Prediction of Heart Disease

BY

VARA PRASAD SHEELA

UIN: 00764783

AND

MAHESH CHEETIRALA

UIN: 01103978

ABSTRACT

Heart disease is one of the major medical problems and lots of people are losing their lives because of one or other form of heart disease whether it might be a Cardiovascular disease, Hypertension, Heart arrhythmia or a Heart failure. In the current world there is a lot of research going on in prediction of heart disease so that it can be prevented at earlier stages. UCI laboratory has a repository of heart disease data where researches can get the information and further their studies. On this note we have decided to work on the data provided by the UCI laboratory and on exploring the research that was already conducted we were interested in

a) Why is that all the research that was conducted always revolved around the 13 parameters that were mentioned in the UCI repository and with out any reference.

b) Why researchers always used Cleveland data in most of the cases or all the data from all the 4 sources mentioned. These "WHY" questions have provoked us to get the more insights.

In our work we have used an Analytics tool from Parametric Technology Corporation (PTC) for the feature selection using Cleveland and Hungarian datasets. Based on the mutual information (0.04) we have picked 15 parameters for Cleveland dataset and mutual information (0.05) for Hungarian dataset we have picked 11 parameters for prediction of heart disease. For further analysis we have used J48, Naive Bayes and Random Forest classifiers from Weka 3.8 tool with a 10 fold cross validation. From our experimental results we were able to conclude that Naive Bayes algorithm gives us accurate prediction both in cleveland and hungarian datasets. Our results have also confirmed that due to missing values in some of the important parameters Hungarian dataset might have not been used in the literature for any analysis.

TABLE OF CONTENTS

LIST OF TABLES AND IMAGES

CHAPTER 1
INTRODUCTION

1.1 Problem Statement:

Heart disease is one of the major reasons for the death of the people [1]. Research states that at least 1 death is caused due to heart disease for every minute in United States alone. According to CDC [2] more than 630,000 Americans die from heart disease each year that's 1 in every 4 deaths. There are plethora of studies out there using various techniques predicting heart disease [3],[4],[5],[6], but still there is an ample scope for research due to evolving heart diseases.

1.2 Goals:

To our knowledge all the studies that were published was always confined to cleveland database [7],[8] of UCI repository with 303 instances and 14 different attributes out of 76 attributes or a mix of all the datasets that are available in the UCI machine learning repository [9]. In this project besides predicting the heart disease using feature selection and implementing a classification algorithm we will be using both hungarian dataset and cleveland dataset separately and besides using the well known 14 attributes we will be working on exploring other attributes that might be the cause for the heart disease which will be a novel research than the published work [3],[4],[5],[6].

1.3 Database and Datasets:

We have used UCI machine learning repository [10] which consists of heart disease datasets from 4 different sources.

1) Cleveland Clinic Foundation (cleveland.data)
2) Hungarian Institute of Cardiology, Budapest (hungarian.data)
3) V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4) University Hospital, Zurich, Switzerland (switzerland.data)

The above mentioned datasets contains 76 attributes each with missing values. Below are the various attributes that were defined in the datasets.

Complete attribute documentation:[10]

1 id: patient identification number

2 ccf: social security number (I replaced this with a dummy value of 0)

3 age: age in years

4 sex: sex (1 = male; 0 = female)

5 painloc: chest pain location (1 = substernal; 0 = otherwise)

6 painexer (1 = provoked by exertion; 0 = otherwise)

7 relrest (1 = relieved after rest; 0 = otherwise)

8 pncaden (sum of 5, 6, and 7)

9 cp: chest pain type

-- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain

-- Value 4: asymptomatic

10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)

11 htn

12 chol: serum cholesterol in mg/dl

13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)

14 cigs (cigarettes per day)

15 years (number of years as a smoker)

16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

17 dm (1 = history of diabetes; 0 = no such history)

18 famhist: family history of coronary artery disease (1 = yes; 0 = no)

19 restecg: resting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

20 ekgmo (month of exercise ECG reading)

21 ekgday(day of exercise ECG reading)

22 ekgyr (year of exercise ECG reading)

23 dig (digitalis used during exercise ECG: 1 = yes; 0 = no)

24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)

25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)

26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)

27 diuretic (diuretic used during exercise ECG: 1 = yes; 0 = no)

28 proto: exercise protocol

1 = Bruce 2 = Kottus 3 = McHenry 4 = fast Balke 5 = Balke 6 = Naughton

7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)

8 = bike 125 kpa min/min 9 = bike 100 kpa min/min 10 = bike 75 kpa min/min

11 = bike 50 kpa min/min 12 = arm ergometer

29 thaldur: duration of exercise test in minutes

30 thaltime: time when ST measure depression was noted

31 met: mets achieved

32 thalach: maximum heart rate achieved

33 thalrest: resting heart rate

34 tpeakbps: peak exercise blood pressure (first of 2 parts)

35 tpeakbpd: peak exercise blood pressure (second of 2 parts)

36 dummy

37 trestbpd: resting blood pressure

38 exang: exercise induced angina (1 = yes; 0 = no)

39 xhypo: (1 = yes; 0 = no)

40 oldpeak = ST depression induced by exercise relative to rest

41 slope: the slope of the peak exercise ST segment

-- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping 42 rldv5: height at rest

43 rldv5e: height at peak exercise

44 ca: number of major vessels (0-3) colored by fluoroscopy

45 restckm: irrelevant

46 exerckm: irrelevant

47 restef: rest radionuclide (sp?) ejection fraction

48 restwm: rest wall (sp?) motion abnormality

0 = none 1 = mild or moderate 2 = moderate or severe 3 = akinesis or dyskmem (sp?)

49 exeref: exercise radinalid (sp?) ejection fraction

50 exerwm: exercise wall (sp?) motion

51 thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

52 thalsev: not used

53 thalpul: not used

54 earlobe: not used

55 cmo: month of cardiac cath (sp?) (perhaps "call")

56 cday: day of cardiac cath (sp?)

57 cyr: year of cardiac cath (sp?)

58 num: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing

(in any major vessel: attributes 59 through 68 are vessels)

59 lmt

60 ladprox

61 laddist

62 diag

63 cxmain

64 ramus

65 om1

66 om2

67 rcaprox

68 rcadist

69 lvx1: not used

70 lvx2: not used

71 lvx3: not used

72 lvx4: not used

73 lvf: not used

74 cathef: not used

75 junk: not used

76 name: last name of patient (I replaced this with the dummy string "name")

Table 1. Four datasets with number of instances.

