A Hybrid Model for Diagnosis of Heart Disease

Progress report

M.Tech Thesis Evaluation-4

May 2018

by

Sandyala Rajesh Varma

(2013IPG-095)

under the supervision of

Dr.P.K.Singh

विश्वजीवनामृतं ज्ञानम्

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR-474 010

2017-18

CANDIDATE'S DECLARATION

I hereby certify that I have properly checked and verified all the items as prescribed in the checklist and ensure that my thesis/report is in proper format as specified in the guideline for thesis preparation. I also declare that the work containing in this report is my own work. I understand that plagiarism is defined as any one or combination of the following:

- To steal and pass off (the ideas or words of another) as ones own
- To use (anothers production) without crediting the source
- To commit literary theft
- To present as new and original an idea or product derived from an existing source.

I understand that plagiarism involves an intentional act by the plagiarist of using someone elses work/ideas completely/partially and claiming authorship/originality of the work/ideas. Verbatim copy as well as close resemblance to some elses work constitute plagiarism. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programmes, experiments, results, websites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources. I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report/dissertation/thesis are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable. My faculty supervisor(s) will not be responsible for the same.

Signature:

Name :
Roll No :
Date :

ABSTRACT

Heart disease prediction is treated as most complicated task in the field of medical sciences. Heart disease is the dominant cause of death in the world over the past 10 years. Researchers have been utilizing several data mining techniques to assist health care professionals in the diagnosis of heart disease. Naive Bayes is one of the data mining techniques used in the diagnosis of heart disease showing ample success. Although researchers are investigating enhancing naive bayes performance, less research is done on enhancing naive bayes performance in the analysis of heart disease patients. K-means clustering is one of the most well known clustering techniques. However, initial centroid selection strongly influences its results. This paper demonstrates the effectiveness of an unsupervised learning technique such as k-means clustering with mean based initial centroid selection method in improving supervised learning technique which is naive bayes in the diagnosis of heart disease patients. The results show that naive bayes performance in diagnosing heart disease patients has been improved through integrating clustering with mean based initial centroid selection method as a pre-processing step to naive bayes classification.

Key words: K-Means Clustering, Naive Bayes, Mean Based Initial Centroid Selection Method, Heart Disease Diagnosis.

ACKNOWLEDGEMENTS

I am highly indebted to **Dr.P.K.Singh** and obliged for giving me the autonomy of functioning and experimenting with ideas. I would like to take the opportunity to express my profound gratitude to him not only for his academic guidance but also for his interest in my project and constant support coupled with confidence boosting and motivating sessions which proved very fruitful and were instrumental in infusing self-assurance and trust within me. Finally, I am grateful to all my friends, whose constant encouragement served to renew my spirit, refocus my attention and energy and helped me in carrying out this work.

Date :

Rajesh Varma   Signature of Supervisor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and Motivation

1.1 Introduction

Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. Naive bayes, Hybrid model of KMNB algorithm with different initial centroid selection methods such as Random, Inlier, Outlier, Range, Random attribute, Random row methods are used in the diagnosis of heart disease patients[1]. Naive Bayes is one of the successful data mining techniques used in the analysis of heart disease patients[2]. Despite of improving naive bayes efficiency in classification problems, a little work had been done on improving naive bayes efficiency for diagnosis of a heart disease[11]. This research investigates enhancing naive bayes performance in the diagnosis of heart disease patients through integrating clustering as a pre-processing step to naive bayes classification. K-means clustering is one of the most popular and well known clustering techniques. Its simplicity and good behaviour made it popular in many applications. Initial centroid selection is a critical issue in k-means clustering and strongly affects its results. This paper investigates integrating k-means clustering using mean based initial centroid selection method with naive bayes in the analysis of heart disease patients. The rest of the paper is divided as follows: the literature review section investigates applying data mining techniques in the diagnosis of heart disease; the methodology section explains k-means clustering, mean based initial centroid selection method and naive bayes used in the diagnosis of heart disease patients; the results section presents the results of integrating k-means clustering with naive bayes and the comparison of proposed algorithm performance with naive bayes, kmnb algorithm with different initial centroid selection methods; followed by the conclusion section.

1.2 Problem Statement

An approach for enhancing naive bayes classifier's performance by k-means clustering algorithm with initial centroid selection method in the diagnosis of heart disease patients. Given data sets with the attributes, instances for which the proposed algorithm has to be applied. Calculate the accuracies by comparing the output class labels with the original class labels of the input data set. Compare the accuracies of the proposed algorithm with the existing algorithms accuracies.

1.3 Objectives

- To enhance the naive bayes performance by integrating it with k-means clustering algorithm with initial centroid selection method in the diagnosis of heart disease patients.
- To implement proposed method on heart disease data set in order to make good decision about presence or absence of heart disease.

1.4 Report Layout

Information about various sections of this report are given here in Report Layout.

Chapter 2 explains the detailed research in choosing this particular topic. It shows how the literature survey has been done to analyze the topic.

Chapter 3 discusses about the methodology that has been applied in order to get the results.

Chapter 4 elaborates about the results and discussion.

Chapter 5 elaborates about the conclusion.

CHAPTER 2

Literature Review

Data mining gives the procedure and technology needed to convert huge amounts of data into helpful information for decision making. It is an effective procedure utilized to extract knowledge and find new patterns embedded in large data sets. Data mining has been progressively used in medicine, especially in cardiology[3]. In fact, Data Mining applications can significantly benefit each one of those associated with cardiology, for example, patients and cardiologists. K-means clustering is one of the prominent clustering techniques. Initial centroid selection is a critical issue that strongly affects its results. M.Goyal and S. Kumar have investigated the improvement of Initial Centroids of k-means Clustering Algorithm[4].Clustering of data points is done effectively by taking mean of the data points as initial centroids. The results showed that mean based k-means algorithm performs better than mid point based k-means and original k-means algorithms. Once the initial centroids are systematically determined by mean based k-means algorithm, the number of iterations required to reach the convergence criteria is reduced.

Palaniappan and Awang investigated comparing different data mining techniques in the diagnosis of heart disease patients. These techniques involved naive bayes, decision tree, and neural network. The results showed that the naive bayes could achieve the best accuracy in the diagnosis of heart disease patients[12].

Rajkumar and Reena investigated comparing naive bayes, k-nearest neighbour, and decision list in the diagnosis of heart disease patients. The results showed that the naive bayes could achieve the best accuracy in the diagnosis of heart disease patients[6].

Shouman, Turner and Stocker have investigated the integration of naive bayes and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. They have also investigated different methods of initial centroid selection of the K-means clustering such as range, inlier, outlier, random attribute, and random row methods in the diagnosis of heart disease patients. The results show that integrating k-means clustering with naive bayes with different initial centroid selection had enhanced the naive bayes accuracy in diagnosing heart disease patients. It also showed that the two clusters random row initial centroid selection method had achieved higher accuracy than other initial centroid selection methods in the diagnosis of heart disease.

CHAPTER 3

METHODOLOGY

The methodology section discusses k-means clustering with mean based initial centroid selection method. It also discusses about integrating k-means clustering and naive bayes classifier used in the diagnosis of heart disease patients.

