks. These experiences nothing but reviews are categorized as positive and negative, which help marketers or industries to improve the product and also help other users in reviewing the product.

Sentiment analysis, also called opinion mining is a field of study that analyzes people's opinions or sentiments about the entities such as products, services and organizations[1].Previous studies of opinion mining provide limitless opportunities for patients to discuss their experiences with drugs. Therefore, even companies get limitless opportunities to receive feedback on their products and services[2-4]because of minority groups of patients on the Internet. Furthermore, recent studies have shown that patient opinions are also useful and important with medical professional opinions[5-8], especially for drugs with afflicting side effects.

Many patients hope to get more information from other patients with similar conditions. They can also share their experience and propose practical ways to alleviate symptoms and side effects of the drugs. These online communities were found to have positive impacts on patient health[9-11].Sentiments (opinions) are of two types: regular and comparative opinions. Regular opinion is often referred to as a simple opinion and further divided into direct opinion and indirect opinion. A direct opinion is an opinion which is expressed directly on an entity or an entity aspect, whereas indirect opinion is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. Comparative opinions express a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities.

These sentiments further classified using supervised learning and unsupervised learning. Supervised learning discovers patterns in data and relates data attributes to class attributes. The values of class attribute for further data instances are then predicted by utilizing these patterns.

However, in some applications, there are no class attributes. Then the sentiments can be classified using unsupervised learning also called clustering, which organizes data instances into groups called clusters such that the data instances in same cluster are similar to each other and data instances in different clusters are very different from each other.

In this paper, a model is designed and an algorithm is proposed in which the drug reviews are crawled from the web and classified drug reviews as positive and negative. A user interface is designed which helps patients and medical professionals in analyzing the results for a particular drug.

The rest of the paper is organized as follows. Section II covers the related work. Implementation of the proposed model is presented in Section III. In Section IV, we compare the proposed model with other algorithms and accuracy, precision and recall is calculated for each drug. Finally, we conclude in Section V.

1.1 Aim and Objective:

1.1.1 Aim:

To develop a drug reviews sentiments classification project.

1.1.2 Objective:

We are developing a project to classify the best drug from the sentiments of drugs reviews using machine learning technique.

1.2 Existing System

Opinion mining (or sentiment analysis) deals with the extraction of specified information (e.g., positive or negative sentiments of a product) from a large amount of text opinions or reviews authored by Internet users. In many situations, solely an overall rating for a review cannot reflect the conditions of different features of a drug products or a service.

1.2.1 Disadvantages:

- Unlike general products or services, drugs have a very limited number of kinds of aspects: price, ease of use, dosages, effectiveness, side effects and people's experiences.
- There are other more technical aspects such as chemical or molecular aspects, but they are almost not mentioned in drug reviews in the external websites.

1.3 Problem Statement

The reviews given in the internet sites by the drug users are not accurate. Hence the recommending of the drugs usage implemented using machine learning is required using drugs.com website.

1.4 Proposed System/Solution

We have proposed a model to in which the drug reviews are crawled from the web and classified drug reviews as positive and negative. A user interface is designed which helps patients and medical professionals in analyzing the results for a particular drug.

LITERATURE SURVEY CHAPTER 2

2.1 Intelligent data mining technique of social media for improving health care

A Data offers many facilities to the end users such as software, organization and platform go on. In this paper, we study about the wisely mining knowledge of social media. Social media becomes much popular from the health care information and Biomedical. This information is commonly shared so healthcare is improves and costs is decrease using opinion which is generated by user. We suggest investigation framework that give attentions on side effects of drugs and also focus on positive and negative response. To improve health care some Clinical documents are mostly useful because it's are free-text data sources. Clinical documents containing information related to symptoms and valuable medications. To extract a Data from large dataset it's become a very popular because users get various ideas from this filtered data. All Data Mining and Knowledge mining become popular because user are process on data and getting information of different area like health, Social, etc. After data processing we focus on users positive and negative opinions. We count this opinions and find out which medication is good, to decide this we also find out the side effects of the medications. Further we focus on the symptoms of the cancer patient. By taking the expert doctors suggestion, we list out the medication of the cancer according to the symptoms and we provide this medication or treatment to the user on our forum. We can expand our research into Data and Knowledge mining of social media and takes the users' views on various drugs of cancer. This daily updated data helps to pharmaceutical industry, doctors, hospitals, and medical staff, for effective future treatments.

2.2 How valuable is medical social media data? Content analysis of the medical web

It is still an open question where to search for complying a specific information need due to the large amount and diversity of information available. In this paper, a content analysis of health-related information provided in the Web is performed to get an overview on the medical content available. In particular, the content of medical Question & Answer Portals, medical weblogs, medical reviews and Wikis is compared. For this purpose, medical concepts are extracted from the text material with existing extraction technology. Based on these concepts, the content of the different knowledge resources is compared. Since medical weblogs describe experiences as well as information, it is of large interest to be able to distinguish between informative and affective posts. For this reason, a method to classify blogs based on their information content is presented, which exploits high-level features describing the medical and affective content of blog posts. The results show that there are substantial differences in the content of various health-related Web resources. Weblogs and answer portals mainly deal with diseases and medications. The Wiki and the encyclopedia provide more information on anatomy and procedures. While patients and nurses describe personal aspects of their life, doctors aim to present health-related information in their blog posts. The knowledge on content differences and information content can be exploited by search engines to improve ranking, search and to direct users to appropriate knowledge sources

2.3 Drug Reviews using Characteristic Mining Model of Probabilistic Analysis

Nowadays finding chronic diseases and drugs are becoming more important for supporting the patient resource information. Extracting patient information from the text is most challenging and also critical. So for extracting patient information from these substantial bodies of texts we are using so many opinion mining techniques. In this paper we are extracting information from these substantial bodies of texts using one of the mining models of classification approach. The classification technique used is Naïve Bayesian classifier which is used for finding the causes that occur by using over doses of drugs and also to find the type of side effect that will occur. After completion of classification we are grouping the related drugs which are causing the same side effects by using Word Comparator Clustering algorithm. By implementing this application we can improve the efficiency and also provide more classification accuracy.

2.4 Do virtual communities matter for the social support of patients?: Antecedents and effects of virtual relationships in onlinecommunities.

The purpose of this paper is to explore whether online communities meet their potential of providing environments in which social relationships can be readily established to help patients cope with their disease through social support. The paper aims to develop and test a model to examine antecedents of the formation of virtual relationships of cancer patients within virtual communities (VCs) as well as their effects in the form of social assistance. Design/methodology/research - Data were collected from members of virtual patient communities in the German-speaking internet through an online survey to which 301 cancer patients responded. The data were analyzed with partial least square (PLS) structural equation modeling. Findings - Virtual relationships for patients are established in VCs and play an important role in meeting patients' social needs. Important determinants for the formation of virtual relationships within virtual communities for patients are general internet usage intensity (active posting vs lurking) and the perceived disadvantages of CMC. The paper also found that virtual relationships have a strong effect on virtual support of patients; more than 61 per cent of the variance of perceived social assistance of cancer patients was explained by cancer-related VCs. Emotional support and information exchange delivered through these virtual relationships may help patients to better cope with their illness. Research limitations/implications - In contrast to prior research, known determinants for the formation of virtual relationships (i.e. marital status, educational status, gender, and disease-related factors such as the type of cancer as control variables, as well as general internet usage motives, and perceived advantages of CMC as direct determinants) played a weak role in this study of German cancer patients. Studies on other patient populations (i.e. patients with other acute illnesses in other cultures) are needed to see if results remain consistent. Practical implications - Participants and administrators of patient VCs have different design criteria for the improvement of VCs for patients (e.g. concerning community management, personal behaviour and the usage of information in online communities). Once the social mechanisms taking place in online communities are better understood, the systematic redesign of online communities according to the needs of their users should be given priority. Originality/value - Little research has been conducted examining the role of VCs for social relationships and social networks in general and for patients in particular. Antecedents and effects of virtual social relationships of patients have not been sufficiently theoretically or empirically researched to be better understood. This research combines various determinants and effects of virtual relationships from prior related research. These are integrated into a conceptual model and applied empirically to a new target group, i.e. VCs for patients.

2.5 Investigating Web Search Strategies and Forum Use to Support Dietand Weight Loss

Healthcare is shifting from being reactive to preventive, with a focus on maintaining general wellness through positive decisions on diet, exercise, and lifestyle. In this paper, we investigate search behavior as people navigate the Web and find support for dietary and weight loss plans. Inspecting the Web search logs of nearly 2,000 users, we show that people progressively narrow their searches to support their progress through these plans. Interestingly, people that visit online health forums seem to progress through the plans‟ phases more quickly. Based on these results, we conducted a survey to further explore the roles and importance of online forums in supporting dieting and weight loss.