Dataset	**Number of Instances**
Cleveland dataset	303
Hungarian dataset	294
Switzerland dataset	123
VA Long Beach dataset	200

CHAPTER 2
METHODOLOGY

2.1 Data cleansing/Preprocessing:

Data from UCI repository is not in a usable format and most importantly this data has span into multiple lines with space separation and each record ending with a value “name” in the last name field of the patient. Hence it is strongly recommended that we need to clean and process the data into a usable format. To achieve this we have took a step to convert the data into csv format and import it into pandas dataframe in python library.

We have first copied the data from UCI repository into different text files and called the text files inside python code and have used some of the inbuilt functions in python to clean the data. During this process we have also encountered a situation where the dataset from cleveland data is corrupt and has to delete some of the rows. We have also observed that the “NULL” is represented by “-9” which we have changed it to “NaN”.

Although we have decided to use all the 76 attributes and predict the most important ones (unlike 14 attributes that were described in the literature [3],[4],[5],[6]) we have to remove some of the attributes such as ID of the patient, ccf(social security number), name(last name) etc., which will not be relevant in predicting the heart disease. Besides these fields it has also been reported in the dataset that some of the columns were irrelevant/not used hence we have decided to remove those columns as well. Besides the above mentioned there were columns from 59 through 68 which is about blood vessels which will not be used in predicting the heart disease hence those columns were deleted. Below is the list of columns that we have decided to remove based on the above criteria.

1 id: patient identification number
2 ccf: social security number (I replaced this with a dummy value of 0)
3 ekgmo (month of exercise ECG reading)
4 ekgday(day of exercise ECG reading)
5 ekgyr (year of exercise ECG reading)

6 proto: exercise protocol

7 dummy

8 restckm: irrelevant

9 exerckm: irrelevant

10 thalsev: not used

11 thalpul: not used

12 earlobe: not used

13 cmo: month of cardiac cath (sp?) (perhaps "call")

14 cday: day of cardiac cath (sp?)

15 cyr: year of cardiac cath (sp?)

16 lmt

17 ladprox

18 laddist

19 diag

20 cxmain

21 ramus

22 om1

23 om2

24 rcaprox

25 rcadist

26 lvx1: not used

27 lvx2: not used

28 lvx3: not used

29 lvx4: not used

30 lvf: not used

31 cathef: not used

32 junk: not used

33 name: last name of patient (I replaced this with the dummy string "name").

After deleting the above columns we ended up having 43 attributes of 76 attributes that we assume is relevant to predict the heart disease and have exported the data into csv format for further analysis. Image 1 represents the code snippet for the work done to preprocess/clean hungarian data. Similar steps were followed to clean cleveland data.

Image 1. Preprocess/Cleaning the raw data.

```
import os
import pandas as pd
import numpy as np

# changing the directory to the path where the data.txt file exists in general.
os.chdir("C:/Users/vsheela/Downloads")

# code used to converge mulitple psan rows into a single row per patientid.
lines = tuple(open("Hungarian Data.txt", 'r'))

list_raw = []
one_row = []

for i in range(len(lines)):
    row = lines[i].strip('\n')
    split_row = row.split(" ")
    if "name" in split_row:
        for i in split_row:
            if i != "name":
                one_row.append(float(i))
        list_raw.append(one_row)
        one_row=[]
    else:
        for i in split_row:
            one_row.append(float(i))

singlerow_df = pd.DataFrame(list_raw)

# Name unnamed columns
singlerow_df.columns = ["id", "ccf", "age", "sex", "painloc", "painexer", "relrest", "pncaden", "cp","trestbps","htn","chol

# Replace -9 to NaN
replace_data = singlerow_df.applymap(lambda x: np.NaN if x==-9 else x)

# delete unnecessary/unused columns.
usedcolumns_df = replace_data[replace_data.columns[:58]]
for feat in ['id','ccf','ekgday','ekgmo','ekgyr','proto','dummy','restckm','exerckm',
             'thalsev','thalpul','earlobe','cmo','cday','cyr']:
    del usedcolumns_df[feat]

usedcolumns_df.to_csv("Hungarian used columns.csv", index=False)
```

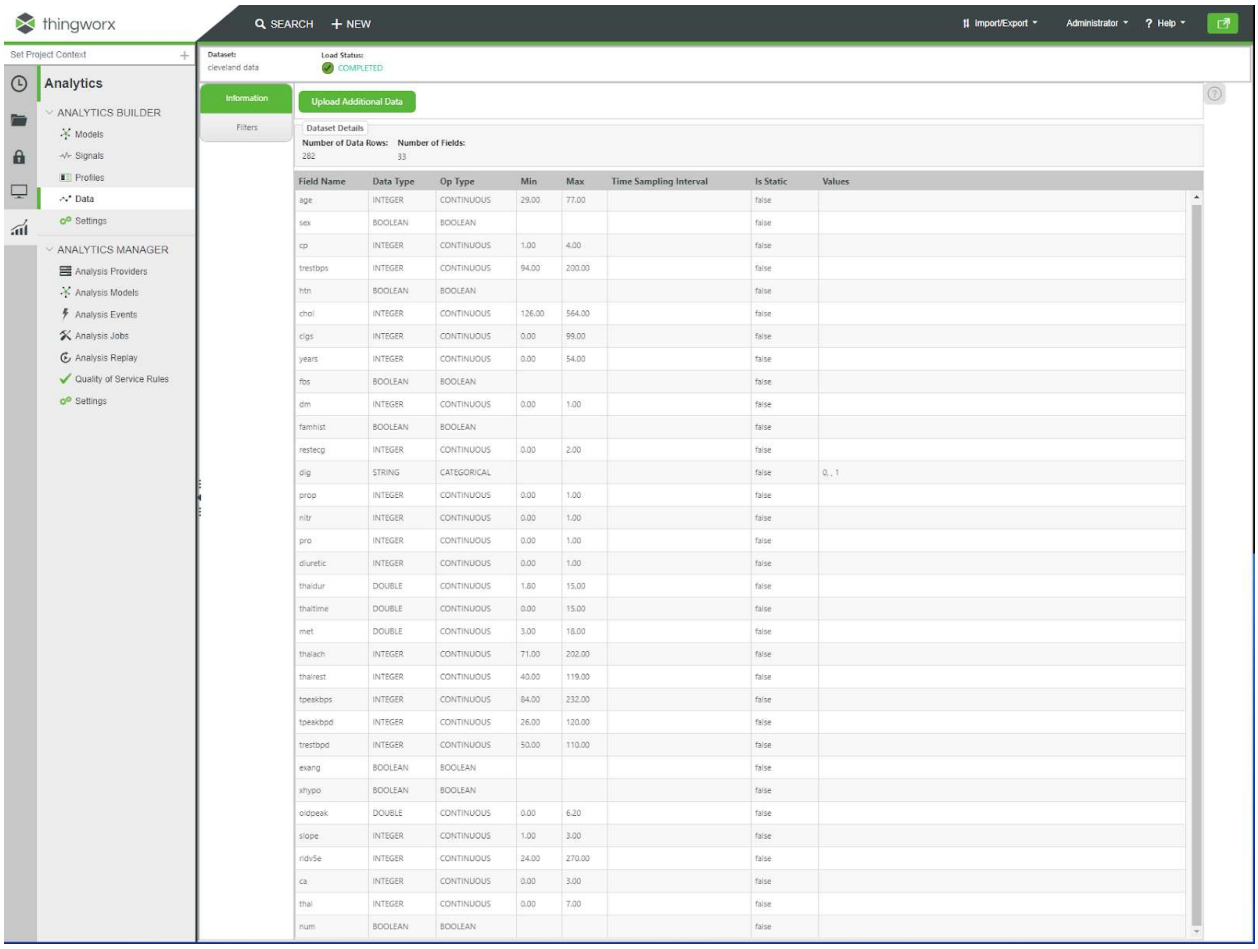
2.2 Feature Selection:

Even after extensive review of literature we were not able to figure out why any of the literature was only using the aforementioned (UCI Machine Learning Repository) 13 parameters as features and continuing with their work. To this extent we have used a trial version of

Analytics tool from Parametric Technology Corporation (PTC) to identify the parameters that contribute to heart disease.

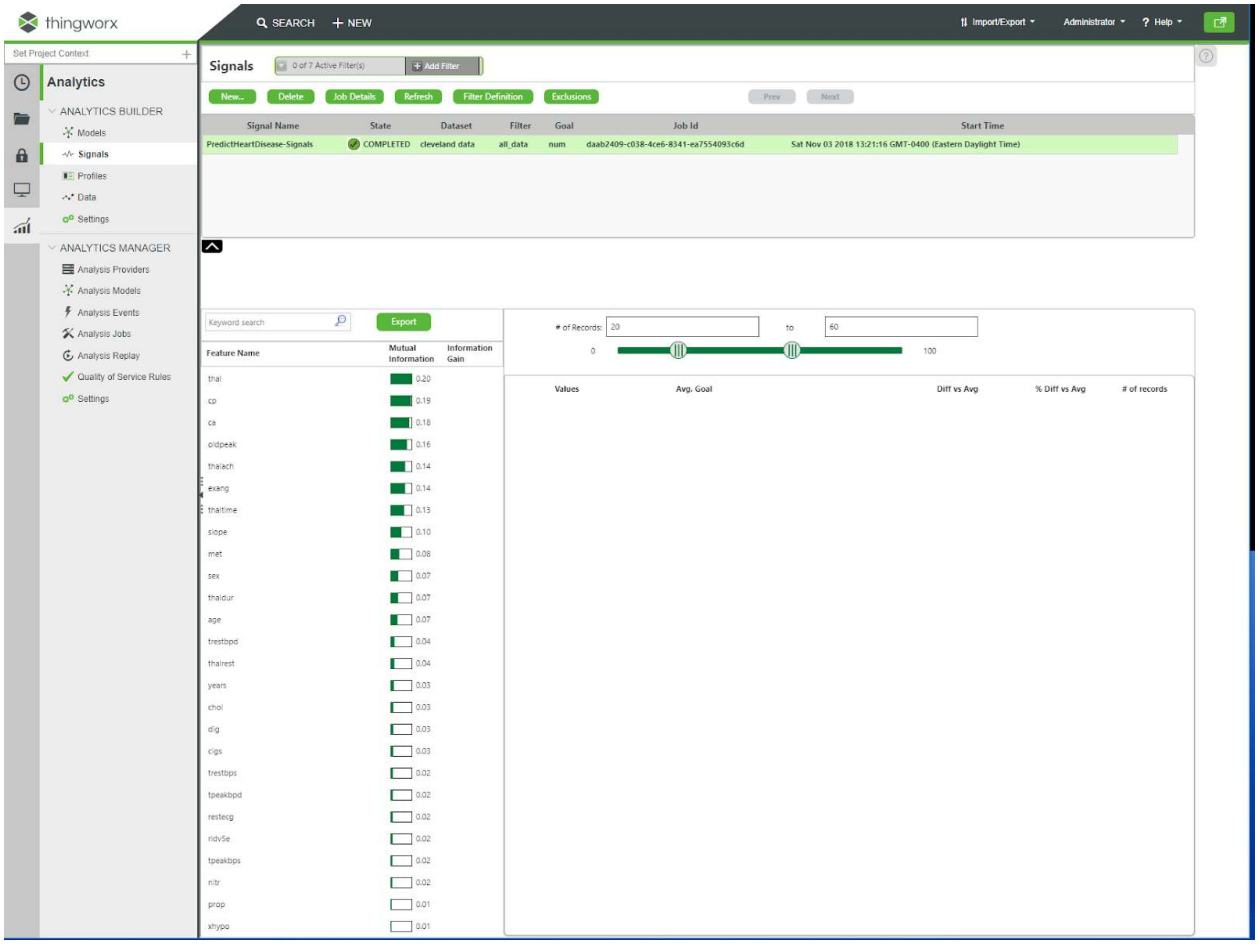
In order for us to use the tool we have to first create a json file which has the data structure and then use the output that we retrieved from cleanup. Image 2 shows the dataset details.

Image 2. Dataset Details from PTC Thingworx.



The reason behind picking this tool is that it has a smart feature where it can provide us with the mutual information column (Image 3) which is an important data point from which we can understand the contribution of the feature towards predicting the heart disease and hence select the features based on the mutual information.

Image 3. Signal Data from PTC Thingworx.



Based on the mutual information provided from the signals on the Cleveland dataset we have decided to use the features that has a mutual information above and equal to 0.04 which now boils down to using 15 data points for further analysis. Below are the 15 parameters that we have used for further analysis using Weka tool.