3.1 Mean Based Initial Centroid Selection Method

Inspired from M. Goyal and S. Kumar's work in improving the initial centroids of k-means clustering algorithm[4] an approach to systematically choosing the initial centroids has been described here and integrated with naive bayes to improve the performance of naive bayes classifier. The centroids are decided following a systematic approach so that various runs of the algorithm on same dataset gives the same and good quality outcomes. All the data samples should have positive valued attributes. If not then the negative value attributes should first be converted to positive by subtracting each data sample attribute with the minimum attribute value in the given dataset. This conversion is required because in mean based initial centroid selection method, the distance of each data sample from the origin has to be evaluated. If data samples are not converted there is a chance that for different data samples, the same Euclidean distance from the origin is attained, which will result in incorrect selection of initial centroids.
The algorithm is as follows:

```
Algorithm 1 Mean based initial centroid selection method
Input: A dataset D containing n training data samples.
       $D = \{d_1, d_2.....d_n\}$
       Number of desired clusters k.
Output: k number of initial centroids.
1: Calculate the distance for each data sample in D from the origin.
2: Sort the distances obtained in the previous step. Sort the original data points in accordance with these distances.
3: Divide the sorted data samples into k number of equal partitions.
4: Calculate the mean of the data samples in each partition. These mean values will be considered as the initial centroids to be used in the k-means algorithm.
```

The following Euclidean distance measure has been used to calculate the distance for each data sample from the origin.

Origin: O (0,0)

Data point: D(x,y)

Euclidean distance between O-D will be: $\sqrt{(x-0)^2 + (y-0)^2}$

3.2 Integrating K-means and Naive bayes with Mean Based Initial Centroid Selection Method

K-means and Naive bayes are integrated with mean based initial centroid selection method to enhance the performance of naive bayes classifier in the analysis of heart disease patients. Several researchers have analyzed that blood pressure, age and cholesterol are critical risk factors related with heart disease[7, 8, 9, 10]. In identifying the attributes that will be utilized as a part of the clustering, these attributes are obvious clustering attributes for heart disease patients. The number of clusters used in the K- means in this investigation ranged between two and five clusters.

The process of integrating K-means algorithm with mean based initial centroid selection method and Naive bayes is as follows:

Algorithm 2 Integration of K-Means and Naive Bayes with Mean based initial centroid selection method

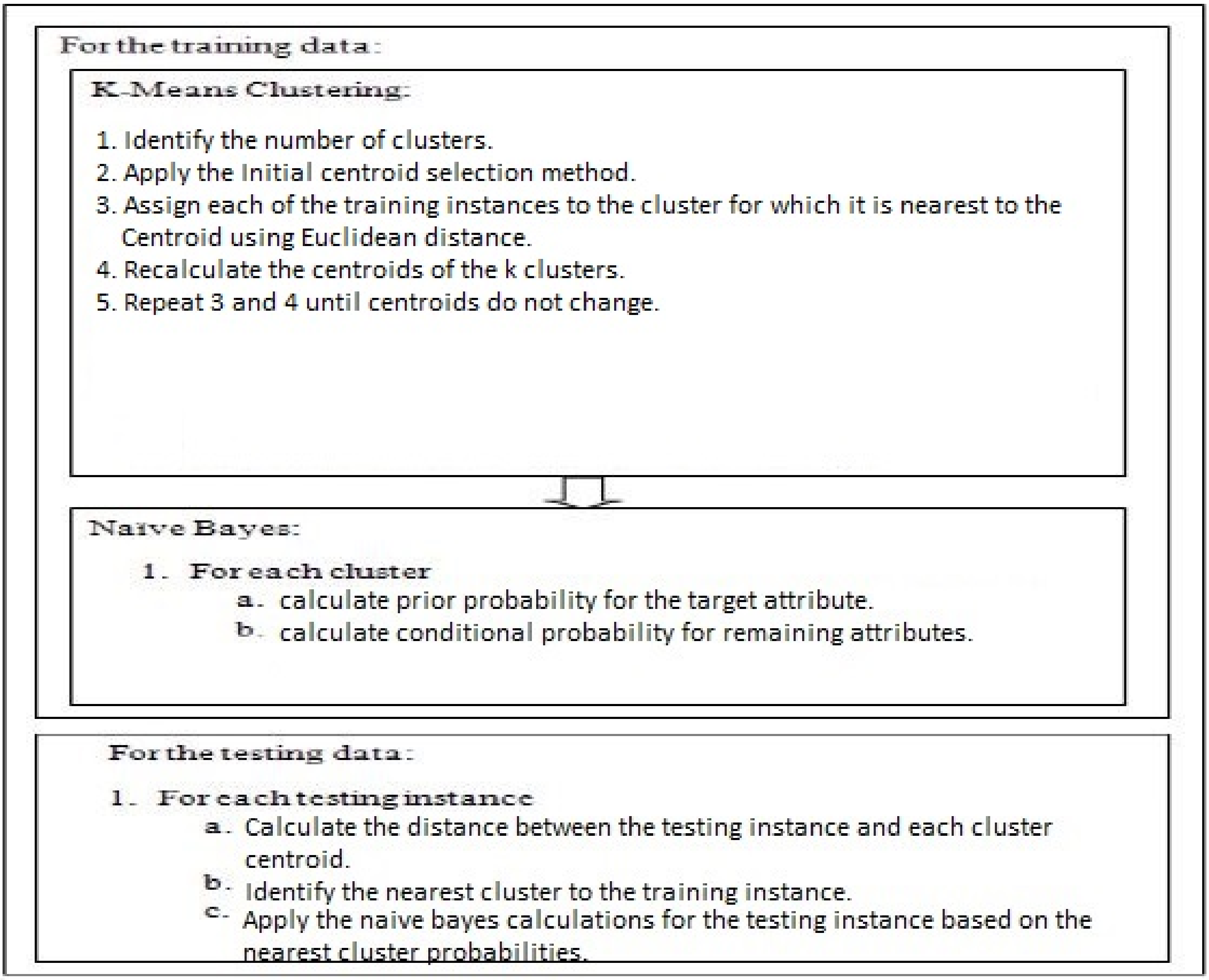
```
Input: A dataset D containing n samples.
    $D = \{d_1, d_2.....d_n\}$
Output: Class labels of testing data
1: Identify the number of clusters.
2: Apply the mean based initial centroid selection method.
3: Assign each of the training instance to the cluster for which it is nearest to
   the centroid using Euclidean distance.
4: Recalculate the centroids of the k clusters.
5: Repeat 4 and 5 until centroids do not change.
6: For each cluster, calculate prior probability for the target attribute and conditional probability for remaining attributes.
7: Calculate the distance between the testing instance and each cluster centroid.
8: Identify the nearest cluster to the testing instance.
9: Apply the naive bayes calculations for the testing instance based on the
   nearest cluster probabilities.
```

3.3 Naive Bayes

Naive bayes is one of the data mining techniques that show considerable success in classification problems and specially in diagnosing heart disease patients[6]. Naive bayes is based on probability theory to find the most likely possible classifications. For a given set of training samples with class labels and a test instance T represented by n attribute values (a1, a2... an), naive Bayesian classifier uses the following equation to classify T:

$$\mathrm{c}_{NB}(T) = arg_{c \in C} max P(c) \prod_{i=1}^{n} P(a_i|c)(3.1)$$

Where, $c_{NB}(T)$ represents the classification given by NB on test instance T[5]. It is based on prior probability of the target attribute and the conditional probability of the remaining attributes. For the training data the prior and conditional probability are calculated for each cluster. For each testing instance in the testing data, the probability is calculated with each of the target attribute values and the target attribute value with the largest probability is then selected.



For the training data:

K-Means Clustering:

1. Identify the number of clusters.
2. Apply the Initial centroid selection method.
3. Assign each of the training instances to the cluster for which it is nearest to the Centroid using Euclidean distance.
4. Recalculate the centroids of the k clusters.
5. Repeat 3 and 4 until centroids do not change.