CHAPTER 3

REQUIREMENT SPECIFICATION

3.1 HARDWARE REQUIREMENTS:-

Processor	:	Dual Core
Speed	:	1.1 G Hz
RAM	:	4 GB (min)
Hard Disk	:	20 GB
Key Board	:	Standard Windows Keyboard
Mouse	:	Two or Three Button Mouse
Monitor	:	SVGA

3.2 SOFTWARE REQUIREMENTS:-

Operating System	:	Windows XP,7,8,10
Technology	:	Python
Front End	:	Tkinter
IDLE	:	Python 2.7 or 3.0 or higher
Database	:	MySQL

CHAPTER 4

SYSTEM DESIGN AND ARCHITECTURE

4.1 DESIGN

System design is the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development "blends the perspective of marketing, design, and manufacturing into a single approach to product development," then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

4.1.1 Architecture Design/ System Architecture

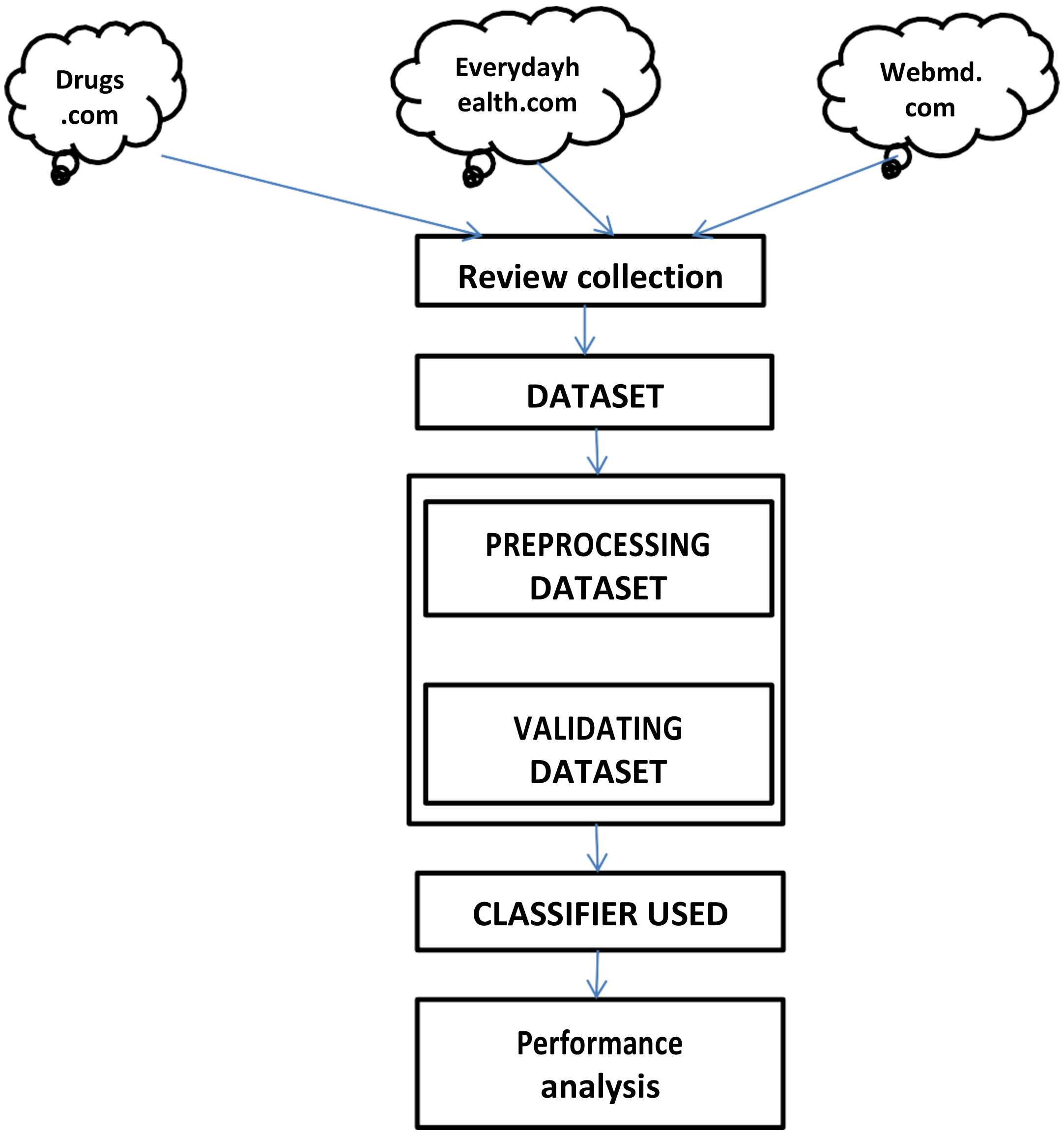


Fig4.1: Architecture Design/ System Architecture

4.2 Flow Chart/ Data Flow Diagram

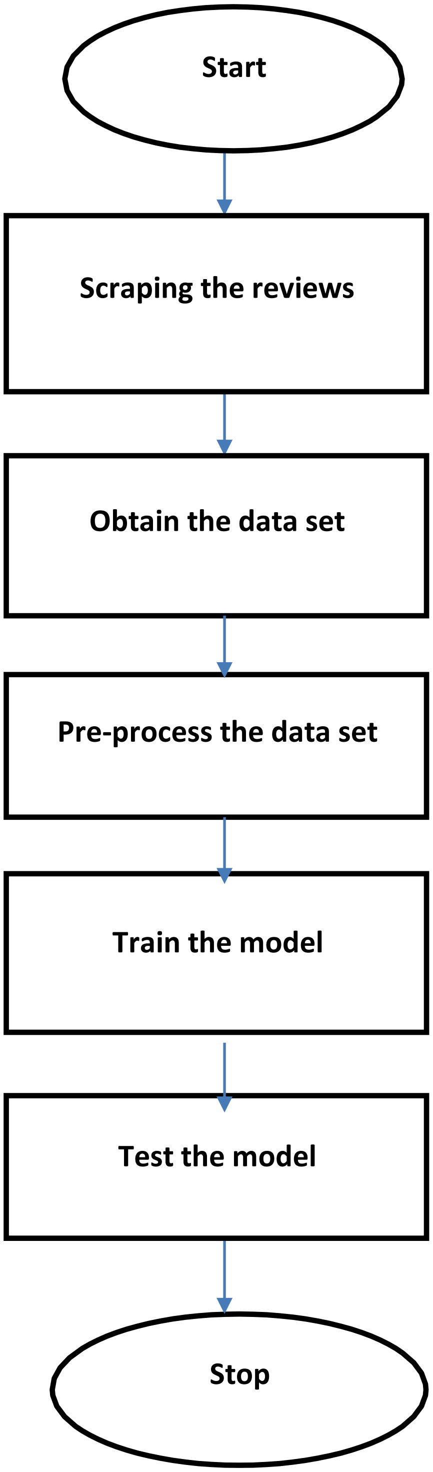


Fig4.2 Flow Chart

UML Diagrams

Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created, by the Object Management Group.

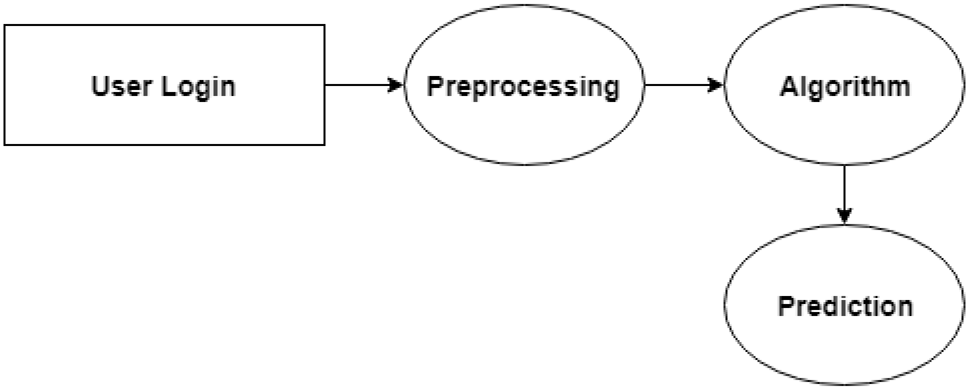
Use Case Diagrams

A use case diagram at its simplest is a graphical representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with thesystem.

4.2.1 Components of DFD Diagram

DATA FLOW DIAGRAM

A data flow diagram is a graphical representation of the "flow" of data through an information system, modeling its *process* aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.



4.2.2 Development Methods/ Algorithm

- Support Vector Machine algorithm
- K-Nearest Neighbor algorithm
- Naïve Bayes algorithm

SVM (Support Vector Machine):

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Decision Tree is a tree where each node represents a feature (attributes), each link (branch) represents a decision and each leaf represents an outcome (categorical or continuous value).A Decision Tree is a Supervised Machine Learning algorithm which looks like an inverted tree, wherein each node represents a predictor variable (feature), the link between the nodes represents a Decision and each leaf node represents an outcome (response variable).

KNN (K-Nearest Neighbors):

In this project I used KNN because it is simple to implement & very straight forward. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data .Here I am given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Naive – Bayes Classifier:

Naive Bayes is a surprisingly powerful algorithm for predictive modeling. It is a statistical classifier which assumes no dependency between attributes attempting to maximize the posterior probability in determining the class. Theoretically, this classifier has the minimum error rate, but may not be the case always. Inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. This model is associated with two types of probabilities which can be calculated from the training dataset directly:

a) The probability of everyclass.

a. b) The conditional probability of each class with each x value. According to Bayesian theorem

$P(A|B)=P(A)* P(B/A)/P(B)$,

Where

$P(B|A)=P(A\cap B)/P(A)$ .Bayesian classifier calculates conditional probability of an instance belonging to each class, based on the above formula, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. If these probabilities are calculated, then the probabilistic model can be implemented to make predictions with new data using Naïve Bayes Theorem. When the data is real-valued it is likely to assume a Gaussian distribution (bell curve). Thus, these probabilities can easily be estimated. Naive Bayes is called naive because of assuming each input variable independent.