1 thal

2 cp

3 ca

4 oldpeak

5 thalach

6 slope

7 exang

8 thaltime

9 slope

10 met

11 sex

12 thaldur

13 age

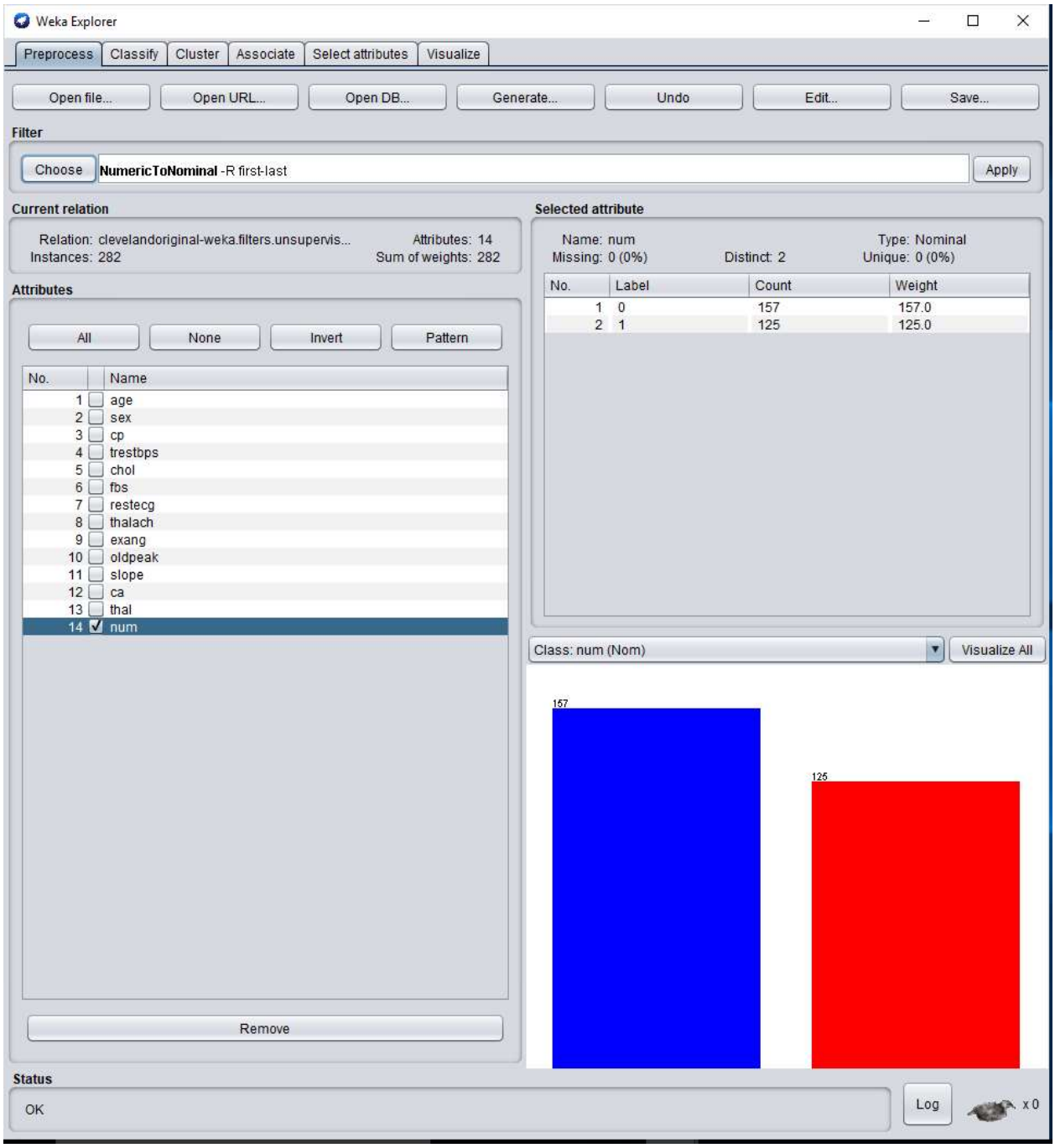
14 trestbpd

15 thalrest

2.3 Experimental Results: Following the literature [7],[8] we have used Weka 3.8 for further analysis in prediction of heart disease using the parameters that were generated using the PTC Thingworx tool. In order for us to use the weka tool we have to modify the csv files into arff files, to this end we have first removed unnecessary columns from the csv files (data obtained from cleanup step using Python code) that we generated. Using Weka tool we modified the csv files into arff files when modifying the csv to arff files missing/Null columns are replaced with "?" in Weka tool.