Naive Bayes:

1. For each cluster
   a. calculate prior probability for the target attribute.
   b. calculate conditional probability for remaining attributes.

For the testing data:

1. For each testing instance
   a. Calculate the distance between the testing instance and each cluster centroid.
   b. Identify the nearest cluster to the training instance.
   c. Apply the naive bayes calculations for the testing instance based on the nearest cluster probabilities.

Figure 3.1: Integrating K-means and Naive bayes with Mean Based Initial Centroid Selection Method

CHAPTER 4

Results and Discussion

4.1 Computation Environment

To implement Naive bayes and other hybrid models of KMNB algorithm, we have used python language in spyder ide. 10-fold cross-validation is being utilized to calculate the accuracies of different algorithms. Original Sample is partitioned into 10 equal size sub samples. Out of these 10 subsamples one subsample is used for Testing and other 9 subsamples are used as Training data. This cross-validation process is repeated 10 times, with each of the subsample used exactly once as the validation data. All accuracies are acquired by taking the average of the outcomes from 10 executions of 10-fold cross-validation.

4.2 Data Sets

Statlog heart disease data set and Cleveland Heart disease data set are used in this study which are available at UCI Machine Learning Repository. The Statlog data set contains 13 attributes and 270 rows. The Cleveland data set contains 13 attributes and 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment. Statlog-Cleveland data set which is formed by integrating both Statlog and Cleveland data sets contain 13 attributes and 567 rows.

4.3 Performance measures

Accuracy, Sensitivity and Specificity are considered as performance measures for the evaluation of the execution of naive bayes and other hybrid models of KMNB algorithm. Sensitivity is measured as the proportion of positive instances that are correctly classified as positive (example: proportion of sick

people that are classified as sick). Specificity is measured as the proportion of negative instances that are correctly classified as negative (example: proportion of healthy people that are classified as healthy). Accuracy is measured as the proportion of instances that are correctly classified.
Sensitivity = True Positive / Positive
Specificity = True Negative / Negative
Accuracy = (True Positive + True Negative) / (Positive + Negative)

4.4 Results for Statlog Data

Proposed KMNB algorithm with mean based initial centroid selection method outperforms all other algorithms and equally performs with the random attribute method of kmnb algorithm. Proposed algorithm performs better with two number of clusters for this data set, So it's performance is compared with other hybrid models of kmnb algorithm with two number of clusters. Accuracy, Sensitivity and Specificity comparison of all algorithms for this data set is shown in Table 4.1. For the representation of table as plots refer to Figure 4.1, 4.2, 4.3.

Table 4.1: Table representing Performance measures comparison for data set Statlog

Algorithm	Accuracy(%)	Sensitivity(%)	Specificity(%)
NB	75.55	75	20
KMNB(Random)	75.55	67.4	16.66
KMNB(Inlier)	77.03	72.09	18.18
KMNB(Outlier)	76.66	68.18	16.36
KMNB(Range)	76.66	68.18	16.36
KMNB(Random attribute)	78.51	72.09	14.54
KMNB(Random row)	77.77	70.45	14.81
KMNB(Proposed)	78.51	72.72	16.36

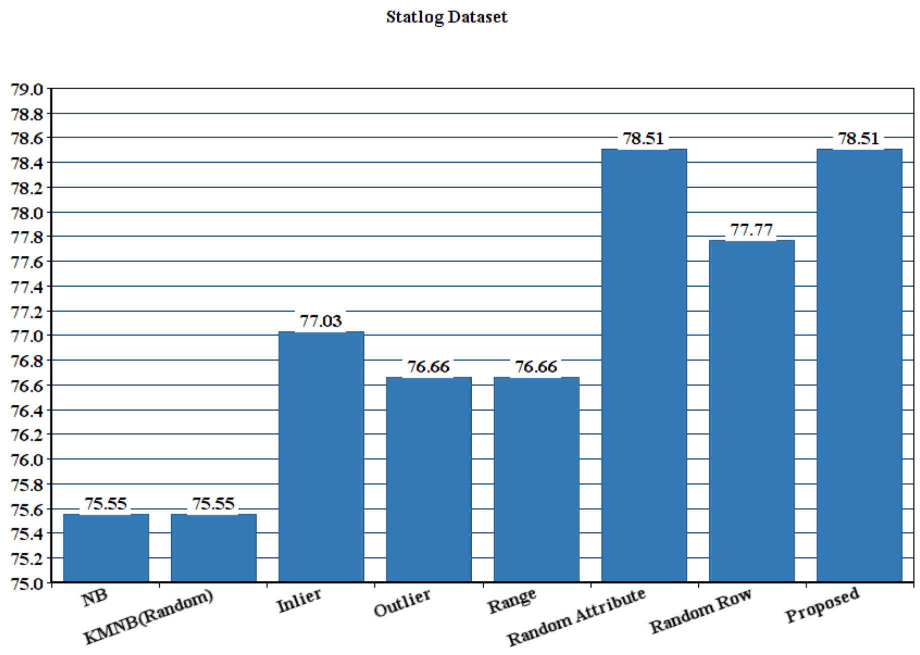


Figure 4.1: Graph representing accuracy comparison for Statlog data set

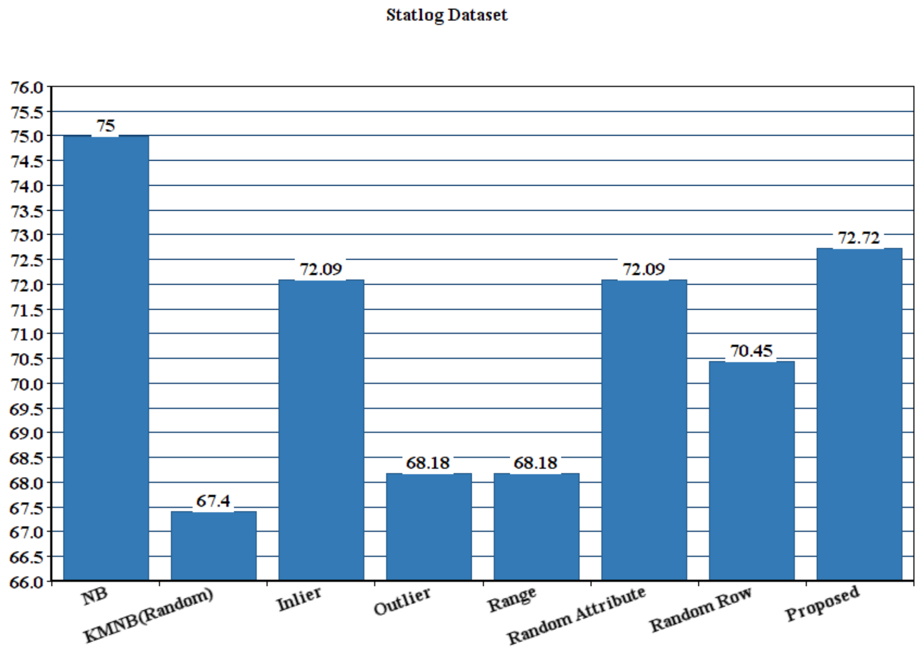


Figure 4.2: Graph representing sensitivity comparison for Statlog data set

4.5 Results for Cleveland Data

Proposed KMNB algorithm with mean based initial centroid selection method outperforms all other algorithms. Proposed algorithm performs better with three number of clusters for this data set, So it's performance is compared with other

