CSE3505 - Foundations of Data Analytics

J Component Report

A project report titled

Cardio Disease Prediction

By

17MIS122 – G.Jai Surya Gowd
17MIS1152-K.Sai Prakash

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGENEERING

Submitted to

Dr. R. Rajalakshmi

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

November 2020

DECLARATION BY THE CANDIDATE

I hereby declare that the report titled "**Cardio Disease Prediction** **"** submitted by me to VIT Chennai is a record of bona-fide work undertaken by me under the supervision of **Dr. R. Rajalakshmi, Associate Professor, SCOPE, Vellore Institute of Technology, Chennai.**

G.Jai Surya Gowd

Signature of the Candidate

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Jagadeesh Kannan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report entitled "**Cardio Disease Prediction ” is a bona-fide work of** G.Jai Surya Gowd(17MIS1122**),** K.Sai Prakash (17MIS1152**) carried out the "J"-Project work under my supervision and guidance for CSE3505 - Foundations of Data Analytics**

.

Dr. R. Rajalakshmi

SCOPE

TABLE OF CONTENTS

ABSTRACT

Cardiovascular disease (CVD) is the number one cause of death globally. This has
raised many concerns and research to find ways to prevent CVD among people. While it is said that up to 90% of CVD may be preventable, it is difficult to predict and prevent this disease because it involves many risk factors such as sex, family history, smoking and many more. Data analysis and machine learning methods seem to be a reliable way to explore the patients' data, identify risk factors, and predict if a person is likely to have CVD or not. I chose this subject due to its importance, and the integral role of data analysis methods to solve it.

INTRODUCTION

Hospitals are generating huge volume of data regarding their patients. With the advancement made in data analysis over big data, the hospital data is useful in building disease predictive models. Data mining techniques can predict the hidden pattern lying in the voluminous hospital data and helps us to build an effective medical diagnosis system. One or other form of heart disease is found to be the major reason for the death of a patient [1]. Irrespective of the region, country and age group, heart disease is the leading death factor. Heart related diseases needs continuous monitoring and treatment based on it. But for rural people frequent medical checkups are not easily accessible and viable. For the people who are suffering from serious heart disease this condition is a life threatening situation. According to a 2010 survey, for every $6 spent on health care, $1 is for Cardiovascular diseases. Coronary Heart Disease(CHD) is the leading cause of death of 370,000 people annually. But the cost associated with their treatment is estimated to be around $444 billion(US) dollars. The chances of survival are more when it is predicted before an emergency situation occurs. Also, it is observed from the data that the survival of sudden out of hospital heart attack is very low. This paper surveys the different kinds of heart diseases

predictive modeling developed based on machine learning, data mining and artificial intelligence techniques. There are various range of heart diseases apart from heart attack which are collectively called as Cardiovascular diseases. There are many reasons for the development of heart diseases such as smoking, blood sugar, obesity, depression, high cholesterol, poor diet and genetically descendant. There are many types of heart diseases such as angina, arrhythmia, congenital heart disease, fibrillation, coronary artery disease, heart failure, fibrillation. When a person is under heart attack, the tests to be done are CPR, bypass surgery, Value disease treatment, Cardio, Pace makers, heart transplant and so on. The prediction of heart diseases helps us in treating the patient before the patient reaches heart failure.

LITERATURE SURVEY

A brief survey is done over the existing works done for heart disease prediction using data mining and machine learning techniques.

Latha et al [3] have developed a neuro-fuzzy based heart disease prediction model named Co-Active Neuro Fuzzy Inference System(CANFIS). They used the Cleveland data source which has 13 attributes related to heart. Further, to optimize the membership functions, momentum coefficient values, learning rate Genetic Algorithm is used. \

Martin Gjoreski [4] have implemented a chronic heart failure detection system using heart sounds. The proposed method involves filtering, segmentation, feature extraction and stacking of ML classifiers. The data set is collected from a total of 152 heart sounds obtained from 122 different subjects out of which 23 were previously diagnosed with heart abnormalities by the physician. The accuracy obtained from the proposed technique was 96% and in addition, it detects 87% of "unhealthy" instances with a precision of 87% . The heart sounds which were collected with the help of digital stethoscope proves that chronic heart failure can be detected.