CHAPTER 5

IMPLEMENTATION AND CODING

5.1 LANGUAGE USED FOR IMPLEMENTATION

5.1.1 Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.[28]

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.[29]

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released in 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python 3.

The Python 2 language, i.e. Python 2.7.x, was officially discontinued on 1 January 2020 (first planned for 2015) after which security patches and other improvements will not be released for it.[30][31] With Python 2's end-of-life, only Python 3.5.x[32] and later are supported.

Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, an open source[33] reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development.

5.1.2 Features of Python

There are many features in python,some of which are discussed below:

- Easy to code
- Free and open source
- Object oriented language
- GUI programming support
- High level language
- Extensible feature
- Python is portable language

- Python is integrrated language
- Interpreted language
- Large standard library

5.1.3 Advantages of python

Python is a high level,interpreted and general purpose dynamic programming language that focuses on code readability.

- Presence of third-party modules
- Extensive support libraries
- Open source and community development
- Easy to learn
- User-friendly data structures

5.1.4 Applications

- GUI based desktop applications
- Web frameworks and applications
- Enterprise and business applications
- Operating systems
- Language development
- Prototyping

5.2 CODING

5.2.1SVM CODE

```
def SVM():
print('\n\n--------SVM algorithm--------- ')
import numpy as np #multi-dimensional arrays and matrices
import pandas as pd #data manipulation and analysis
import matplotlib.pyplot as plt #graph plot
import warnings
warnings.filterwarnings('ignore')

#To read the dataset
dataset = pd.read_csv('Reviews4.csv', encoding='latin1')

#to take the reviews/text from dataset
reviews = [ i for i in dataset['Comments']]
#print(tweet)

#classification of positive and negative sentiments
d1 = {'positive':1,'negative':0}

#stores accuracy of all the models
l=[]

#(Tokenization is the act of breaking up a sequence of strings into pieces such as words,keywords, phrases,
symbols and other elements called tokens)
#CountVectorizer converts a collection of text documents to a matrix of token counts
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 10)

#initial fitting of parameters on the training set
#toarray() can be used to populate a numpy array
X = cv.fit_transform(reviews).toarray()

#creating rows and columns
y = dataset.iloc[:, 1].values

#split data into training and testing set
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test= train_test_split(X, y, train_size=0.7, random_state=42)
```

```
#To print accuracy
 from sklearn.metrics import accuracy_score
 def print_score(clf, X_train, Y_train, X_test, Y_test,train=True):
     if train:
         print("\naccuracy_score: \t {0:.4f}".format(accuracy_score(Y_train,clf.predict(X_train))))
         #{0:.4f} used for printing 4 values after decimal point

 #SVM algorithm
 from sklearn.svm import SVC
 classifier = SVC(kernel = 'linear', random_state = 0)
 classifier.fit(X_train, Y_train)

 #function call- def print_score
 #Scores for training data
 print_score(classifier,X_train,Y_train,X_test, Y_test,train=True)

 #Total count of all sentiments

 print('\n\n\n---------Total count of all sentiments -----')

 import pandas as pd
 import matplotlib.pyplot as plt

 data = pd.read_csv('Reviews4.csv',encoding='latin1')

 df = pd.DataFrame(data, columns= ['Sentiment'])

 print('\nTotal count of all sentiments:\n\n')

 print(df['Sentiment'].value_counts())
```

5.2.2 KNN CODE

```
def KNN():
print('\n---------KNN algorithm----------- ')
import numpy as np #multi-dimensional arrays and matrices
import pandas as pd #data manipulation and analysis
import matplotlib.pyplot as plt #graph plot
import warnings
warnings.filterwarnings('ignore')

#To read the dataset
dataset = pd.read_csv('Reviews4.csv', encoding='latin1')

#to take the reviews/text from dataset
tweet = [ i for i in dataset['Comments']]
#print(tweet)

#classification of positive and negative sentiments
d1 = {'positive':1,'negative':0}

#stores accuracy of all the models
l=[]

#(Tokenization is the act of breaking up a sequence of strings into pieces such as words,keywords, phrases,
symbols and other elements called tokens)
#CountVectorizer converts a collection of text documents to a matrix of token counts
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 10)

#initial fitting of parameters on the training set
#toarray() can be used to populate a numpy array
X = cv.fit_transform(tweet).toarray()

#creating rows and columns
y = dataset.iloc[:, 1].values

#split data into training and testing set
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test= train_test_split(X, y, train_size=0.7, random_state=42)
```