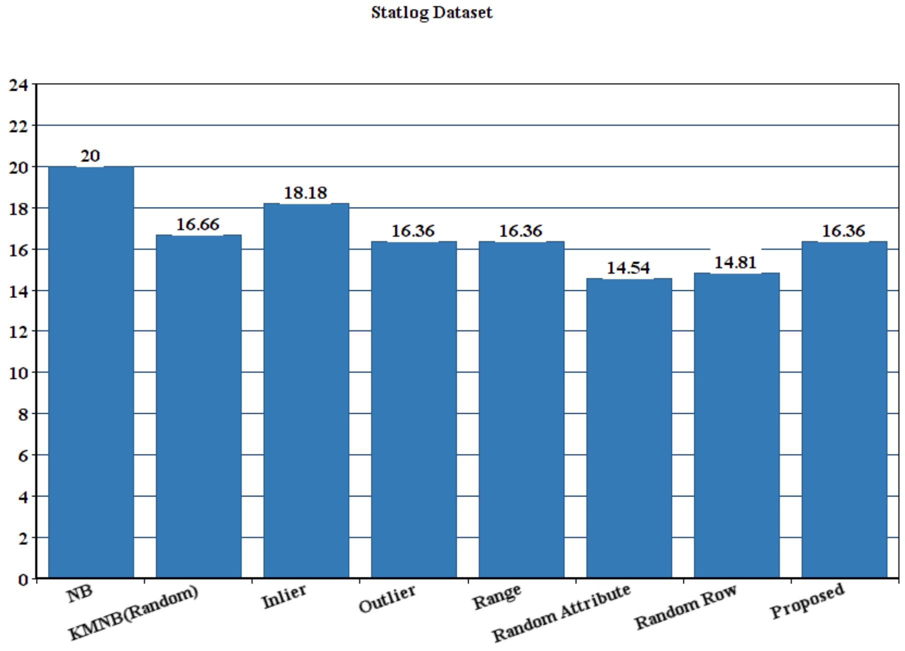


Figure 4.3: Graph representing specificity comparison for Statlog data set

hybrid models of kmnb algorithm with three number of clusters. The accuracy comparison of all algorithms for this data set is shown in Table 4.2. For the representation of table as plots refer to Figure 4.4, 4.5, 4.6.

Table 4.2: Table representing accuracy comparison for data set Cleveland

Algorithm	Accuracy(%)	Sensitivity(%)	Specificity(%)
NB	77.10	73.33	16.98
KMNB(Random)	79.43	75.55	16.98
KMNB(Inlier)	78.09	75.55	18.86
KMNB(Outlier)	79.44	75.55	16.98
KMNB(Range)	79.77	76.08	16.98
KMNB(Random attribute)	78.78	77.77	16.98
KMNB(Random row)	79.12	77.77	16.98
KMNB(Proposed)	80.10	76.08	15.09

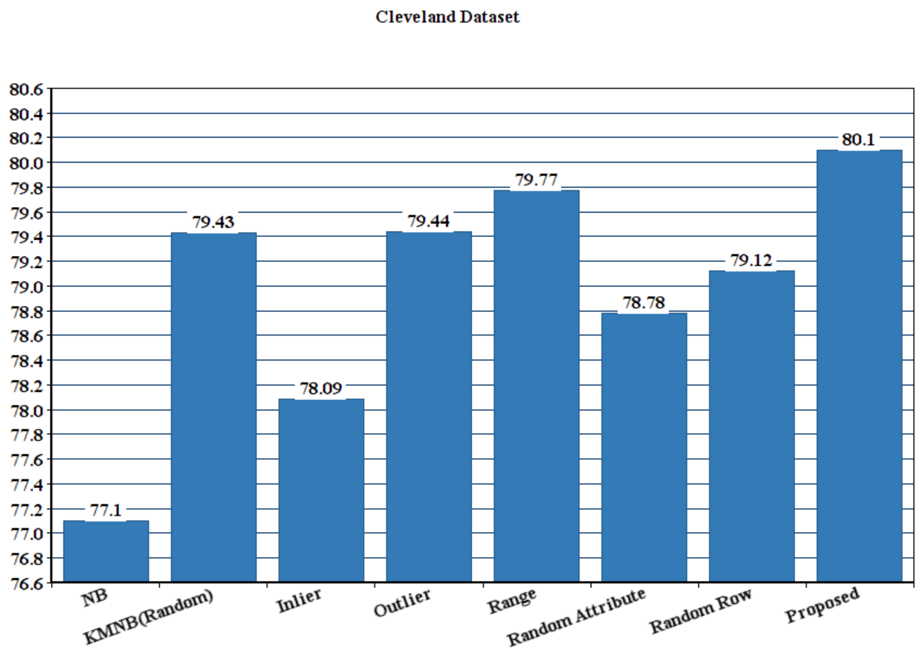


Figure 4.4: Graph representing accuracy comparison for Cleveland data set

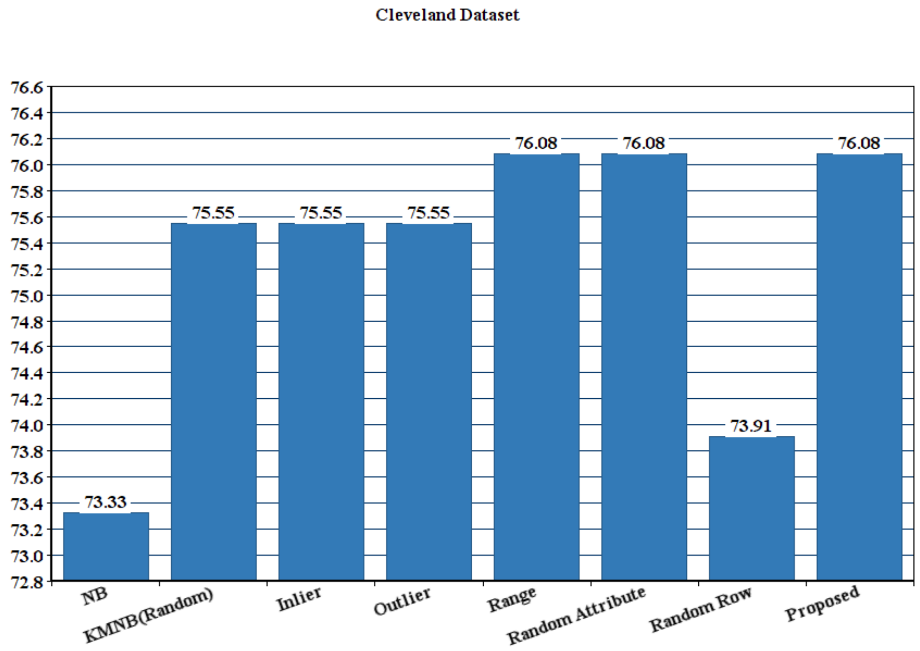


Figure 4.5: Graph representing sensitivity comparison for Cleveland data set

4.6 Results for Statlog-Cleveland Data

Proposed KMNB algorithm with mean based initial centroid selection method outperforms all other algorithms and equally performs with the outlier method of kmnb algorithm. Proposed algorithm performs better with four number of clus-