I Ketut et al [5] proposed heart disease diagnosis system using K-Nearest Neighbors. A real clinical medical record of 450 data sets consisting 15 important parameters such as sex, age, chest pain, shortness of breath and so on were selected. The expected result for the system was to predict what type of heart disease the patient was suffering. The results out of experiments shows that KNN was the best algorithm with 75.11% accuracy (without parameter) and 74.89%, 74.44% and 73.11% accuracy with parameter weighting with OneR Attribute Evaluator, SVM Attribute Evaluator, and ReliefF Attribute Evaluator, respectively. The other classifier Naïve Bayes yields only 50.44% and SVM achieved 45.11% accuracy.

Abdallah et al [6] have created a platform which consists of an electronic device connected with a smart phone for acquiring the electrocardiogram(ECG) signals

around the clock. In addition, inter-beat(R-R) interval analysis instantly alerts the pre-programmed emergency service number whenever there was an abnormal values in signals more than the threshold. Then the patient's ECG report, heart rate and patient's location was sent to the nearby doctor by the system. The hardware system consists of three leads, a micro controller and a smartphone with users choice as to either being connected or disconnected from the entire system.

Sushmita Manikandan [7] has proposed a Naïve Bayes based predictive heart disease system. They used a binary classifier system supporting patients with the graphical user interface through web. The proposed system used a dataset of having 13 predator variables with one binary response variable which were taken from UCI's machine learning repository. Initially, the data from medical records converted into structured data. For data pre-processing, Rapid Miner was used and the performance of the proposed system was compared with K-Nearest Neighbor, Decision Trees and Random Forest. Anaconda v2.7 packages were used to construct the classifier. The results proved that Gaussian Naïve Bayes gave the highest accuracy of 81.25% than other approaches. In the upcoming classification algorithms the accuracy can be improved.

Proposed System / Module(s) description

1. Analyse what are the significant features related to cardiovascular diseases
2. Suggest a model that can classify and make predictions on whether an individual is susceptible to cardiovascular diseases
3. Bring up some insights about cardiovascular diseases

Models

1. Logistic Regression Model
2. KNN
3. Naive Bayes

Logistic Regression Model

Logistic regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
1 / (1 + e^-value)

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.
Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

KNN

K-Nearest Neighbor or K-NN is a Supervised Non-linear classification algorithm. K-NN is a Non-parametric algorithm i.e it doesn't make any assumption about underlying data or its distribution. It is one of the simplest and widely used algorithm which depends on it's k value(Neighbors) and finds it's applications in many industries like finance industry, healthcare industry etc.

In the KNN algorithm, K specifies the number of neighbors and its algorithm is as follows:

- Choose the number K of neighbor.
- Take the K Nearest Neighbor of unknown data point according to distance.
- Among the K-neighbors, Count the number of data points in each category.
- Assign the new data point to a category, where you counted the most neighbors.

For the Nearest Neighbor classifier, the distance between two points is expressed in the form of **Euclidean Distance.**

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the **overlap metric** (or Hamming distance). In the context of gene expression microarray data, for example, *k*-NN has been employed with correlation coefficients, such as Pearson and Spearman, as a metric.[6] Often, the classification accuracy of *k*-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighbourhood components analysis.

A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the *k* nearest neighbors due to their large number.[7] One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its *k* nearest neighbors. The class (or value, in regression problems) of each of the *k* nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a self-organizing map (SOM), each node is a representative (a center) of a cluster of similar points, regardless of their density in the original training data. *K*-NN can then be applied to the SOM.

Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

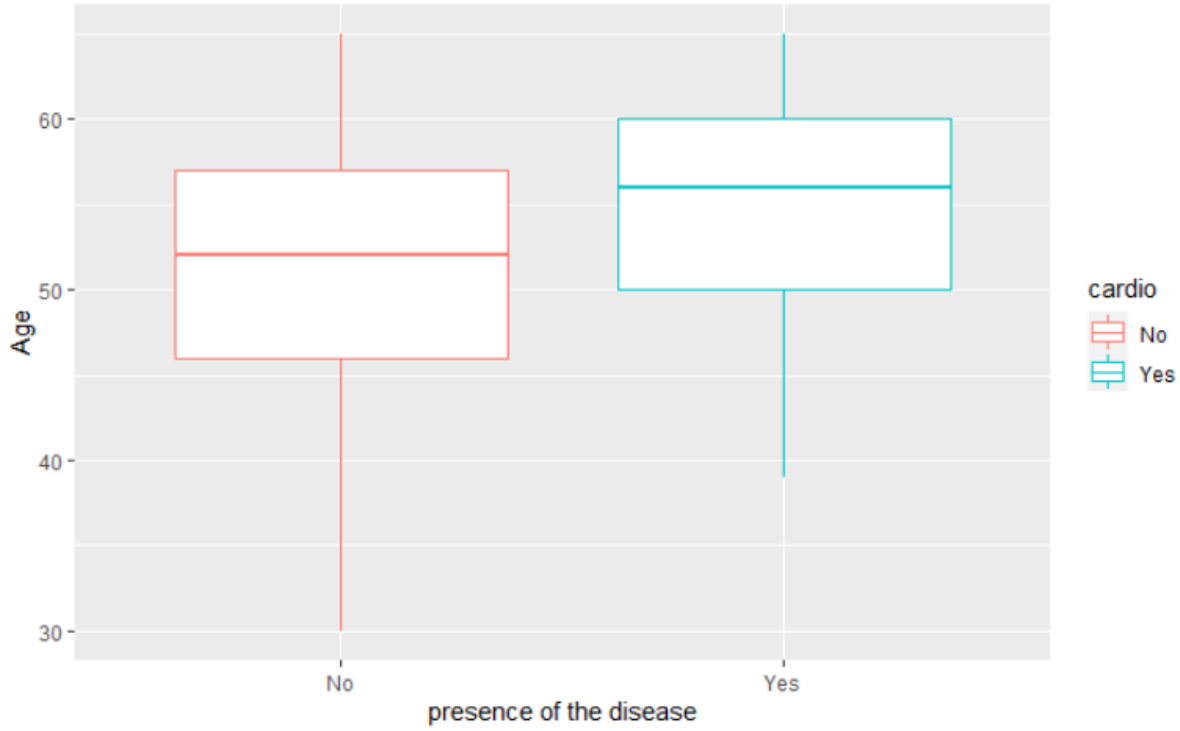
For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Results and Discussion

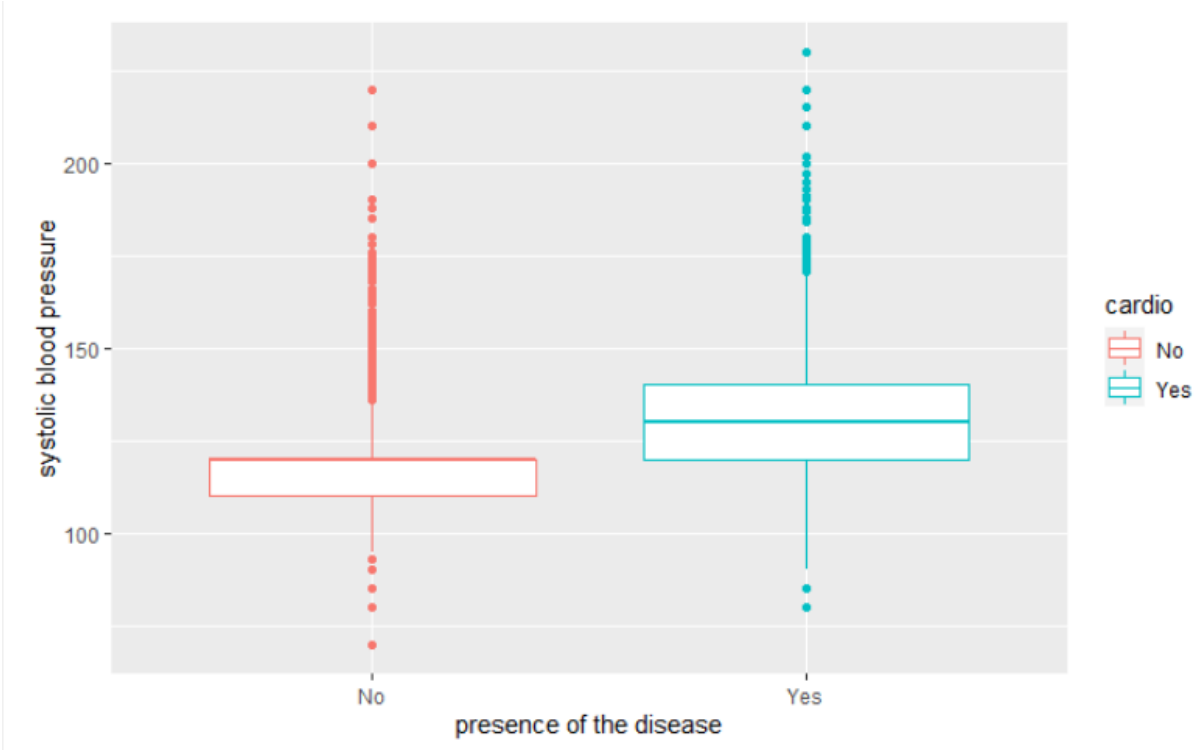
Comparing attributes:

1.Age vs Presence of the Disease



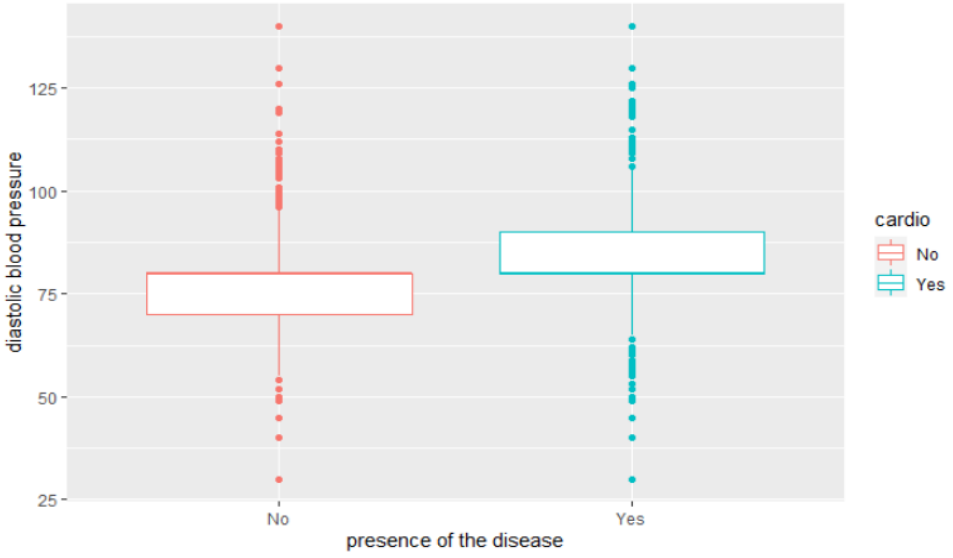
Elder people tend to have Cardio disease more than younger people.

2. Systolic Blood pressure vs Presence of the Disease



Median Systolic blood pressure is higher for the people with Cardio Disease than for the people without Cardio Disease.

3. Diastolic Blood pressure vs Presence of the Disease



Diastolic blood pressure is also high for the people with CVD.

