Mid Term Progress Report

Predication of cardiovascular disease

Arjun Kumar, Dharmil Shah, Anjana Kethineni, Sakshi Parikh, Pranitha Lolla, Anila Chintalapati

Information on Data Set

It is a dataset with about 70000 instances of heart disease. The target variable is the presence of a heart condition marked as a 0 or 1. There are three types of input features:
Objective: Factual Information;
Examination: results of medical examination;
Subjective: information given by the patient.

These are 12 features that are explained as per requirement:

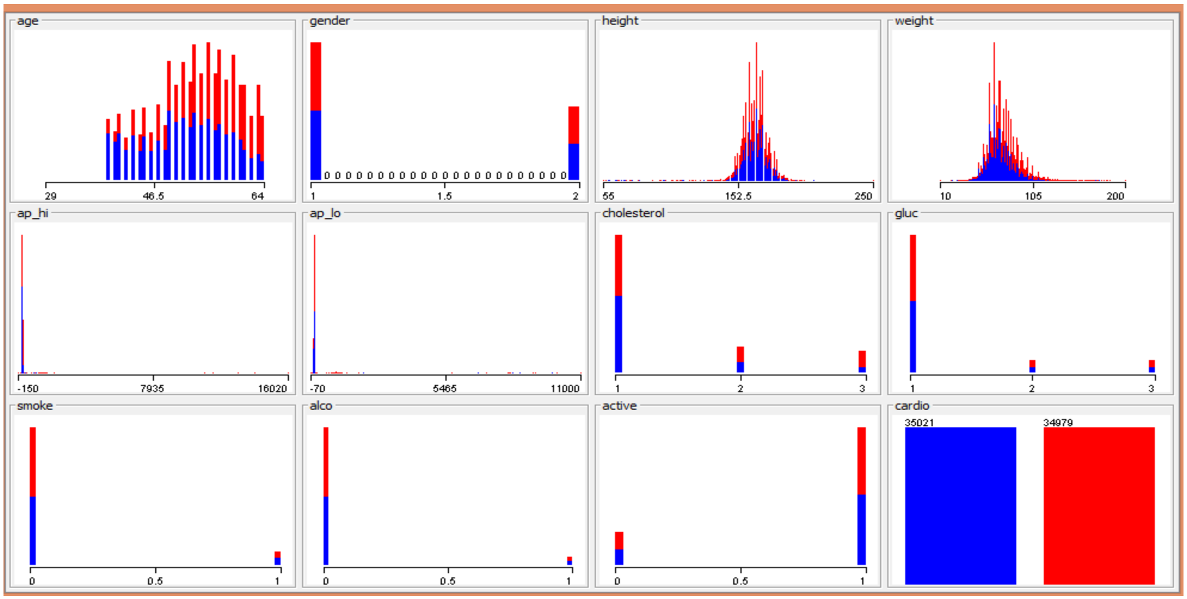
1. Age (Objective Feature) – Number of Days
2. Height (Objective Feature)- Height in Cms
3. Weight (Objective Feature)- In Kg
4. Gender (Objective Feature)- 1 Represent women | 2 Represents men
5. Systolic blood pressure (Examination feature) - A normal systolic pressure is below 120. A reading of 140 or more means high blood pressure
6. Diastolic blood pressure (Examination Feature)-. A normal diastolic blood pressure is lower than 80. A reading of 90 or higher means you have high blood pressure
7. Cholesterol (Examination Feature) 1: normal, 2: above normal, 3: well above normal |
8. Glucose (Examination Feature)1: normal, 2: above normal, 3: well above normal |
9. Smoking (Subjective Feature) 0 ,1 (Binary): This tells if the person smokes or not
10. Alcohol intake (Subjective Feature) Binary (0,1)- This feature tells whether patient drinks alcohol or not
11. Physical activity (Subjective Feature) Binary - This feature tells whether the patient's cardio problem is active or inactive
12. Presence or absence of cardiovascular disease (Target Variable) Binary- This is the target variable which is in binary. If it is a 1, a person has the heart disease if it is 0, then the person does not have a heart disease

All the dataset values were collected during medical examination.

Data Cleaning

Data cleaning was the first step we took to clean the data. Data Cleaning included standardizing and removing spaces between the value which changed the value. The data values had many missing values, which were removed those rows, after removing those data we are having 70000 instances, thus we did not lose much instances during the cleaning process.

Visualization of each attributes

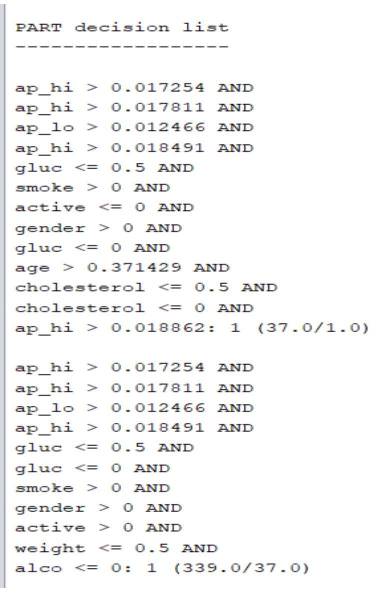


Above image showcase data visualization for each attribute.