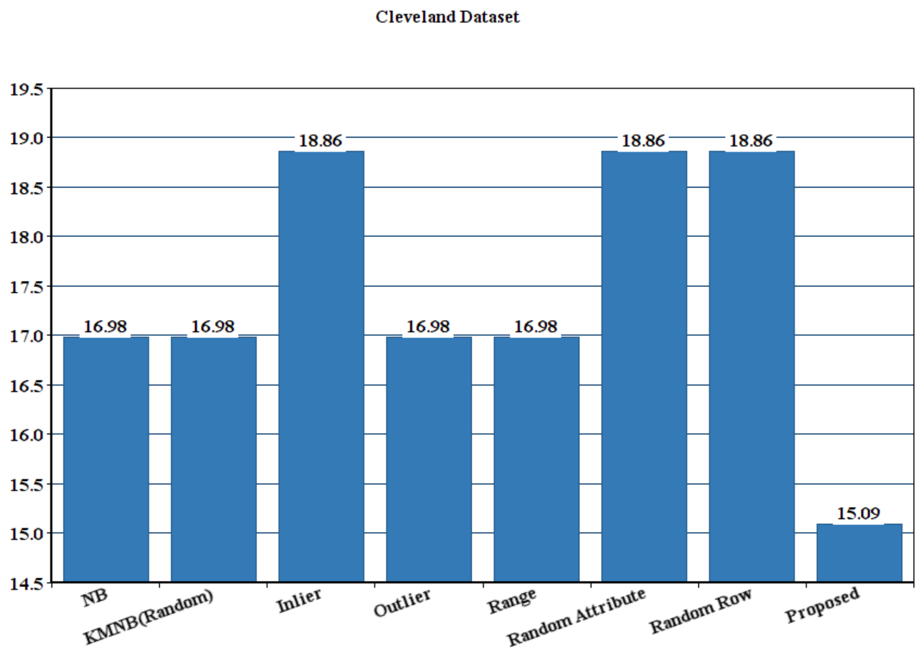


Figure 4.6: Graph representing specificity comparison for Cleveland data set

ters for this data set, So it's performance is compared with other hybrid models of kmnb algorithm with four number of clusters. The accuracy comparison of all algorithms for this data set is shown in Table 4.3. For the representation of table as plots refer to Figure 4.7, 4.8, 4.9.

Table 4.3: Table representing accuracy comparison for data set Statlog-Cleveland

Algorithm	Accuracy(%)	Sensitivity(%)	Specificity(%)
NB	77.07	73.33	20.37
KMNB(Random)	79.71	75	15.09
KMNB(Inlier)	80.60	75.55	14.81
KMNB(Outlier)	81.1	75.55	13.2
KMNB(Range)	80.07	75	14.81
KMNB(Random attribute)	80.41	75	16.66
KMNB(Random row)	80.24	75.55	15.09
KMNB(Proposed)	81.1	77.27	14.81

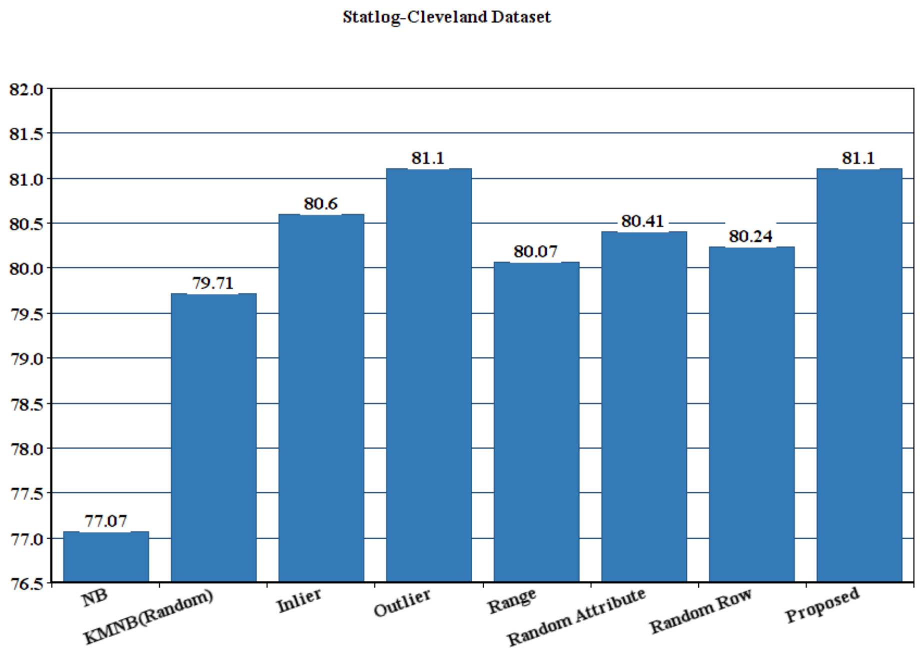


Figure 4.7: Graph representing accuracy comparison for Statlog-Cleveland data set

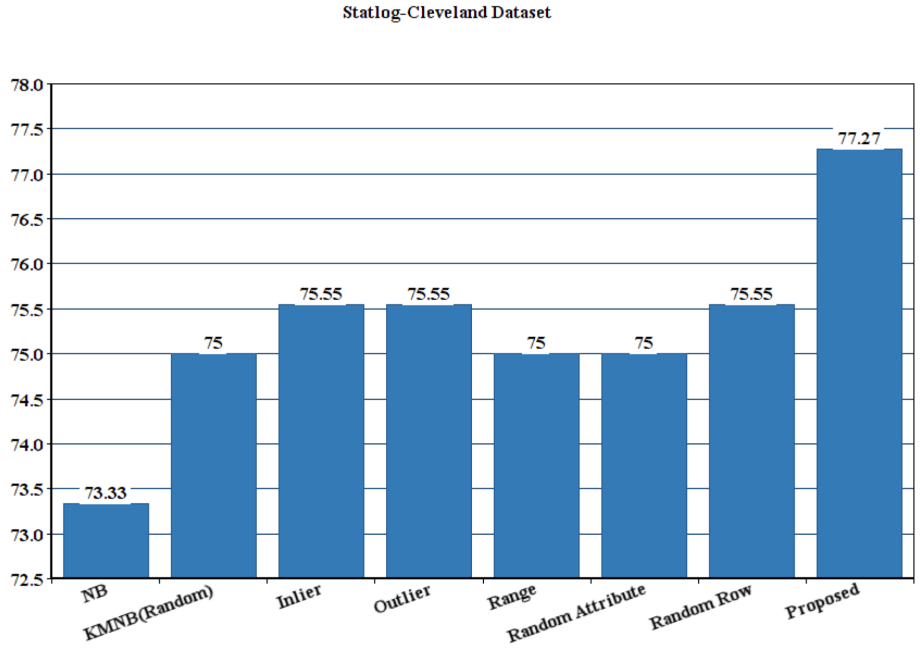


Figure 4.8: Graph representing sensitivity comparison for Statlog-Cleveland data set

4.7 Results

Proposed KMNB algorithm with mean based initial centroid selection method, NB, KMNB(Random) algorithm and other existing hybrid algorithms of KMNB