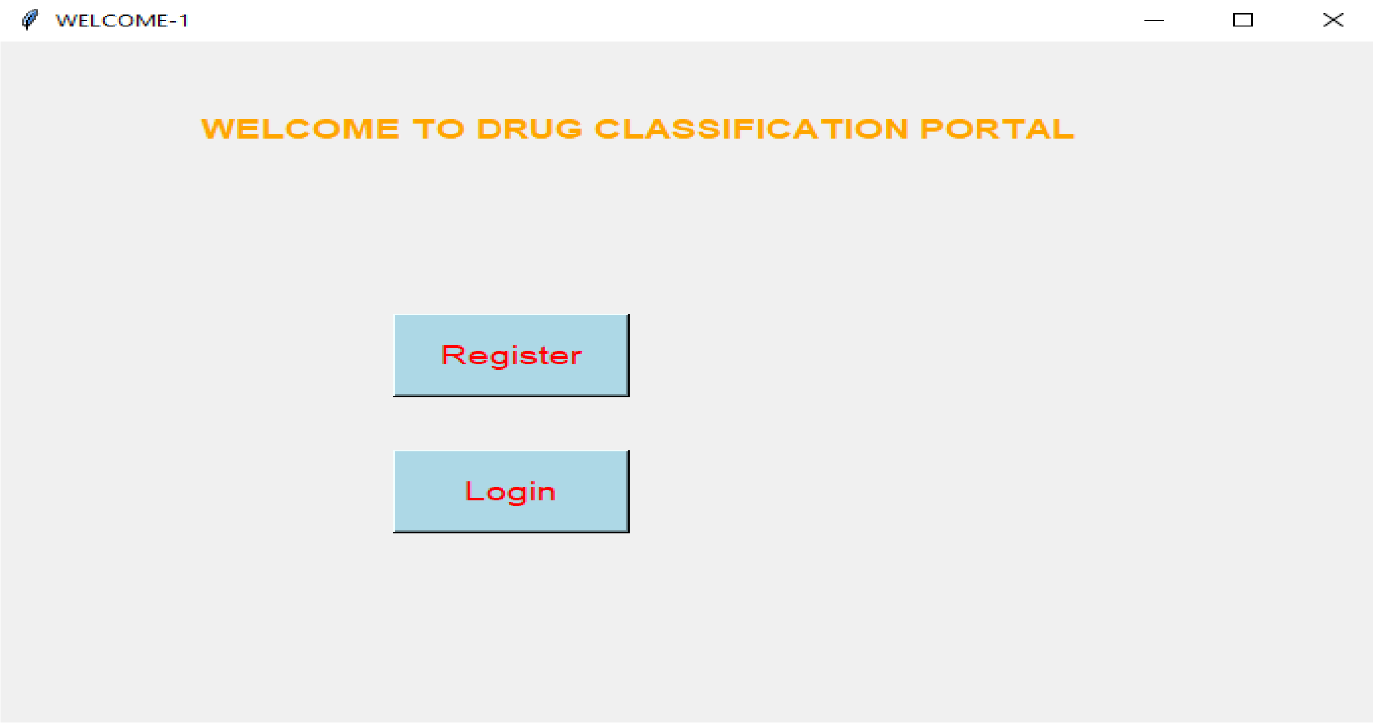
5.2.3 Naive – Bayes Classifier

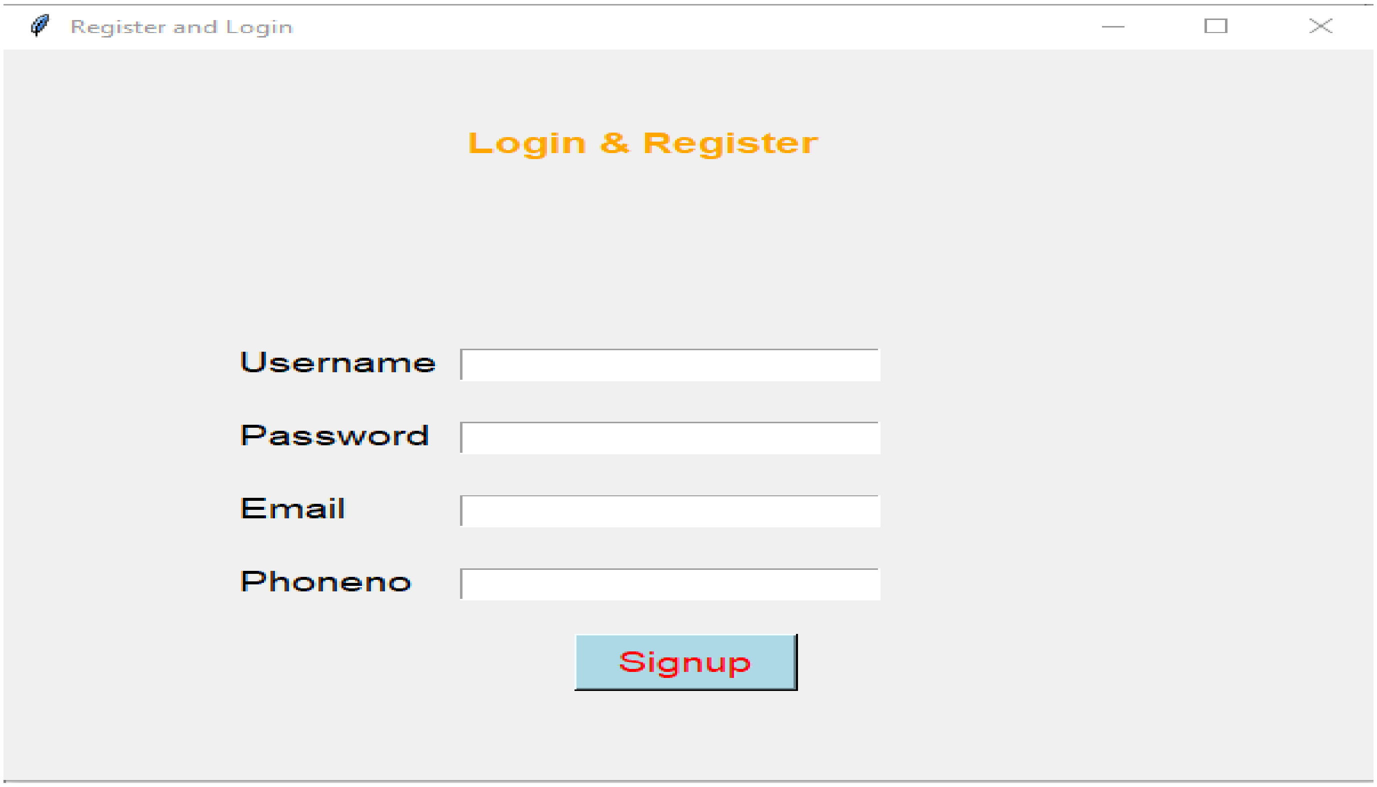
```
def Naive_Bayes():
print('\n\n---------Naive Bayes algorithm --------------')
import numpy as np #multi-dimensional arrays and matrices
import pandas as pd #data manipulation and analysis
import matplotlib.pyplot as plt #graph plot
from sklearn import svm  #
import warnings
warnings.filterwarnings('ignore')

#To read the dataset
dataset = pd.read_csv('Reviews4.csv', encoding='latin1')

#to take the reviews/text from dataset
tweet = [ i for i in dataset['Comments']]
#print(tweet)

d1 = {'positive':1,'neutral':0,'negative':-1}

#stores accuracy of all the models
l=[]

#CountVectorizer provides a simple way to both tokenize a collection of text documents and build a
vocabulary of known words
#(Tokenization is the act of breaking up a sequence of strings into pieces such as words,keywords, phrases,
symbols and other elements called tokens)
#CountVectorizer converts a collection of text documents to a matrix of token counts
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 10)

#initial fitting of parameters on the training set
#toarray() can be used to populate a numpy array
X = cv.fit_transform(tweet).toarray()

#creating rows and columns
y = dataset.iloc[:, 1].values

#split data into training and testing set
from sklearn.model_selection import train_test_split #split data into training and testing set
X_train,X_test,Y_train,Y_test= train_test_split(X, y, train_size=0.7, random_state=42)

#Defining a funtion print_score
#To print accuracy
from sklearn.metrics import accuracy_score
def print_score(clf, X_train, Y_train, X_test, Y_test,train=True):
    if train:
        print("\naccuracy_score: \t {0:.4f}".format(accuracy_score(Y_train,clf.predict(X_train))))
    #{0:.4f} used for printing 4 values after decimal point

#Naive-Bayes algorithm
#Gaussian Naive Bayes (GaussianNB)
from sklearn.naive_bayes import GaussianNB
model = GaussianNB()
#fit the Gaussian model with data
model.fit(X_train,Y_train)

#Scores for training data
#Function call
print_score(model,X_train,Y_train,X_test,Y_test,train=True)

top = {"DrugName":"Sentiment","Traxemic acid":263 , "Lupron":35, 'Ulipristal':2}

with open('top.csv', 'w') as f:
  for key in top.keys():
    f.write("%s,%s\n"%(key,top[key]))

#to read top college(first row)
import pandas as pd
data = pd.read_csv("top.csv", nrows=1)
print("The best drug is:\n\n", data)

import matplotlib.pyplot as plt

labels = 'Traxemic acid', 'Lupron', 'Ulipristal'
sizes = [263,35,2]
colors = ['green', 'yellow','red']
plt.pie(sizes,labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140)
plt.axis('equal')
plt.title('Pie-Chart of all 3 drugs')
plt.show()
```

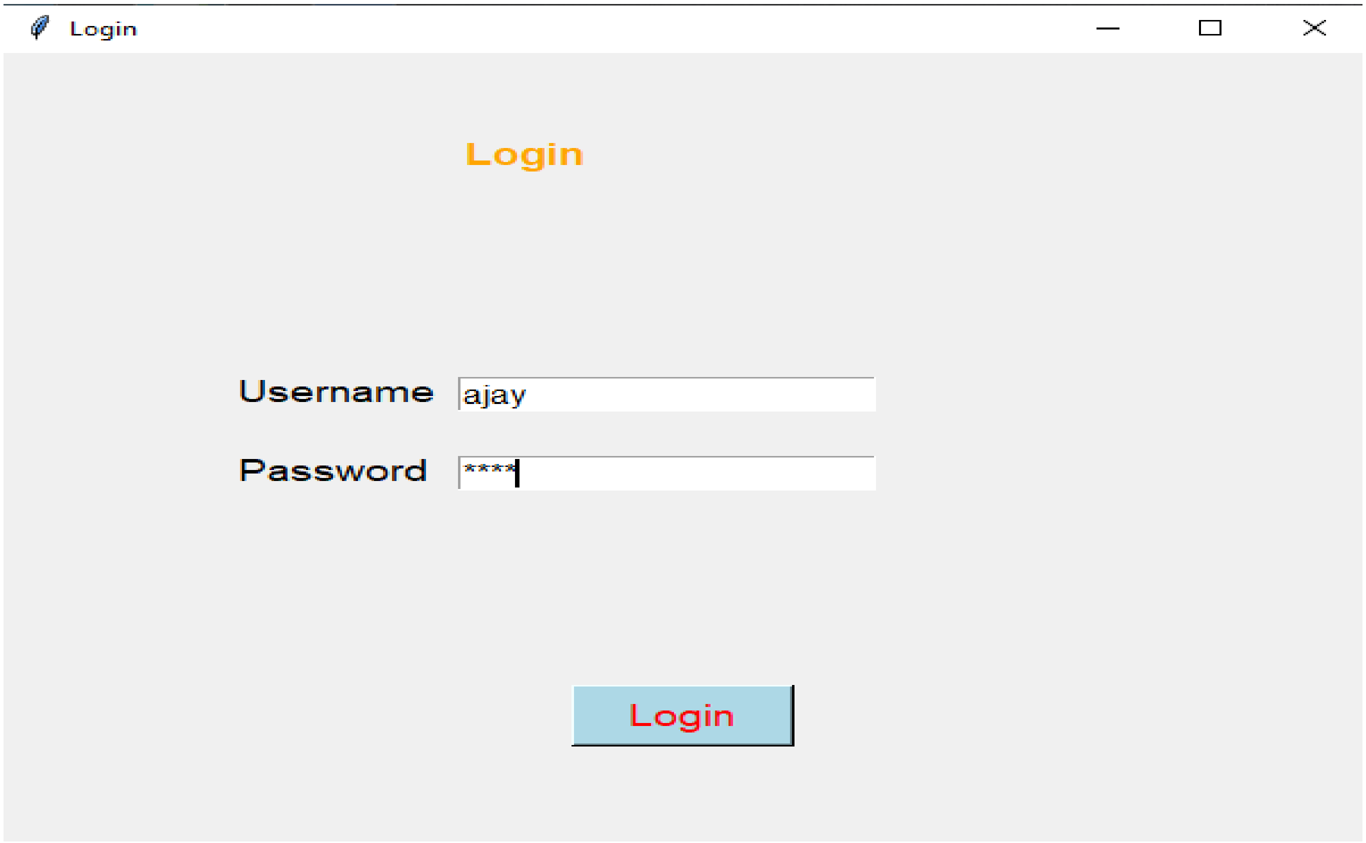
5.2.4 USER INTERFACE

```
def UI():

R1 = Tk()
R1.title('DRUG CLASSIFICATION')
R1.geometry('600x400')

w2 = Label(R1, justify=LEFT, text="Drug Classification using Machine Learning", fg="RED")
w2.config(font=("Elephant", 15))
w2.place(x=80,y=30)

oc = StringVar(R1)
oc.set("----Select tablet ---")


def function2():
    def function1(x):
        if (x == "Traxemic acid"):
            R3 = Tk()
            R3.title('Traxemic acid')
            R3.geometry('500x400')

            w2 = Label(R3, justify=LEFT, text="Drug Classification using Machine Learning-Traxemic acid",
            fg="Green")
            w2.config(font=("Elephant", 10))
            w2.place(x=80,y=18)

            import pandas as pd
            d1 = pd.read_csv('Reviews4.csv',encoding='latin1',index_col ="DrugName")
            first = d1["Sentiment"]

            a = first['Traxemic acid'].value_counts(normalize=True).mul(100).round(1).astype(str)+'%'
            a1 = a.to_string()
            L1=Label(R3, text=str(a1), font=('Times',12,'bold'),fg="orange")
            L1.place(x=280, y=200)

            L2=Label(R3,text="Positive =",font=('Times',12,'bold'))
            L2.place(x=210, y=200)

            L3=Label(R3,text="Negative = ",font=('Times',12,'bold'))
            L3.place(x=210, y=220)

        elif (x=="Lupron"):
            R4 = Tk()
            R4.title('Lupron')
            R4.geometry('500x400')
```