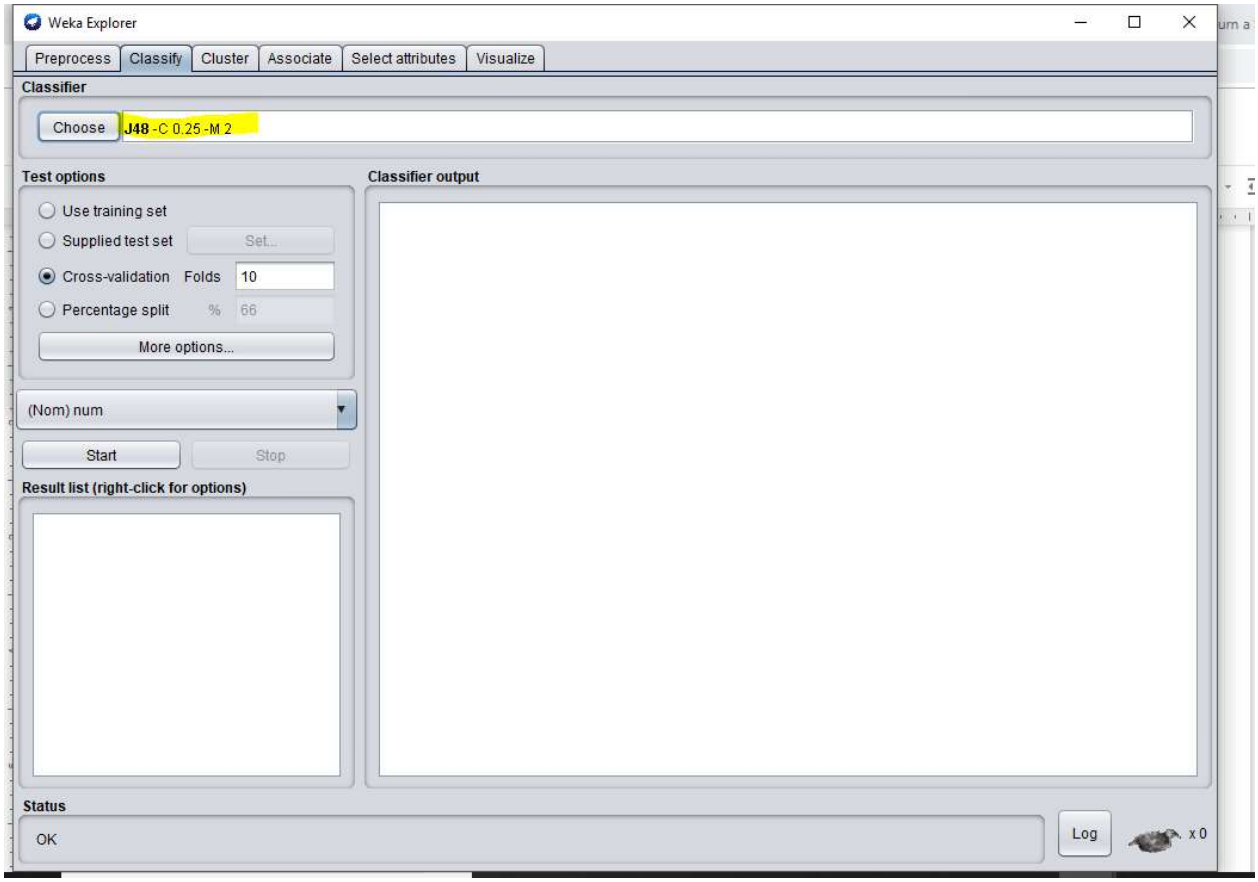
We first have tried to reproduce the results from the literature [7] so that we can set up a baseline. As the "num" feature that we are trying to predict is a numerical value we first have modified "num" into nominal value using unsupervised filter "NumericToNominal" (Image 4). Weka tool gives confusion matrix only for nominal value hence we have converted the predictive class numeric attribute to nominal attribute.

Image 4. Numeric to Nominal conversion using Weka 3.8.



We have used Naive Bayes, J48 and Random Forest classifiers in the Weka tool for our analysis. Image 5 shows an example of selecting a classifier and 10 fold cross validation that we have implemented for J48.

Image 5. Classifier selection using Weka 3.8.



We have used Naive Bayes, J48 and Random Forest classifiers on the 14 parameters that were described in the literature using 282 records (Clean dataset) from our python code to reproduce the results. Image 6 below shows the summary of the data that was produced running the J48 classifier on the 14 parameters specified in the literature on Cleveland dataset.

Image 6. J48 Classifier using literature parameters on Cleveland Data.

=== Summary ===

Correctly Classified Instances	211	74.8227 %
Incorrectly Classified Instances	71	25.1773 %
Kappa statistic	0.4861	
Mean absolute error	0.3129	

Root mean squared error	0.4336
Relative absolute error	63.3883 %
Root relative squared error	87.283 %
Total Number of Instances	282

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.803	0.320	0.759	0.803	0.780	0.487	0.781	0.765	0
	0.680	0.197	0.733	0.680	0.705	0.487	0.781	0.718	1
Weighted Avg.	0.748	0.266	0.747	0.748	0.747	0.487	0.781	0.744	

=== Confusion Matrix ===

a	b	<-- classified as
126	31	a = 0
40	85	b = 1

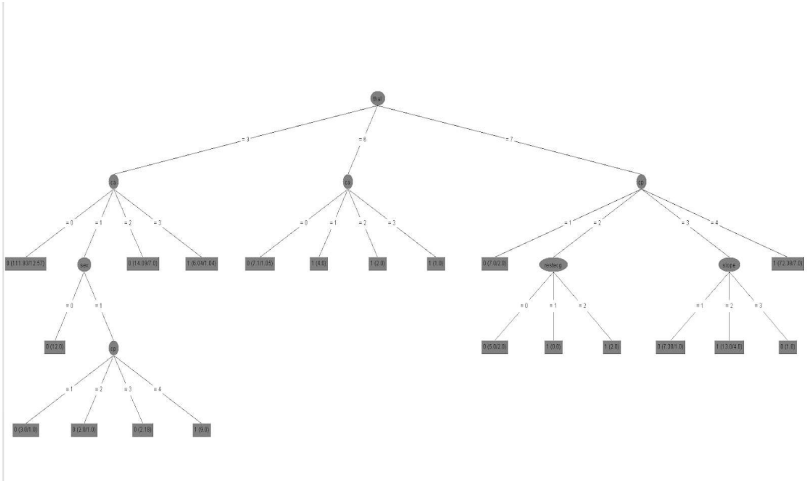


Image 7 below shows the summary of the data that was produced running the Naive Bayes classifier on the 14 parameters specified in the literature on Cleveland dataset.

Image 7. Naive Bayes Classifier using literature parameters on Cleveland Data.

=== Summary ===

Correctly Classified Instances	231	81.9149 %
Incorrectly Classified Instances	51	18.0851 %

Kappa statistic 0.6321
Mean absolute error 0.2138
Root mean squared error 0.3862
Relative absolute error 43.3068 %
Root relative squared error 77.7346 %
Total Number of Instances 282

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.854	0.224	0.827	0.854	0.840	0.632	0.883	0.895	0
	0.776	0.146	0.808	0.776	0.792	0.632	0.883	0.863	1
Weighted Avg.	0.819	0.190	0.819	0.819	0.819	0.632	0.883	0.881	

=== Confusion Matrix ===

a	b	<-- classified as
134	23	a = 0
28	97	b = 1

Image 8 below shows the summary of the data that was produced running the Random Forest classifier on the 14 parameters specified in the literature on Cleveland dataset.

Image 8. Random Forest Classifier using literature parameters on Cleveland Data.