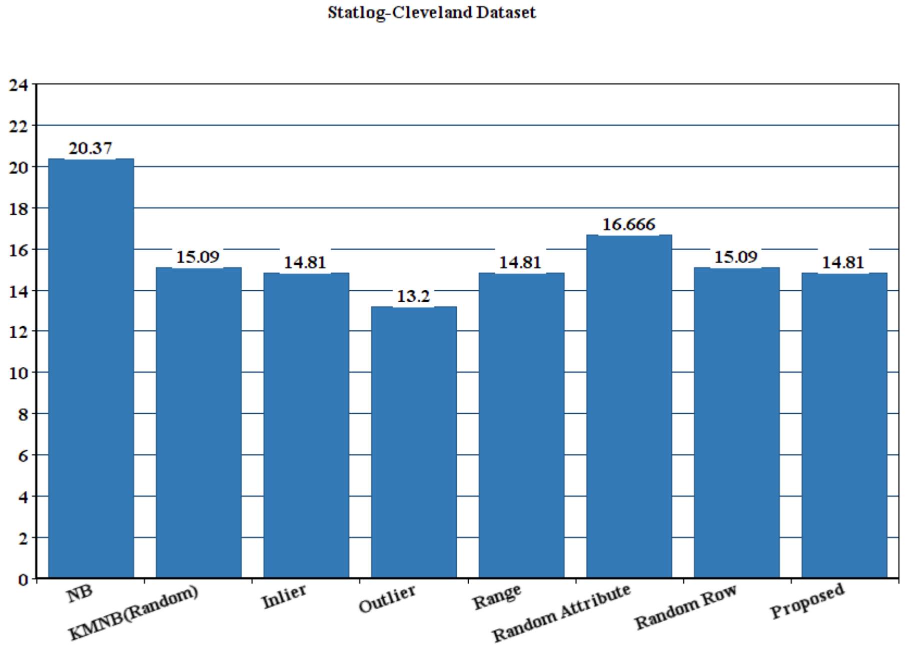
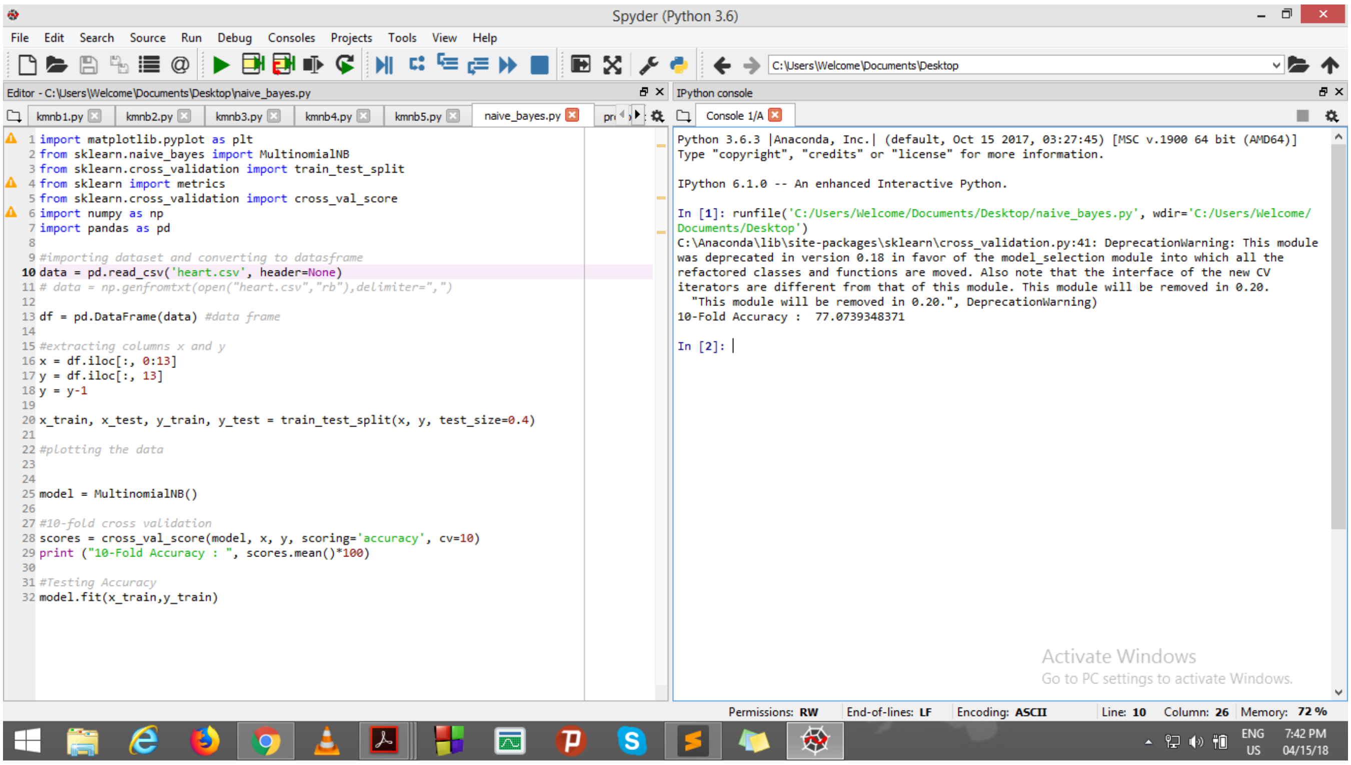


Figure 4.9: Graph representing specificity comparison for Statlog-Cleveland data set