SCREENSHOTS

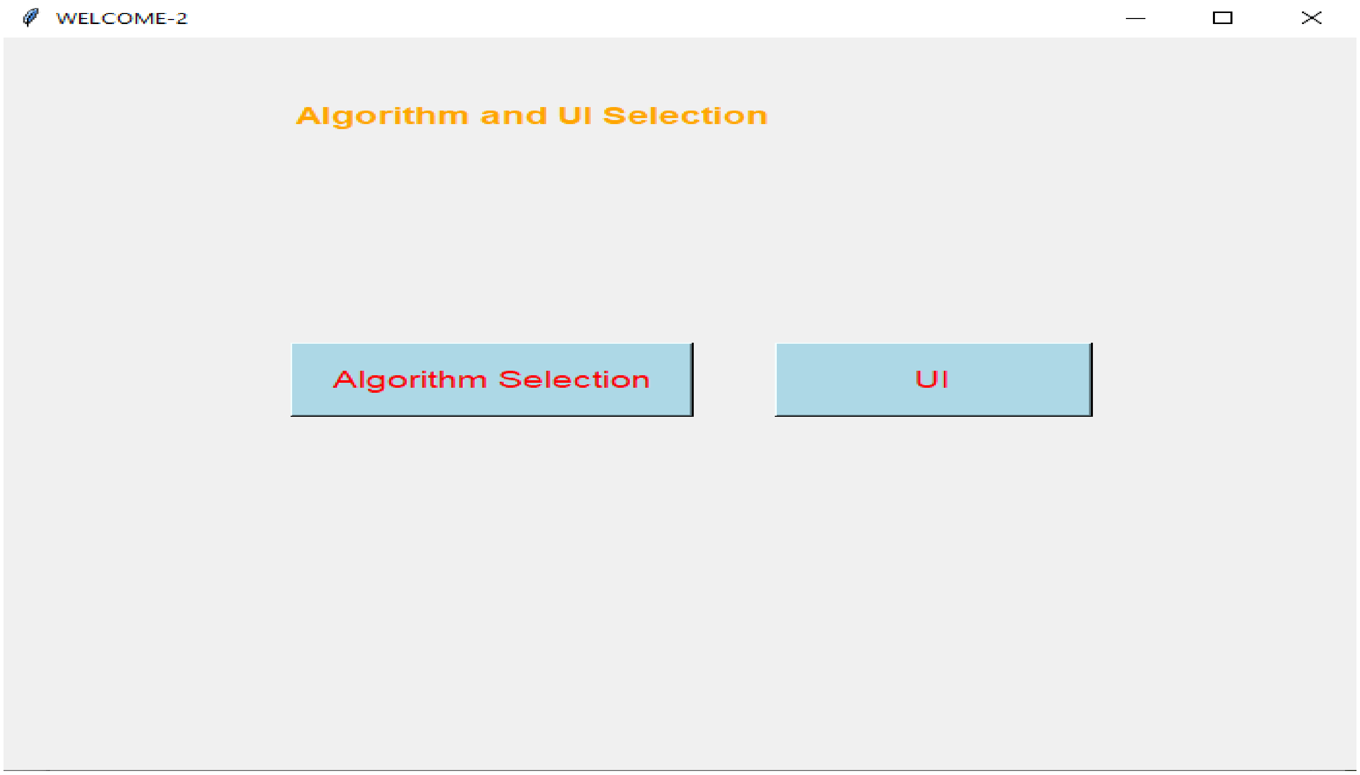


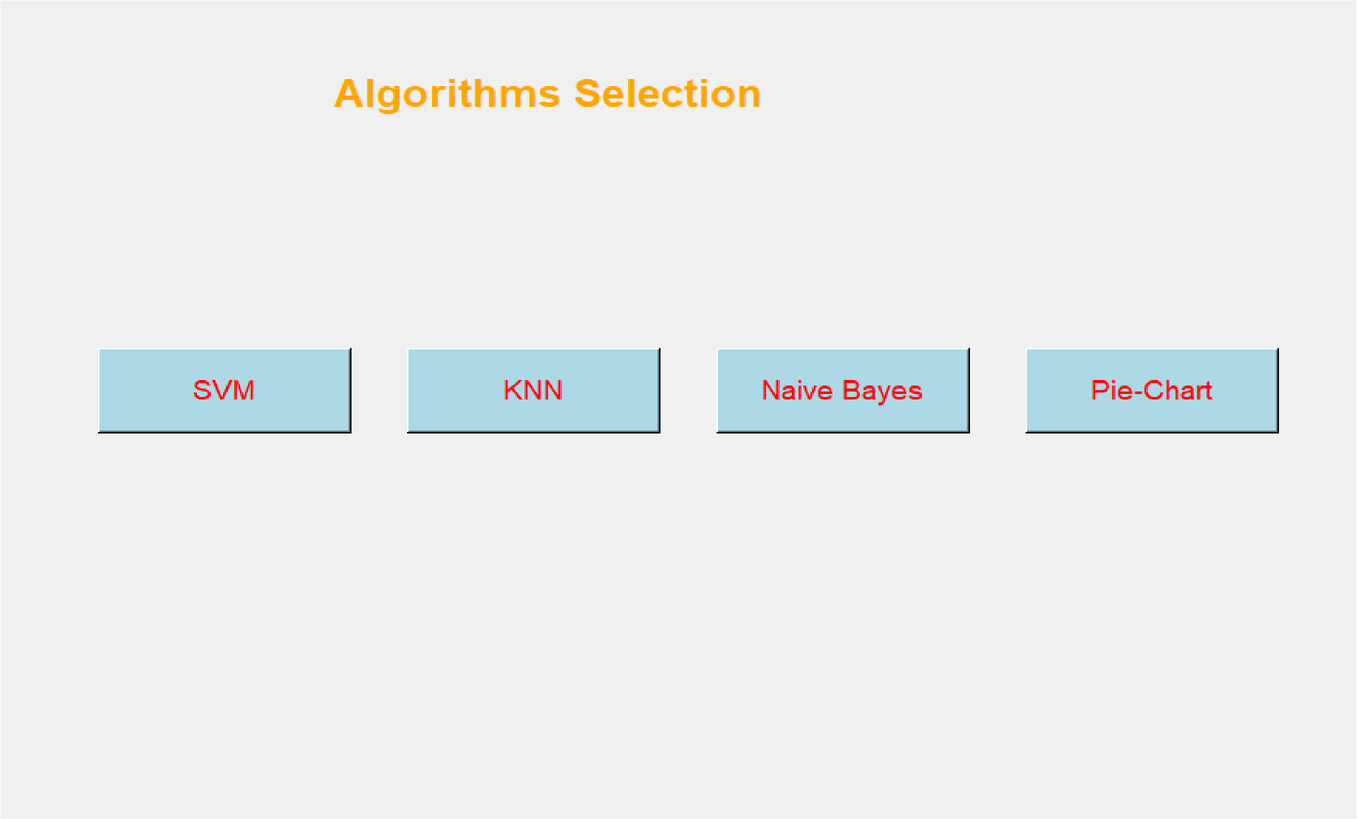


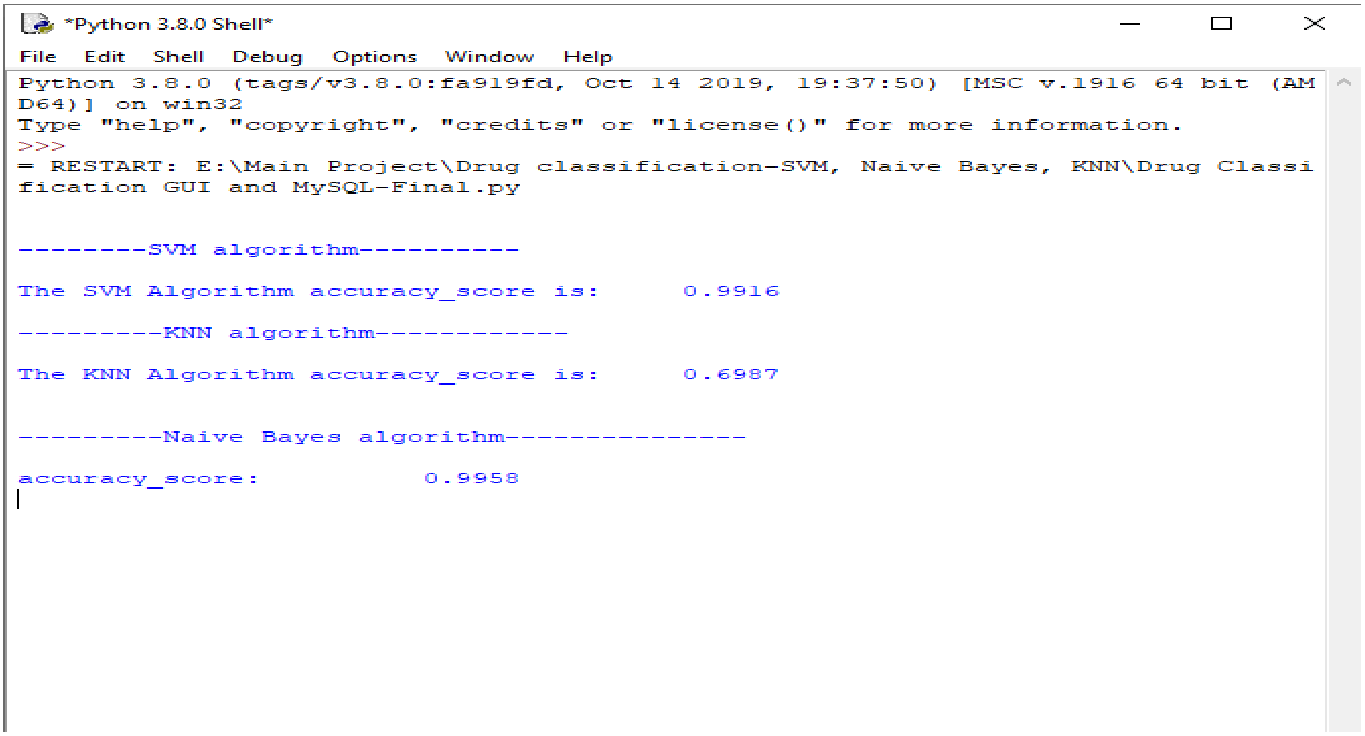


Username ajay

Password ****







```
Python 3.8.0 (tags/v3.8.0:fa919fd, Oct 14 2019, 19:37:50) [MSC v.1916 64 bit (AM
D64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: E:\Main Project\Drug classification-SVM, Naive Bayes, KNN\Drug Classi