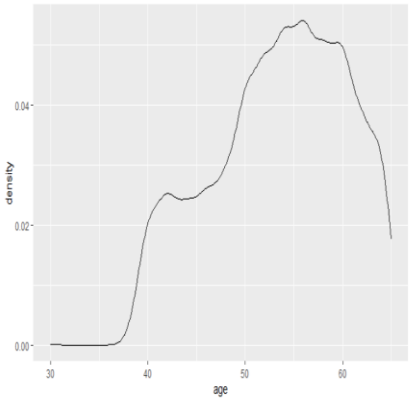
Summary of the dataset

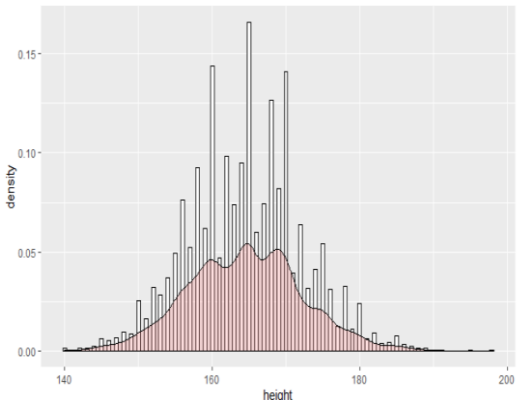
```
      age        gender       height          weight          ap_hi          ap_lo        cholesterol gluc     smoke
 Min.   :30.0   1:4830   Min.   :140.0   Min.   : 34.00   Min.   : 80.0   Min.   : 45.00   1:5501      1:6365   0:6818
 1st Qu.:49.0   2:2652   1st Qu.:159.0   1st Qu.: 65.00   1st Qu.:120.0   1st Qu.: 80.00   2:1057      2: 540   1: 664
 Median :54.0            Median :165.0   Median : 72.00   Median :120.0   Median : 80.00   3: 924      3: 577
 Mean   :53.5            Mean   :164.5   Mean   : 74.35   Mean   :126.6   Mean   : 81.29
 3rd Qu.:59.0            3rd Qu.:170.0   3rd Qu.: 82.00   3rd Qu.:140.0   3rd Qu.: 90.00
 Max.   :65.0            Max.   :198.0   Max.   :169.00   Max.   :230.0   Max.   :140.00
 alco     active   cardio
 0:7056   0:1487   No :3784
 1: 426   1:5995   Yes:3698
```

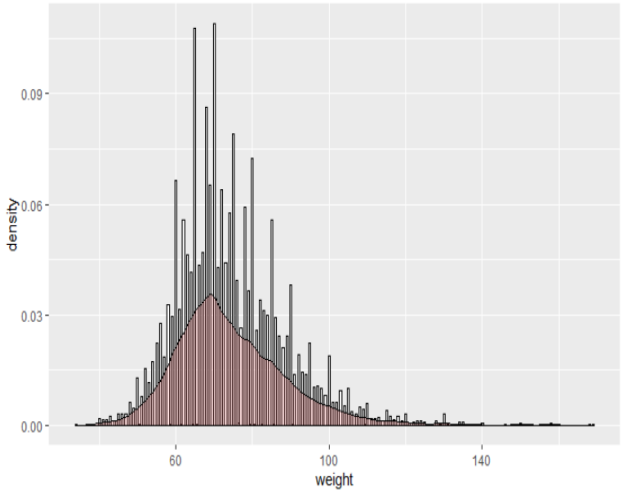
4830 observations are females (gender = 1) and 2652 are males (gender = 2). There is a significant difference in the number of observations between gender. Thus, the data may be slightly biased due to the unequal distribution of gender.

3784 observations are not having cardiovascular diseases and the other 3698 observations suffer from cardiovascular diseases. The approximate ratio is 1:1 which suggests that the dependent variable “cardio” is distributed evenly, and accuracy is thus a reliable method to evaluate how a model performs.

Histogram and density plots are used to identify the distribution of continuous variables.