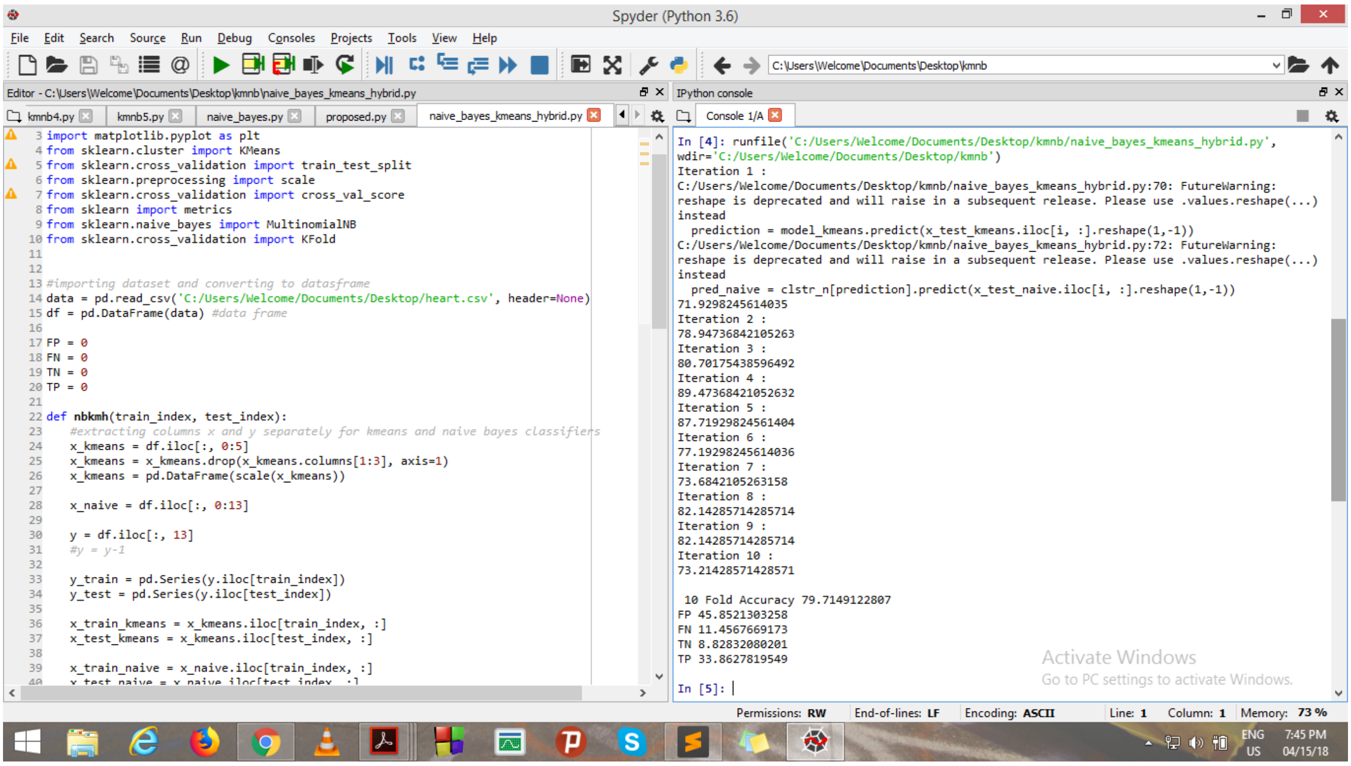
are applied on 3 datasets and on the basis of experimental analysis, predictive accuracy outcomes associated with each data set are presented in Tables. However, by comparing the execution of Proposed algorithm with NB, KMNB(Random) algorithm and other hybrid algorithms of KMNB the attained outcomes prove that Proposed algorithm outperforms NB, KMNB(random), KMNB(Inlier), KMNB(Range), KMNB(Random Row) algorithms in 3 datasets and it outperforms KMNB(Random Attribute), KMNB(Outlier) algorithms in 2 datasets, equally performs KMNB(Random Attribute), KMNB(Outlier) algorithms in 1 dataset. 10-fold cross-validation is being utilized to calculate the accuracies of different algorithms. Original Sample is partitioned into 10 equal size subsamples. Out of these 10 subsamples one subsample is used for Testing and other 9 subsamples are used as Training data. This cross-validation process is repeated 10 times, with each of the subsample used exactly once as the validation data. All accuracies are acquired by taking the average of the outcomes from 10 executions of 10-fold cross-validation. From Table 1, 2 and 3 it is observed that proposed algorithm accuracy is more reliable when we have more number of instances in a data set.



```python
import matplotlib.pyplot as plt
from sklearn.naive_bayes import MultinomialNB
from sklearn.cross_validation import train_test_split
from sklearn import metrics
from sklearn.cross_validation import cross_val_score
import numpy as np
import pandas as pd

#importing dataset and converting to datasframe
data = pd.read_csv('heart.csv', header=None)
# data = np.genfromtxt(open("heart.csv","rb"),delimiter=",")

df = pd.DataFrame(data) #data frame

#extracting columns x and y
x = df.iloc[:, 0:13]
y = df.iloc[:, 13]
y = y-1

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.4)

#plotting the data


model = MultinomialNB()

#10-fold cross validation
scores = cross_val_score(model, x, y, scoring='accuracy', cv=10)
print ("10-Fold Accuracy : ", scores.mean()*100)

#Testing Accuracy
model.fit(x_train,y_train)
```

```
Python 3.6.3 |Anaconda, Inc.| (default, Oct 15 2017, 03:27:45) [MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.1.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/Welcome/Documents/Desktop/naive_bayes.py', wdir='C:/Users/Welcome/
Documents/Desktop')
C:\Anaconda\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWarning: This module
was deprecated in version 0.18 in favor of the model_selection module into which all the
refactored classes and functions are moved. Also note that the interface of the new CV
iterators are different from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
10-Fold Accuracy :  77.0739348371

In [2]: 
```

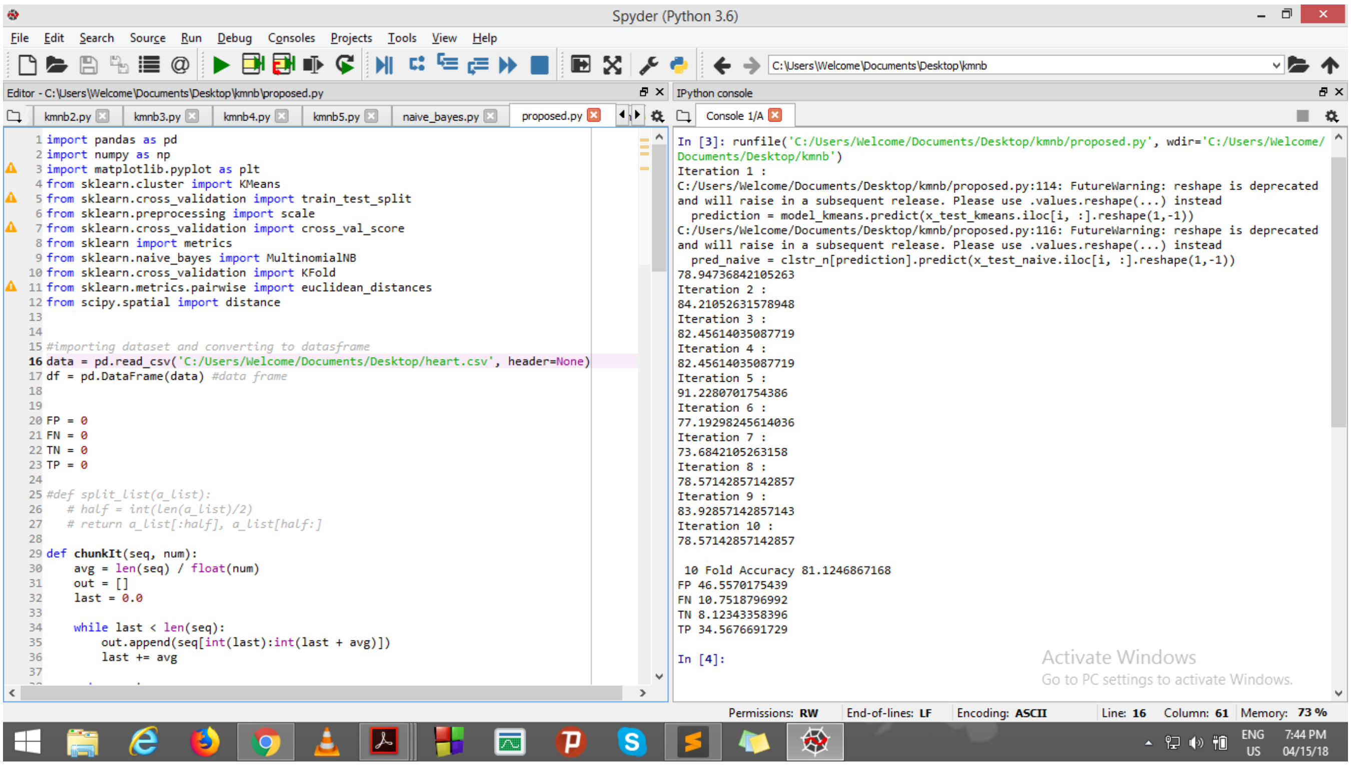
Figure 4.10: Figure depicting implementation of naive bayes algorithm