fication GUI and MySQL-Final.py


--------SVM algorithm----------

The SVM Algorithm accuracy_score is:     0.9916

---------KNN algorithm------------

The KNN Algorithm accuracy_score is:     0.6987


---------Naive Bayes algorithm---------------

accuracy_score:          0.9958
```

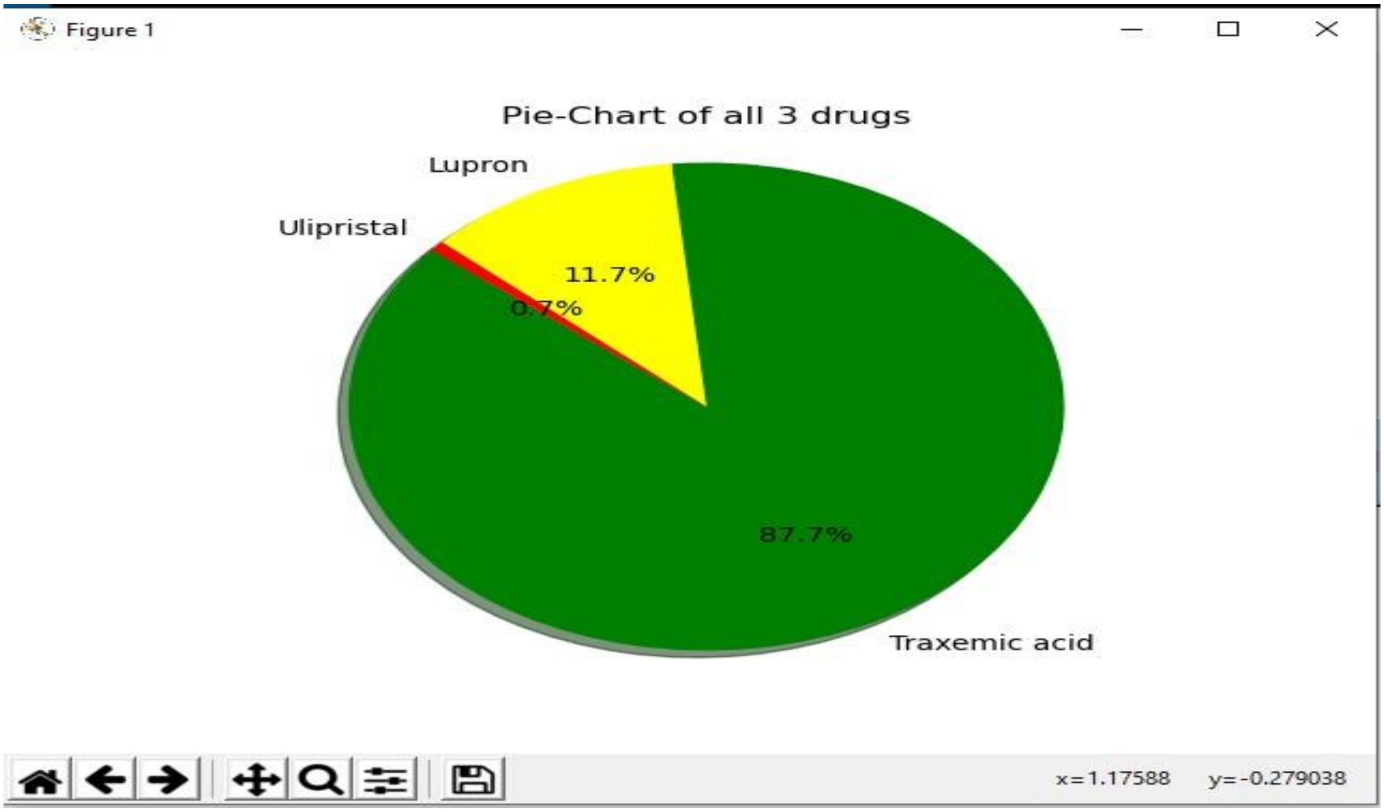
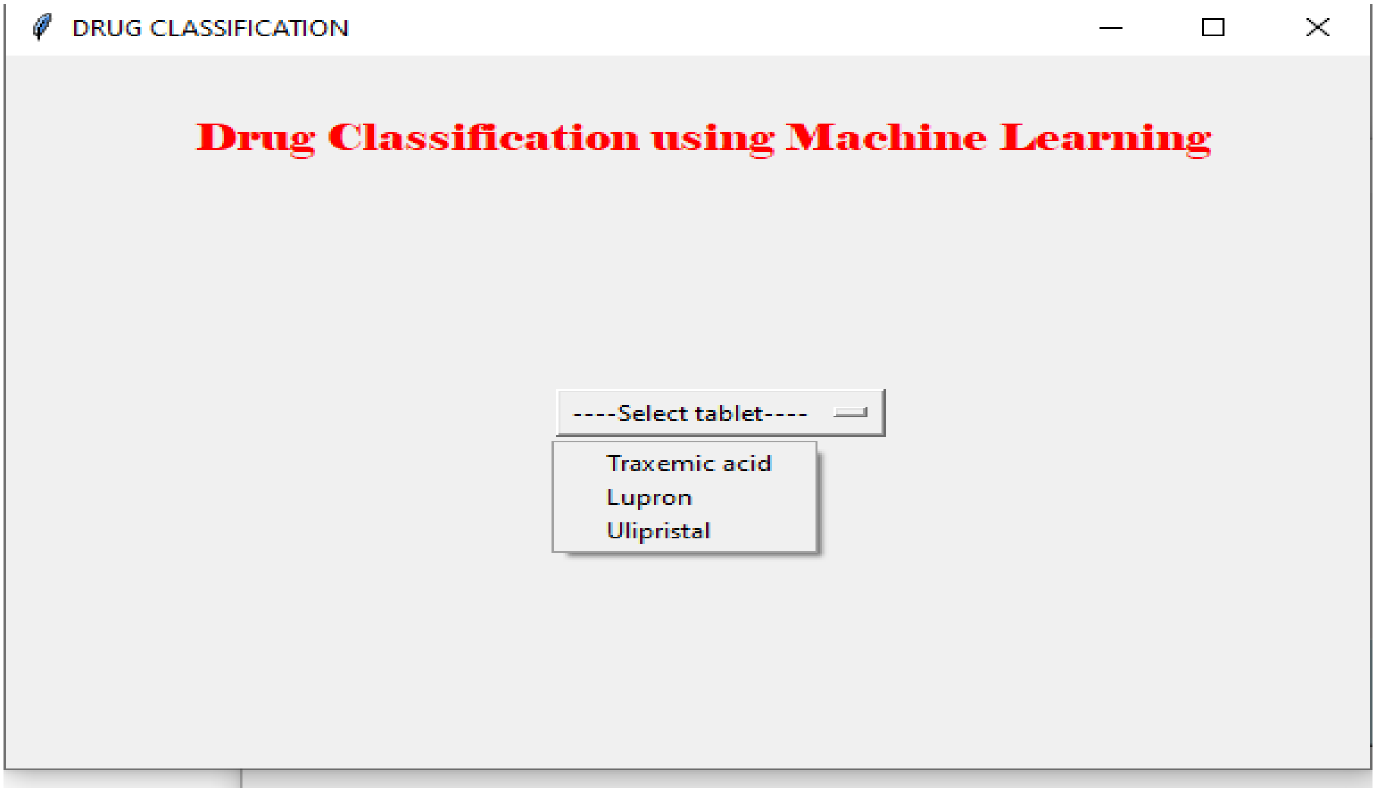
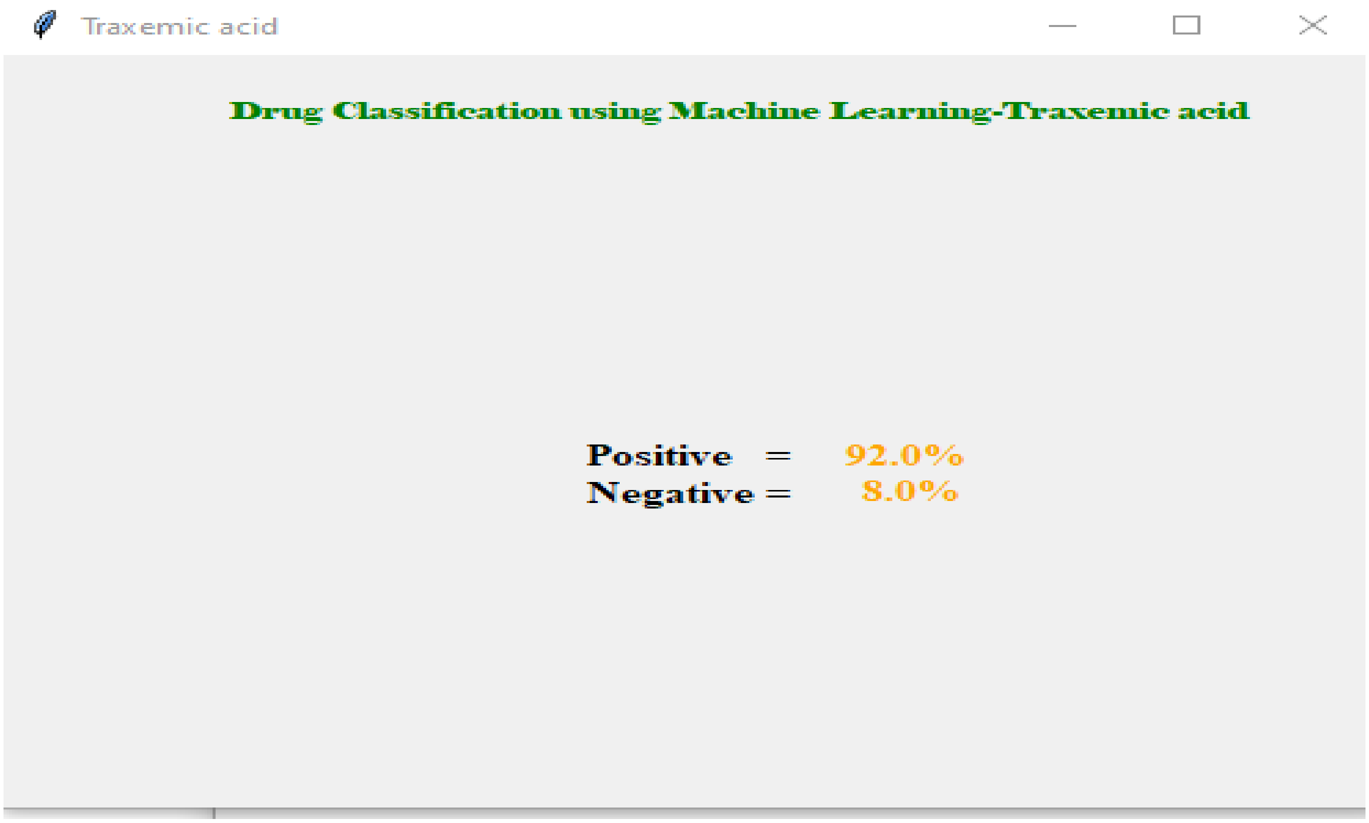