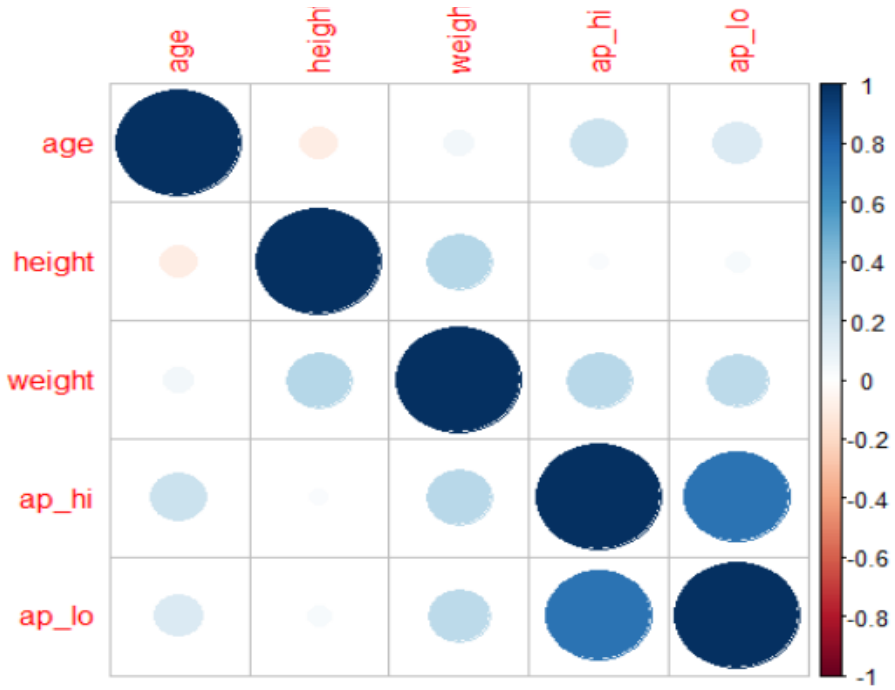




From the above three plots, height, weight and age are not normally distributed and they are all somewhat skewed. This should be taken into account in model selection and validation.

Scatter plots and correlation plot

Scatter plots and correlation plots can visualise and roughly identify the possible correlation between any two continuous variables.



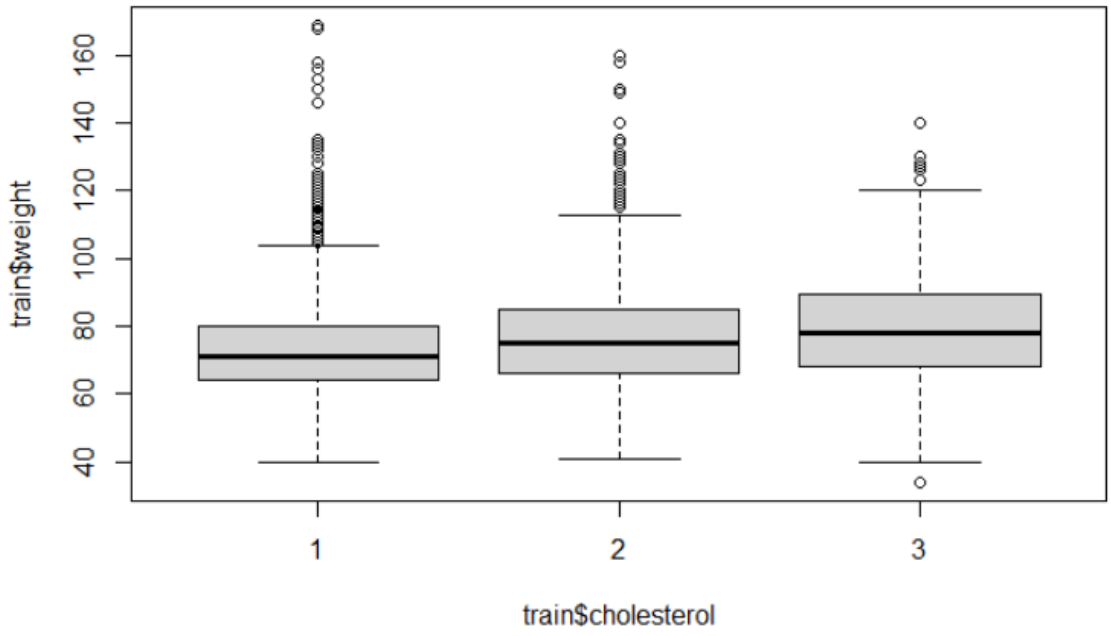
From the above graph, there are correlations between two pairs of continuous variables: ap_hi and ap_low, height and weight. The rest have very low correlation coefficient.