Spyder (Python 3.6)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Welcome\Documents\Desktop\kmnb
Editor - C:\Users\Welcome\Documents\Desktop\kmnb\naive_bayes_kmeans_hybrid.py
kmnb4.py
kmnb5.py
naive_bayes.py
proposed.py
naive_bayes_kmeans_hybrid.py
IPython console
Console 1/A
Activate Windows
Go to PC settings to activate Windows.
Permissions: RW
End-of-lines: LF
Encoding: ASCII
Line: 1
Column: 1
Memory: 73 %
ENG US
7:45 PM 04/15/18

```python
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.cross_validation import train_test_split
from sklearn.preprocessing import scale
from sklearn.cross_validation import cross_val_score
from sklearn import metrics
from sklearn.naive_bayes import MultinomialNB
from sklearn.cross_validation import KFold


#importing dataset and converting to datasframe
data = pd.read_csv('C:/Users/Welcome/Documents/Desktop/heart.csv', header=None)
df = pd.DataFrame(data) #data frame

FP = 0
FN = 0
TN = 0
TP = 0

def nbkmh(train_index, test_index):
    #extracting columns x and y separately for kmeans and naive bayes classifiers
    x_kmeans = df.iloc[:, 0:5]
    x_kmeans = x_kmeans.drop(x_kmeans.columns[1:3], axis=1)
    x_kmeans = pd.DataFrame(scale(x_kmeans))

    x_naive = df.iloc[:, 0:13]

    y = df.iloc[:, 13]
    #y = y-1

    y_train = pd.Series(y.iloc[train_index])
    y_test = pd.Series(y.iloc[test_index])

    x_train_kmeans = x_kmeans.iloc[train_index, :]
    x_test_kmeans = x_kmeans.iloc[test_index, :]

    x_train_naive = x_naive.iloc[train_index, :]
    x_test_naive = x_naive.iloc[test_index, :]
```

```
In [4]: runfile('C:/Users/Welcome/Documents/Desktop/kmnb/naive_bayes_kmeans_hybrid.py',
wdir='C:/Users/Welcome/Documents/Desktop/kmnb')
Iteration 1 :
C:/Users/Welcome/Documents/Desktop/kmnb/naive_bayes_kmeans_hybrid.py:70: FutureWarning:
reshape is deprecated and will raise in a subsequent release. Please use .values.reshape(...)
instead
  prediction = model_kmeans.predict(x_test_kmeans.iloc[i, :].reshape(1,-1))
C:/Users/Welcome/Documents/Desktop/kmnb/naive_bayes_kmeans_hybrid.py:72: FutureWarning:
reshape is deprecated and will raise in a subsequent release. Please use .values.reshape(...)
instead
  pred_naive = clstr_n[prediction].predict(x_test_naive.iloc[i, :].reshape(1,-1))
71.9298245614035
Iteration 2 :
78.94736842105263
Iteration 3 :
80.70175438596492
Iteration 4 :
89.47368421052632
Iteration 5 :
87.71929824561404
Iteration 6 :
77.19298245614036
Iteration 7 :
73.6842105263158
Iteration 8 :
82.14285714285714
Iteration 9 :
82.14285714285714
Iteration 10 :
73.21428571428571

 10 Fold Accuracy 79.7149122807
FP 45.8521303258
FN 11.4567669173
TN 8.82832080201
TP 33.8627819549

In [5]: 
```

Figure 4.11: Figure depicting implementation of traditional kmnb algorithm



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.cross_validation import train_test_split
from sklearn.preprocessing import scale
from sklearn.cross_validation import cross_val_score
from sklearn import metrics
from sklearn.naive_bayes import MultinomialNB
from sklearn.cross_validation import KFold
from sklearn.metrics.pairwise import euclidean_distances
from scipy.spatial import distance


#importing dataset and converting to datasframe
data = pd.read_csv('C:/Users/Welcome/Documents/Desktop/heart.csv', header=None)
df = pd.DataFrame(data) #data frame


FP = 0
FN = 0
TN = 0
TP = 0

#def split_list(a_list):
   # half = int(len(a_list)/2)
   # return a_list[:half], a_list[half:]

def chunkIt(seq, num):
    avg = len(seq) / float(num)
    out = []
    last = 0.0

    while last < len(seq):
        out.append(seq[int(last):int(last + avg)])
        last += avg

```

```
In [3]: runfile('C:/Users/Welcome/Documents/Desktop/kmnb/proposed.py', wdir='C:/Users/Welcome/
Documents/Desktop/kmnb')
Iteration 1 :
C:/Users/Welcome/Documents/Desktop/kmnb/proposed.py:114: FutureWarning: reshape is deprecated
and will raise in a subsequent release. Please use .values.reshape(...) instead
  prediction = model_kmeans.predict(x_test_kmeans.iloc[i, :].reshape(1,-1))
C:/Users/Welcome/Documents/Desktop/kmnb/proposed.py:116: FutureWarning: reshape is deprecated
and will raise in a subsequent release. Please use .values.reshape(...) instead
  pred_naive = clstr_n[prediction].predict(x_test_naive.iloc[i, :].reshape(1,-1))
78.94736842105263
Iteration 2 :
84.21052631578948
Iteration 3 :
82.45614035087719
Iteration 4 :
82.45614035087719
Iteration 5 :
91.2280701754386
Iteration 6 :
77.19298245614036
Iteration 7 :
73.6842105263158
Iteration 8 :
78.57142857142857
Iteration 9 :
83.92857142857143
Iteration 10 :
78.57142857142857

 10 Fold Accuracy 81.1246867168
FP 46.5570175439
FN 10.7518796992
TN 8.12343358396
TP 34.5676691729

In [4]:
```

Figure 4.12: Figure depicting implementation of proposed method

CHAPTER 5

CONCLUSION

5.1 Conclusion

Naive bayes performance in the diagnosis of heart disease patients has been improved through integrating clustering with proposed initial centroid selection method as a pre-processing step to naive bayes classification. Proposed algorithm outperforms NB, KMNB(random) and all other hybrid algorithms of KMNB when applied on real heart disease data sets. With respect to obtained results, We conclude that the utilization of proposed initial centroid selection method in the integration of k-means algorithm and naive bayes algorithm gives better results for the diagnosis of heart disease patients when compared with NB algorithm, KMNB(Random) algorithm, and other hybrid algorithms of KMNB. The KMNB algorithm can be applied on numerical data only. But in everyday life, situations with a combination of both numerical and categorical data values are encounterd. So future work will be carried out in the direction of making the KMNB algorithm suitable for mixed type of data.

REFERENCES

[1] Shouman, Mai, Tim Turner, and Rob Stocker. "Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients." CS IT-CSCP (2012): 125-137.

[2] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2.02 (2010): 250-255.

[3] Kadi, Ilham, Ali Idri, and J. L. Fernandez-Aleman. "Knowledge discovery in cardiology: A systematic literature review." International journal of medical informatics 97 (2017): 12-32.

[4] Goyal, M., and S. Kumar. "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability." Journal of The Institution of Engineers (India): Series B 95.4 (2014): 345-350.

[5] Al-Aidaroos, Khadija, Azuraliza Abu Bakar, and Zalinda Othman. "Data classification using rough sets and naïve Bayes." International Conference on Rough Sets and Knowledge Technology. Springer, Berlin, Heidelberg, 2010.

[6] Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." Global journal of computer science and technology 10.10 (2010): 38-43.

[7] Altayeva, Aigerim, Suleimenov Zharas, and Young Im Cho. "Medical decision making diagnosis system integrating k-means and Naïve Bayes algorithms." Control, Automation and Systems (ICCAS), 2016 16th International Conference on. IEEE, 2016.

[8] Shahwan-Akl, Lina. "Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne." International Journal of Research in Nursing 6.1 (2010): 1-7.

[9] Heller, R. F., et al. "How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project." Br Med J (Clin Res Ed) 288.6428 (1984): 1409-1411.

[10] Din, Salahud, and Fazle Rabbi. "Statistical analysis of risk factors for cardiovascular disease in malakand division." Pakistan journal of statistics and operation research 2.1 (2006).

[11] Ratanamahatana, Chotirat Ann, and Dimitrios Gunopulos. "Scaling up the naive Bayesian classifier: Using decision trees for feature selection." (2002).

[12] Palaniappan, Sellappan, and Rafiah Awang. "Web-based Heart Disease Decision Support System Using Data Mining Classification Modeling Techniques." iiWAS. 2007