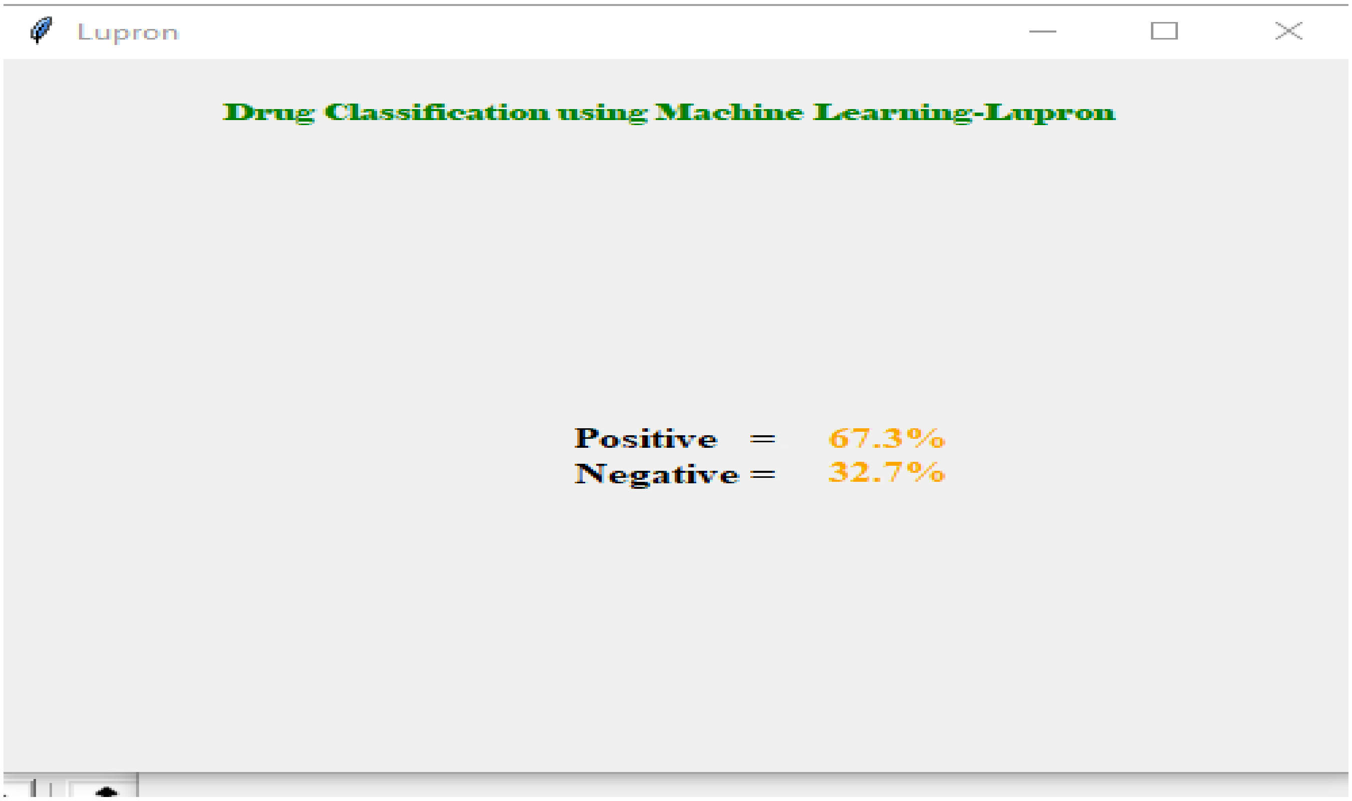
Figure 1
Pie-Chart of all 3 drugs
Lupron
Ulipristal
11.7%
0.7%
87.7%
Traxemic acid
x=1.17588 y=-0.279038


DRUG CLASSIFICATION
Drug Classification using Machine Learning
----Select tablet----
Traxemic acid
Lupron
Ulipristal