Hypothesis Testing

Hypothesis 1: Do people with different cholesterol levels have different values of mean weight? Are cholesterol and weight correlated?

```
        One-way analysis of means

data:  train$weight and train$cholesterol
F = 97.826, num df = 2, denom df = 7479, p-value < 2.2e-16
```



The null hypothesis is that the mean weight for people having different cholesterol levels is the same. Since the p-value is < 2.2e-16, which is smaller than 0.05. We reject the null hypothesis and conclude that there is a difference in the mean weight among people with various cholesterol levels. Also, from the box plots, we can see that median, upper and lower quartile all increase as cholesterol level rises from 1 (normal) to 3 (well above normal). This implies that independent variables of cholesterol and weight are positively correlated.

Hypothesis 2: Is there a correlation between independent variables height and weight?

```
Call:corr.test(x = train$height, y = train$weight)
Correlation matrix
[1] 0.29
Sample Size
[1] 7482
Probability values  adjusted for multiple tests.
[1] 0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

The correlation coefficient between height and weight is 0.29. Hence, the two variables are moderately correlated.

Hypothesis 3: Are systolic blood pressure (ap_high) and diastolic blood pressure (ap_low) correlated?

```
Call:corr.test(x = train$ap_hi, y = train$ap_lo)
Correlation matrix
[1] 0.73
Sample Size
[1] 7482
Probability values  adjusted for multiple tests.
[1] 0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

By applying the correlation test, the correlation coefficient between systolic blood pressure (ap_hi) and diastolic blood pressure (ap_lo) is 0.74, indicating a strong correlation between these two variables. Thus, interaction term ap_hi * ap_lo should be included in the model

Hypothesis 4: Will gender affect someone's smoking habit?

```
        Pearson's Chi-squared test

data:  train$gender and train$smoke
X-squared = 960.39, df = 1, p-value < 2.2e-16
```

By conducting a Chi-Square test with a contingency table, the above result is obtained. The null hypothesis for the Chi-Square test is that the variables are

independent of one another while the alternative hypothesis is that they are correlated in some way. The p-value is less than 2.2e-16, which is smaller than 0.05. There is sufficient evidence at 5% significance level to reject the null hypothesis and conclude that there is a correlation between gender and smoke.

Logistic Regression Model

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0422  -0.9219  -0.3375   0.9517   2.4713

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -15.704562   3.220644  -4.876 1.08e-06 ***
age            0.055819   0.004106  13.595  < 2e-16 ***
gender2       -0.045204   0.067154  -0.673  0.50086
height         0.026394   0.019517   1.352  0.17625
weight        -0.018751   0.021298  -0.880  0.37863
BMI            0.085267   0.056816   1.501  0.13342
ap_hi          0.058893   0.002823  20.865  < 2e-16 ***
ap_lo          0.001365   0.004380   0.312  0.75527
cholesterol2   0.354659   0.081323   4.361 1.29e-05 ***
cholesterol3   0.844825   0.100281   8.425  < 2e-16 ***
gluc2         -0.146338   0.109940  -1.331  0.18317
gluc3         -0.348608   0.116106  -3.003  0.00268 **
smoke1         0.012751   0.107475   0.119  0.90556
alco1         -0.159171   0.126210  -1.261  0.20725
active1       -0.265381   0.066033  -4.019 5.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10371.3  on 7481  degrees of freedom
Residual deviance:  8429.4  on 7467  degrees of freedom
AIC: 8459.4
```

The result is logit(cardio) = -16.2 + 0.05405 *age_y* + *0.1043* gender2 + 0.03021 *BMI* + *0.09376* ap_hi + 0.06603 *ap_lo* + *0.3682* cholesterol2 + 0.9623 *cholesterol3* + *0.0964* gluc2 - 0.3861 *gluc3* + *0.2294* smoke - 0.4771 *alco* - *0.2402* active - 0.0004463 (ap_hi *ap_lo)* - *0.5499 (gender2* smoke).

Logistic regression model performance on test data

The model performance is evaluated using several indicators based on the number of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) in the confusion matrix. The confusion matrix showing the performance of logistic regression model on test data is below.