=== Summary ===
Correctly Classified Instances 191 67.7305 %
Incorrectly Classified Instances 91 32.2695 %
Kappa statistic 0.327
Mean absolute error 0.4311
Root mean squared error 0.4566
Relative absolute error 87.3276 %
Root relative squared error 91.9111 %
Total Number of Instances 282

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.822	0.504	0.672	0.822	0.739	0.339	0.743	0.787	0
	0.496	0.178	0.689	0.496	0.577	0.339	0.743	0.661	1
Weighted Avg.	0.677	0.360	0.679	0.677	0.667	0.339	0.743	0.731	

=== Confusion Matrix ===

a	b	<-- classified as
129	28	a = 0
63	62	b = 1

Once we were able to produce the results using various classifiers mentioned we have then used the 15 parameters that we generated from the PTC tool to predict the heart diseases using the same classifiers. Image 9 below shows the summary of the data that was produced running the J48 classifier.

Image 9. J48 Classifier using PTC Tool parameters on Cleveland Data.

=== Summary ===

Correctly Classified Instances	211	74.8227 %
Incorrectly Classified Instances	71	25.1773 %
Kappa statistic	0.4844	
Mean absolute error	0.346	
Root mean squared error	0.4308	
Relative absolute error	70.0834 %	
Root relative squared error	86.702 %	
Total Number of Instances	282	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.815	0.336	0.753	0.815	0.783	0.487	0.750	0.757	0

	0.664	0.185	0.741	0.664	0.700	0.487	0.750	0.685	1
Weighted Avg.	0.748	0.269	0.748	0.748	0.746	0.487	0.750	0.725	

=== Confusion Matrix ===

a	b	<-- classified as
128	29	a = 0
42	83	b = 1

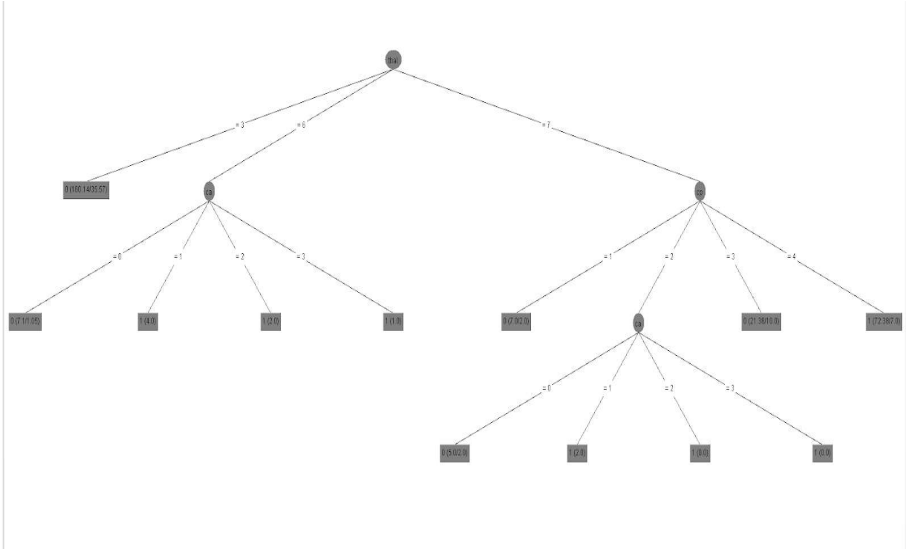


Image 10 below shows the summary of the data that was produced running the Naive Bayes classifier.

Image 10. Naive Bayes Classifier using PTC Tool parameters on Cleveland Data.

=== Summary ===

Correctly Classified Instances	224	79.4326 %
Incorrectly Classified Instances	58	20.5674 %
Kappa statistic	0.5806	
Mean absolute error	0.213	
Root mean squared error	0.3903	
Relative absolute error	43.1443 %	
Root relative squared error	78.5643 %	
Total Number of Instances	282	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.841	0.264	0.800	0.841	0.820	0.582	0.878	0.869	0
	0.736	0.159	0.786	0.736	0.760	0.582	0.878	0.861	1
Weighted Avg.	0.794	0.218	0.794	0.794	0.793	0.582	0.878	0.865	

=== Confusion Matrix ===

a	b	<-- classified as
132	25	a = 0
33	92	b = 1.

Image 11 below shows the summary of the data that was produced running the Random Forest classifier.

Image 11. Random Forest Classifier using PTC Tool parameters on Cleveland Data.

=== Summary ===

Correctly Classified Instances	189	67.0213 %
Incorrectly Classified Instances	93	32.9787 %
Kappa statistic	0.3099	
Mean absolute error	0.4426	
Root mean squared error	0.4661	
Relative absolute error	89.6504 %	
Root relative squared error	93.815 %	
Total Number of Instances	282	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.828	0.528	0.663	0.828	0.737	0.324	0.700	0.736	0
	0.472	0.172	0.686	0.472	0.559	0.324	0.700	0.655	1
Weighted Avg.	0.670	0.370	0.673	0.670	0.658	0.324	0.700	0.700	

=== Confusion Matrix ===

a	b	<-- classified as
130	27	a = 0
66	59	b = 1

After running the above mentioned classifiers we have used Weka 3.8 tool on literature specified 14 parameters (Table 2) and PTC tool specified 15 parameters (Table 3) to interpret the data.

Table 2. Data from various classifiers on literature parameters using Cleveland dataset.

Statistics	**J48 Decision Tree**	**Naive Bayes**	**Random Forest**
Accuracy of correctly predicted Instances	74.8227	81.9149	67.7305
Inaccuracy	25.1773	18.0851	32.2695
Kappa Statistic	0.4861	0.6321	0.327
Mean absolute error	0.3129	0.2138	0.4311
Root Mean Squared error	0.4336	0.3862	0.4566
Relative absolute error	63.38	43.3068	87.3276
Root relative squared error	87.283	77.7346	91.911
Total Number of Instances	282	282	282
Precision	0.747	0.819	0.679
Recall	0.748	0.819	0.677
F Measure	0.747	0.819	0.667

Table 3. Data from various classifiers on PTC tool parameters using Cleveland dataset.

Statistics	J48 Decision Tree	Naive Bayes	Random Forest
Accuracy of correctly predicted Instances	74.8	79.43	67.02
Inaccuracy	25.17	20.56	32.97
Kappa Statistics	0.484	0.58	0.3
Mean absolute error	0.346	0.213	0.44
Root Mean Squared error	0.43	0.39	0.46
Relative absolute error	70.08	43.14	89.65
Root relative squared error	86.7	78.56	93.8
Total Number of Instances	282	282	282
Precision	0.748	0.794	0.673
Recall	0.748	0.794	0.67
F Measure	0.746	0.793	0.658

Accuracy of the data obtained using various classifiers on literature and the parameters produced from the PTC tool is shown in table 4.