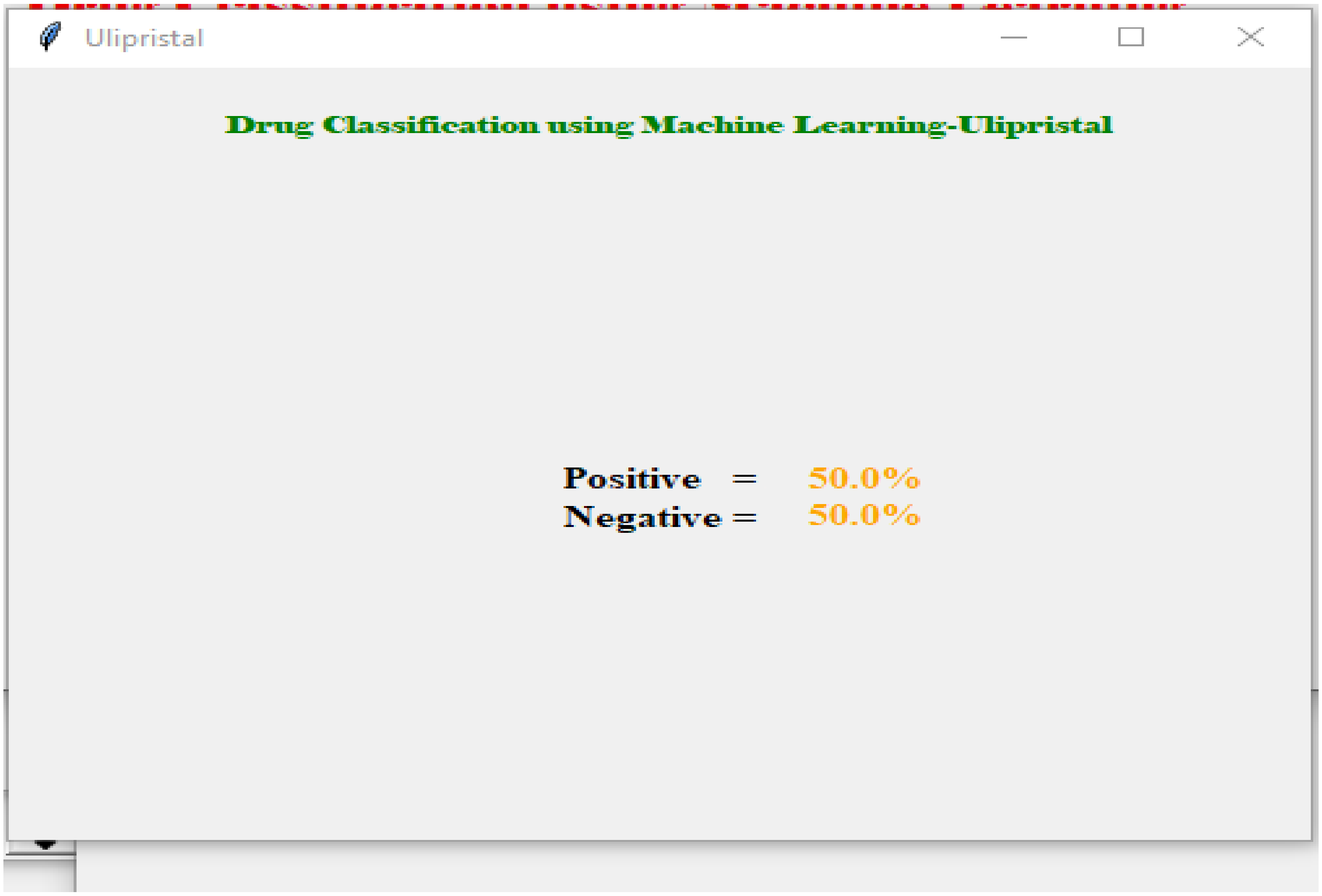
Drug Classification using Machine Learning-Traxemic acid

```
Positive   =    92.0%
Negative =      8.0%
```



Drug Classification using Machine Learning-Lupron

```
Positive   =    67.3%
Negative =    32.7%
```



Drug Classification using Machine Learning-Ulipristal

```
Positive   =    50.0%
Negative =    50.0%
```

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this paper, we have proposed novel model SVM to classify sentiments of drugs reviews. The results and performance analysis shows best performance with SVM classifier ensemble with an accuracy of 97.46%.The proposed model works efficiently for large data set of reviews crawled using a review collection algorithm. A user interface is also designed for the SVM model. Thus the model can be used for any drug reviews analysis.

REFERENCES

[1] Bing Liu,”Sentiment Analysis and Opinion Mining”, liub@cs.uic.edu, *April 22,2012.* [2] A. Ochoa, A. Hernandez, L. Cruz, J. Ponce, F. Montes, L. Li, andL. Janacek, “Artificial societies and social simulation using ant colony,particle swarm optimization and cultural algorithms,” in New Achieve-ments in Evolutionary Computation, P. Korosec, Ed. Rijeka, *m*Croatia:Intech, pp. 267–297, 2010

[2] A. Ochoa, A. Hernandez, L. Cruz, J. Ponce, F. Montes, L. Li, andL. Janacek, “Artificial societies and social simulation using ant colony,particle swarm optimization and cultural algorithms,” in New Achieve-ments in Evolutionary Computation, P. Korosec, Ed. Rijeka, Croatia:Intech, pp. 267–297, 2010

[3] Cornell and W. Cornell. (2013). How Data Mining Drives Pharma: Information as a Raw Material and Product [Webinar]. [Online]. Available:http://acswebinars.org/big-data W.

[4] L. Toldo, “Text mining fundamentals for business analytics,” presented at the 11th Annu. Text Soc. Analytics Summit, Boston, MA, USA, 2013.

[5] J. Sarasohn-Kahn, “The wisdom of patients: Health care meetsonline social media,” California Healthcare Foundation, Tech.Rep., 2009.

[6] K. Denecke and W. Nejdl, “How valuable is medical social media data? content analysis of the medical web,” *J. Inform. Sci.*, vol. 179, no. 12, pp. 1870–1880, 2009.

[7] X. Ma, G. Chen, and J. Xiao, “Analysis on an online health socialnetwork,’’ in Proc. 1st ACM Int. Health Inform. Symp., New York, NY, USA, pp. 297–306, 2010.

[8] A. Névéol and Z. Lu, “Automatic integration of drug indications from multiple health resources,” in *Proc. 1st ACM Int.HealthInform. Symp.*, New York, NY, USA, pp. 666--673, 2010.

[9] J. Leimeister, K. Schweizer, S. Leimeister, and H. Krcmar, “Do virtual communities matter for the social support of patients? Antecedents and effects of virtual relationships in online communities,” *Inform. Technol. People*, vol. 21, no. 4, pp. 350–374, 2008.

[10] R. Schraefel, R. White, P. André, and D. Tan, “Investigating web search strategies and forum use to support diet and weight loss,” in *Proc. 27th CHI EA*, New York, NY, USA, pp. 3829–3834, 2009.

[11] J. Zrebiec and A. Jacobson, “What attracts patients with diabetes to an internet support group? A 21-month longitudinal website stuey,” *Diabetic Med.*, vol. 18, no. 2, pp. 154–158, 2008.

[12] Bing Liu,”Web Data Mining”, liub@cs.uic.edu, Springer-Verlag Berlin Heidelberg 2007

[13] Victor C. Cheng, C.H.C. Leung, Jiming Liu, and Alfredo Milani, “Probabilistic Aspect Mining Model for Drug Reviews”, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 8, August 2014 .

[14] D. Blei and J. Lafferty, “Correlated topic models,” in *Proc. Adv. NIPS*, 2006, pp. 147–154.

[15] D. Mimno and A. McCallum, “Topic models conditioned on arbitrary features with Dirichlet-multinomial regression,” in *Proc.* 24th Conf. Uncertain. Artif. Intell, pp. 411--418, 2008.

[16] S. Lacoste-Julien, F. Sha, andM. Jordan, “DiscLDA: Discriminative learning for dimensionality reduction and classification,” in *Proc. Adv. NIPS*, pp. 897–904, 2008.

[17] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” in *Proc. 26th Annu. Int. ACM* SIGIR Conf. Res. Develop. Inform. Ret., New York, NY, USA, pp. 267–273, 2003.

[18] H. Lee, J. Yoo, and S. Choi, “Semi-supervised nonnegative matrix factorization”, *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 4–7, Jan. 2010.

[19] P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal, Lalit M. Patnaik. “Dynamic Association Rule Mining using Genetic Algorithms” Journal Intelligent Data Analysis, Volume 9, Number 5, pp. 439-453, 2005.

[20] Shenoy, P. Deepa, K. G. Srinivasa, K. R. Venugopal, and Lalit M. Patnaik. "Evolutionary approach for mining association rules on dynamic databases." In Advances in knowledge discovery and data mining,. Springer Berlin Heidelberg, pp. 325-336, 2003.

[21] Srinivasa, K. G., Karthik Sridharan, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. "A dynamic migration model for self-adaptive genetic algorithms." In Intelligent Data Engineering and Automated Learning-IDEAL 2005, Springer Berlin Heidelberg, pp. 555-562, 2005.

[22] Asha S Manek, Pallavi R P, Veena H Bhat, P Deepa Shenoy, Venugopal K R, M Chandramohan, L M Patnaik, “SentReP: Sentiment Classification of Movie Reviews using Efficient Repetitive Pre- Processing”, International Conference IEEE TENCON 2013, 22-25 October 2013, Xi’an China,, ISBN No. 978-1-4799-2825-5, pp. 1-5.

[23] Liu, Xiao, and Hsinchun Chen. “AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums” Smart Health. Springer Berlin Heidelberg, 2013. pp 134-150.

[24] Abeed Sarker , Graciela Gonzalez, “Portable automatic text classification for adverse drug reaction detection via multi-corpus training,” Journal of Biomedical Informatics 53 (2015) pp 196–207.