```
       No  Yes
No   1053  405
Yes   239  821
```

The accuracy on test data is 72.20%. Recall is 68.60%. Precision is 73.03%. F-measure is 0.707.

KNN

In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression.[1] In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression.

```
k-Nearest Neighbors

7482 samples
  10 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 7482, 7482, 7482, 7482, 7482, 7482, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.6534199  0.3064138
  7  0.6654253  0.3304231
  9  0.6721757  0.3438418

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  983 423
       Yes 309 803

               Accuracy : 0.7093
                 95% CI : (0.6911, 0.727)
    No Information Rate : 0.5131
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4168
```

The accuracy when handling unseen test data is 69.42%

Naive Bayes

Naive Bayes classifies observations based on posterior probability, prior probability and the conditional probability of test data. Also, it assumes that the value of a feature in a given class is independent of the values of other features. The confusion matrix showing the performance of decision tree on test data is below.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1081  464
       Yes  211  762

               Accuracy : 0.7319
                 95% CI : (0.7142, 0.7492)
    No Information Rate : 0.5131
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4606

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6215
            Specificity : 0.8367
         Pos Pred Value : 0.7831
         Neg Pred Value : 0.6997
             Prevalence : 0.4869
         Detection Rate : 0.3026
   Detection Prevalence : 0.3864
      Balanced Accuracy : 0.7291

       'Positive' Class : Yes
```

The computed performance indicators are shown below. The accuracy when handling unseen test data is 71.82%, which is lower than the accuracy in the logistic model on the test data i.e. 72.20%.

Conclusion

The elderly are more likely to suffer from various types of cardiovascular diseases. Cholesterol level is a very important determinant in leading to cardiovascular diseases. The risks of getting cardiovascular diseases climb significantly when the cholesterol content in human body rises above the normal. People with high alcohol consumption have a lower likelihood to have cardiovascular diseases. Physical activities help people become less susceptible to cardiovascular diseases. A female non-smoker is less likely to have cardiovascular diseases, while a male smoker is less likely to get cardiovascular diseases. As long as people have systolic blood pressure ≥ 126, they are classified as those susceptible to cardiovascular diseases. Therefore, high systolic blood pressure plays an important role in predicting that one experiences greater risks of suffering from cardiovascular diseases. Among the three splitting factors, two of them are age. This suggests that apart from systolic blood pressure, age is the second most significant factor in deciding whether a person will have a chance of getting cardiovascular disease.

References :

1.M. Ati, "Knowledge capturing in autonomous system design for chronic disease risk assessment, "Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on, pp. 62-66, 2014.

2. Sowmiya, C., and P. Sumitra. "Analytical study of heart disease diagnosis using classification techniques." In Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017 IEEE International Conference, pp. 1-5. IEEE, 2017.

3. Parthiban, Latha, and R. Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3, no. 3 (2008).

4. Martin Gjoreski, Anton Gradisˇek, Matjazˇ Gams, Monika Simjanoska, Ana Peterlin, Gregor Poglajen et al, “Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers”, 13th International IEEE Conference on Intelligent Environments, 2017.

5. I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan al, “Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records”, 4th International Conference, June 2018.

6. Abdallah Kassem, Mustapha Hamad, Chady El Moucary and Elie Fayad, “A Smart Device for the Detection of Heart Abnormality using R-R Interval”, 28th IEEE International conference on Microelectronics(ICM), 2016

7. Sushmita Manikandan, “Heart Attack Prediction System”, International Conference on Energy, Communication, Data Analytics and Soft Computing, ICCET 2017

8. Anitek Bhattacharya,Mohan Mishra, AnushikhaSingh & Malay Kishore Dutta, “Machine Learning Based Portable Device for Detection of Cardiac Abnormality”, International IEEE conference on Emerging Trends in Computing and Communication Technologies (ICETCCT 2017).

9. Jagdeep Singh, Amit Kamra and Harbhag Singh, “Prediction of Heart Diseases Using Associative Classification”, 5th International Conference on Wireless Networks and Embedded System(WECON 2016).

10. Tahira Mahboob, Rida Irfan, Bazelah Ghaffar, “Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics”, IEEE Internet Technologies and Application(ITA 2017).