Table 4. Information on Accuracy using various classifiers on Cleveland dataset.

Cleveland Data Source	J48 Decision Tree	Naive Bayes	Random Forest
Literature Accuracy	72.5	77.5	77
Reproducing Literature work*	74.8	81.9	67.7
Accuracy from PTC tool parameters	74.8	79.43	67

***As we have deleted the corrupt data from the cleveland dataset hence when calculating the accuracy for reproducing the literature work we have used the baseline of 303 records.**

One of the other important goals that we have discussed in the beginning of this work was to analyze why there was only literature work on cleveland dataset but not any other dataset. Hence, to this respective we have used the signals obtained from the PTC tool using the Hungarian dataset and analyzed the data using the same classifiers that we have mentioned above. Image 12 below shows the summary of the data that was produced running the J48 classifier on the 14 parameters specified in the literature on Hungarian dataset.

Image 12. J48 Classifier using literature parameters on Hungarian Data

```
=== Summary ===
Correctly Classified Instances         238               81.2287 %
Incorrectly Classified Instances        55               18.7713 %
Kappa statistic                          0.5789
Mean absolute error                      0.3065
Root mean squared error                  0.3922
Relative absolute error                 66.3326 %
Root relative squared error             81.6093 %
Total Number of Instances              293

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.898    0.340    0.824      0.898    0.859      0.584    0.736     0.777     0
                 0.660    0.102    0.787      0.660    0.718      0.584    0.736     0.614     1
Weighted Avg.    0.812    0.254    0.810      0.812    0.808      0.584    0.736     0.718

=== Confusion Matrix ===
   a   b   <-- classified as
 168  19 |   a = 0
  36  70 |   b = 1
```

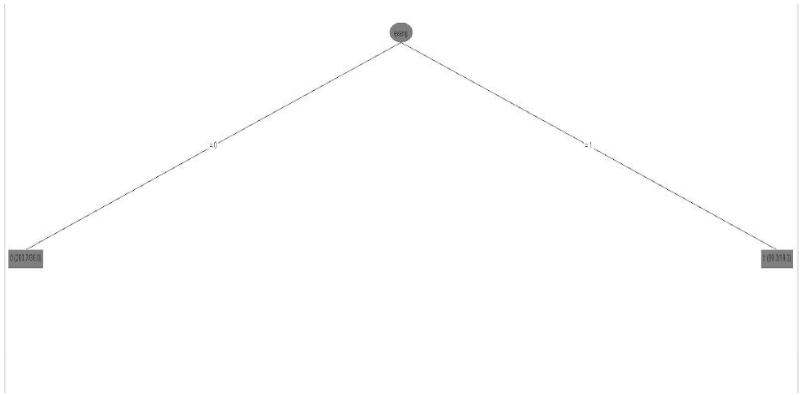


Image 13 below shows the summary of the data that was produced running the Naive Bayes classifier on the 14 parameters specified in the literature on Hungarian dataset.

Image 13. Naive Bayes Classifier using literature parameters on Hungarian Data.

```
=== Summary ===
Correctly Classified Instances         249               84.9829 %
Incorrectly Classified Instances        44               15.0171 %
Kappa statistic                          0.6652
Mean absolute error                      0.1803
Root mean squared error                  0.3503
Relative absolute error                 39.0077 %
Root relative squared error             72.9001 %
Total Number of Instances              293

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.274    0.856      0.920    0.887      0.669    0.904     0.931     0
                 0.726    0.080    0.837      0.726    0.778      0.669    0.904     0.854     1
Weighted Avg.    0.850    0.204    0.849      0.850    0.847      0.669    0.904     0.903

=== Confusion Matrix ===

   a   b   <-- classified as
 172  15 |   a = 0
  29  77 |   b = 1
```

Image 14 below shows the summary of the data that was produced running the Random Forest classifier on the 14 parameters specified in the literature on Hungarian dataset.

Image 14. Random Forest Classifier using literature parameters on Hungarian Data.

```
=== Summary ===
Correctly Classified Instances         220               75.0853 %
Incorrectly Classified Instances        73               24.9147 %
Kappa statistic                          0.4032
Mean absolute error                      0.3692
Root mean squared error                  0.4109
Relative absolute error                 79.8832 %
Root relative squared error             85.4951 %
Total Number of Instances              293

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.930    0.566    0.744      0.930    0.827      0.437    0.840     0.887     0
                 0.434    0.070    0.780      0.434    0.558      0.437    0.840     0.759     1
Weighted Avg.    0.751    0.386    0.757      0.751    0.729      0.437    0.840     0.841

=== Confusion Matrix ===

   a   b   <-- classified as
 174  13 |   a = 0
  60  46 |   b = 1
```

Once we were able to produce the results using various classifiers mentioned we have then used the 15 parameters that we generated from the PTC tool to predict the heart diseases using the same classifiers on Hungarian dataset. Image 15 below shows the summary of the data that was produced running the J48 classifier.

Image 15. J48 Classifier using PTC Tool parameters on Hungarian Data.

=== Summary ===

Correctly Classified Instances	238	81.2287 %
Incorrectly Classified Instances	55	18.7713 %
Kappa statistic	0.5789	
Mean absolute error	0.3065	
Root mean squared error	0.3922	
Relative absolute error	66.3326 %	
Root relative squared error	81.6093 %	
Total Number of Instances	293	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.898	0.340	0.824	0.898	0.859	0.584	0.736	0.777	0
	0.660	0.102	0.787	0.660	0.718	0.584	0.736	0.614	1
Weighted Avg.	0.812	0.254	0.810	0.812	0.808	0.584	0.736	0.718	

=== Confusion Matrix ===

a	b	<-- classified as
168	19	a = 0
36	70	b = 1

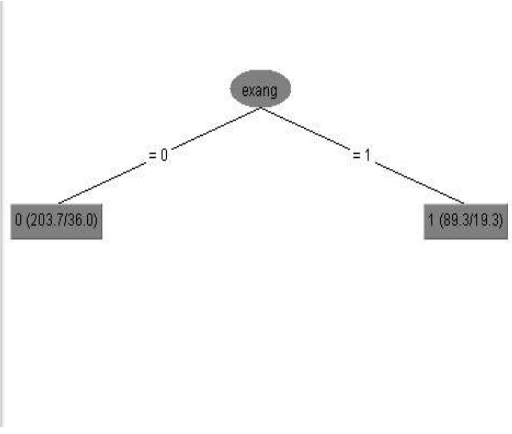


Image 16 below shows the summary of the data that was produced running the Naive Bayes classifier on Hungarian data using PTC tool specified parameters.