Cross Validation Decision Table:

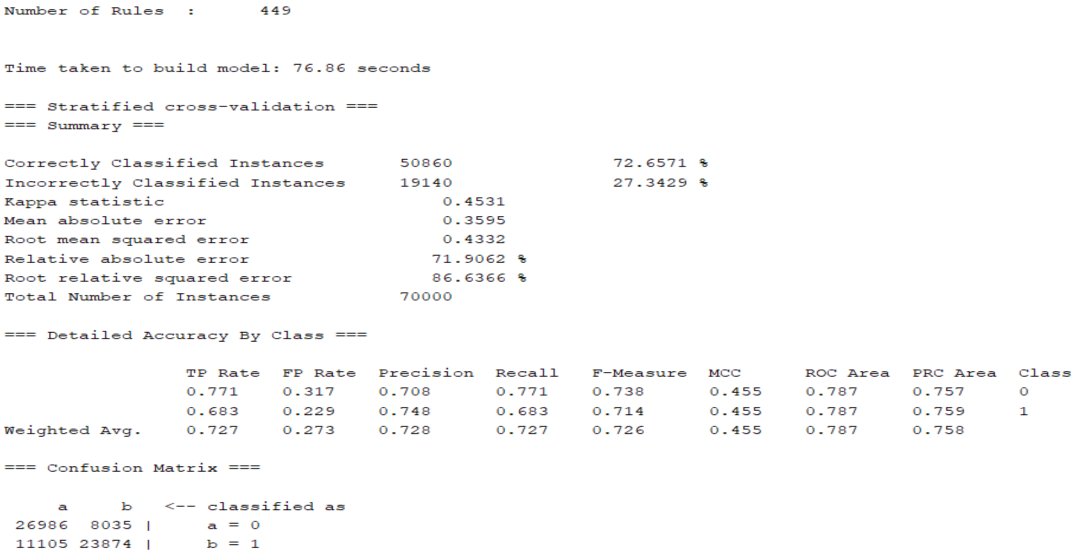
```
Correctly Classified Instances       51123               73.0329 %
Incorrectly Classified Instances     18877               26.9671 %
Kappa statistic                          0.4606
Mean absolute error                      0.3664
Root mean squared error                  0.4281
Relative absolute error                 73.273  %
Root relative squared error             85.6267 %
Total Number of Instances            70000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.774    0.313    0.712      0.774    0.742      0.462
                 0.687    0.226    0.752      0.687    0.718      0.462
Weighted Avg.    0.730    0.270    0.732      0.730    0.730      0.462

=== Confusion Matrix ===

     a     b   <-- classified as
 27092  7929 |     a = 0
 10948 24031 |     b = 1
```

With Cross Validation Decision Table accuracy rate is 73.0329%



```
PART decision list
------------------

ap_hi > 0.017254 AND
ap_hi > 0.017811 AND
ap_lo > 0.012466 AND
ap_hi > 0.018491 AND
gluc <= 0.5 AND
smoke > 0 AND
active <= 0 AND
gender > 0 AND
gluc <= 0 AND
age > 0.371429 AND
cholesterol <= 0.5 AND
cholesterol <= 0 AND
ap_hi > 0.018862: 1 (37.0/1.0)

ap_hi > 0.017254 AND
ap_hi > 0.017811 AND
ap_lo > 0.012466 AND
ap_hi > 0.018491 AND
gluc <= 0.5 AND
gluc <= 0 AND
smoke > 0 AND
gender > 0 AND
active > 0 AND
weight <= 0.5 AND
alco <= 0: 1 (339.0/37.0)
```

Some of the rules using PART- decision list (filter- Normalization)



```
Number of Rules  :      449


Time taken to build model: 76.86 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       50860               72.6571 %
Incorrectly Classified Instances     19140               27.3429 %
Kappa statistic                          0.4531
Mean absolute error                      0.3595
Root mean squared error                  0.4332
Relative absolute error                 71.9062 %
Root relative squared error             86.6366 %
Total Number of Instances            70000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.771    0.317    0.708      0.771    0.738      0.455    0.787     0.757     0
                 0.683    0.229    0.748      0.683    0.714      0.455    0.787     0.759     1
Weighted Avg.    0.727    0.273    0.728      0.727    0.726      0.455    0.787     0.758

=== Confusion Matrix ===

     a     b   <-- classified as
 26986  8035 |     a = 0
 11105 23874 |     b = 1
```

Accuracy is 72.6571% using PART-decision list

Challenges:

The challenging part faced until now is converting the data to weka format. The initial dataset has age in days and challenging part is to convert the days into years for 70000 values. We converted days to years using Excel by using formulas in excel.

Future Work:

We will be doing more visual analysis to analyze the data which will help to find exact relation between the attributes and the target attribute which is if the person is having disease or not. and then will be performing different algorithms like Random Forest, KNN, Logistic Regression to predict the maximum accuracy. To improve the performance, we will be using different parameters across each model.