Image 16. Naive Bayes Classifier using PTC Tool parameters on Hungarian Data.

```
=== Summary ===
Correctly Classified Instances         244               83.2765 %
Incorrectly Classified Instances        49               16.7235 %
Kappa statistic                          0.6295
Mean absolute error                      0.1933
Root mean squared error                  0.374
Relative absolute error                 41.8326 %
Root relative squared error             77.8235 %
Total Number of Instances              293

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.898    0.283    0.848      0.898    0.873      0.632    0.883     0.913     0
                 0.717    0.102    0.800      0.717    0.756      0.632    0.883     0.809     1
Weighted Avg.    0.833    0.217    0.831      0.833    0.831      0.632    0.883     0.875

=== Confusion Matrix ===

   a   b   <-- classified as
 168  19 |   a = 0
  30  76 |   b = 1
```

Image 17 below shows the summary of the data that was produced running the Random Forest classifier on Hungarian data using PTC tool specified parameters.

Image 17. Random Forest Classifier using PTC Tool parameters on Hungarian Data.

=== Summary ===

Correctly Classified Instances	219	74.744 %
Incorrectly Classified Instances	74	25.256 %
Kappa statistic	0.3853	
Mean absolute error	0.3764	
Root mean squared error	0.4167	
Relative absolute error	81.442 %	
Root relative squared error	86.7059 %	
Total Number of Instances	293	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.947	0.604	0.734	0.947	0.827	0.431	0.823	0.876	0
	0.396	0.053	0.808	0.396	0.532	0.431	0.823	0.745	1
Weighted Avg.	0.747	0.405	0.761	0.747	0.720	0.431	0.823	0.829	

=== Confusion Matrix ===

a	b	<-- classified as
177	10	a = 0
64	42	b = 1

After running the above mentioned classifiers we have used Weka 3.8 tool on literature specified 14 parameters (Table 5) and PTC tool specified 15 parameters (Table 6) to interpret the data.

Table 5. Data from various classifiers on literature parameters using Hungarian dataset.

Statistics	**J48 Decision Tree**	**Naive Bayes**	**Random Forest**
Accuracy of correctly predicted Instances	81.2287	84.9829	75.0853
Inaccuracy	18.7713	15.0171	24.9147

Kappa Statistic	0.5789	0.6652	0.4032
Mean absolute error	0.3065	0.1803	0.3692
Root Mean Squared error	0.3922	0.3503	0.4109
Relative absolute error	66.3326	39.0077	79.8832
Root relative squared error	81.6093	72.9001	85.4951
Total Number of Instances	293	293	293
Precision	0.810	0.849	0.757
Recall	0.812	0.850	0.751
F Measure	0.808	0.847	0.729

Table 6. Data from various classifiers on PTC tool parameters using Hungarian dataset.

Statistics	**J48 Decision Tree**	**Naive Bayes**	**Random Forest**
Accuracy of correctly predicted Instances	81.2287	83.2765	74.744
Inaccuracy	18.7713	16.7235	25.256
Kappa Statistic	0.5789	0.6295	0.3853
Mean absolute error	0.3065	0.1933	0.3764
Root Mean Squared error	0.3922	0.374	0.4167
Relative absolute error	66.3326	41.8326	81.442
Root relative squared error	81.6093	77.8235	86.7059
Total Number of Instances	293	293	293

Precision	0.810	0.831	0.761
Recall	0.812	0.833	0.747
F Measure	0.808	0.831	0.720

Accuracy of the data obtained using various classifiers on literature and the parameters produced from the PTC tool is shown in table 7.

Table 7. Information on Accuracy using various classifiers on Hungarian dataset.

Hungarian Data Source	**J48 Decision Tree**	**Naive Bayes**	**Random Forest**
Reproducing Literature work*	81.2287	84.9829	75.0853
Accuracy from PTC tool parameters	81.2287	83.2765	74.744

CHAPTER 3
CONCLUSIONS

1. We were able to reproduce the literature work on Cleveland dataset and was able to almost match the results with a slight percentage difference.
2. We have also discovered new parameters using PTC tool in prediction of heart disease and the results almost matched with a slight percentage difference compared to literature work.
3. Based on the results from the Weka tool and accuracy of correctly predicted instances we can conclude that Naive Bayes Algorithm predicts heart disease more accurately than other 2 classifiers in both Cleveland and Hungarian datasets.
4. Although Hungarian dataset was able to predict the heart disease with highest accuracy, from the J48 decision tree we have seen that the prediction was only based on a single parameter which underfits the data prediction.
5. We have also concluded with the results from Hungarian data that due to the missing values there was not much of a research work in the literature.

REFERENCES

[1] Jae-Hong Eom and Sung-Chun Kim, et al, “AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction”, Journal of Expert Systems with Applications, Vol. 34 2465, PP.2479, 2008.

[2]. Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2015 on CDC WONDER Online Database, released December 2016. Data are from the Multiple Cause of Death Files, 1999-2015, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10.html.

[3]. Carlos Ordonez, “Association Rule Discover with the Train and Test Approach for the Heart Disease Prediction”, IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, PP. 334-343, April 2006.

[4]. Jesmin Nahar and Tasadduq Imam et al, “Association rule mining to detect factors which contribute to heart disease in males and females”, Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013.

[5]. Peter C. Austin and Jack V. Tu, et al, “Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes” Journal of Clinical Epidemiology, Vol.66, PP. 398-407, 2013.

[6]. P.K. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules”, Journal of Computer Sciences, Vol.24, PP. 27–40, 2012.

[7]. Keerthana TK “Heart Disease Prediction System using Data Mining Method”. International Journal of Engineering Trends and Technology, 47, 361-363.

[8]. Chaitrali S. Dangare and Sulabha S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques” International Journal of Computer Applications, Volume 47– No.10, June 2012.

[9]. Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M.Essam Khalifa, “Feature Analysis of Coronary Artery Heart Disease Data Sets” Procedia Computer Science 65 (2015) 459 – 468, 2015.

[10]. UCI Machine Learning Laboratory https://archive.ics.uci.edu/ml/datasets/heart+Disease.