Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools

Submitted in partial fulfillment of the requirements for the course of

CSE3020
Data Visualization

by

Kartikay Gupta,
(18BCE2199)

Under the guidance of
RAMANI S
SCOPE VIT, VELLORE

November, 2020

I. Chapter I – INTRODUCTION

The heart is the hardest working muscle in the body. The average heart beats 100,000 times a day, day and night, to supply oxygen and nutrients throughout the body. Blood pumped by the heart also shuttles waste products such as carbon dioxide to the lungs so it can be eliminated from the body. Proper heart function is essential to support life. Coronary artery disease (CAD), commonly known as heart disease, is a condition in which cholesterol, calcium, and other fats accumulate in the arteries that supply blood to the heart. This material hardens forming a plaque that blocks blood flow to the heart. When a coronary artery narrows due to plaque build up or some other cause, the heart muscle is starved for oxygen and a person experiences chest pain known as angina.

1.1 Importance of the study proposed

Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

II. Chapter II –BACKGROUND

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. In almost every 34 seconds the heart disease kills one person in world.

Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer-based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes. Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc. [3].

2.1 Objectives of the proposed study and definition of the problem

Data mining is the most popular knowledge extraction method for knowledge discovery (KDD). Machine learning is used to enable a program to analyze data, understand correlations and make use of insights to solve problems and/or enrich data and for prediction. Data mining techniques and machine learning algorithms play a very important role in medical area. The health care industry contains a huge amount of data. But most of it is not effectively used. Heart disease is one of the main reason for death of people in the world. Nearly 47% of all deaths are caused by heart diseases. We use 4 algorithms including Decision Tree, Logistic model tree algorithm, Random Forest algorithm, Support Vector Machine, to predict the heart diseases. Accuracy of the prediction level is high when using more number of attributes. Our aim is to perform predictive analysis using these data mining, machine learning algorithms on heart diseases and analyze the various mining, Machine Learning algorithms used and conclude which techniques are effective and efficient.

Keywords: Data Mining, Machine Learning, Decision Tree, Heart Disease.

2.2 DATA MINING

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns [2]. The Data mining may be accomplished using classification, clustering, prediction, association and time series analysis.

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods. Thus, data mining refers to mining or extracting knowledge from large amounts of data. Data mining applications will be used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths. Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients [1]. Hence by implementing a heart disease prediction system using Data Mining techniques and doing some sort of data mining on various heart disease attributes, it can able to predict more probabilistically that the patients will be diagnosed with heart disease. This project presents a different model that like the Decision Tree, Random Forest, Logistic Regression, SVM accuracy in identifying heart disease patients. It uses the different algorithm of Machine Learning.

2.3 LITERATURE SURVEY (Identification and exact definition of the problem)

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, random forest logistic regression and support vector machine showing different levels of accuracies on multiple databases of patients from around the world. One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Decision Tree in the diagnosis of heart disease. In [9], the decision making process of heart disease is effectively diagnosed by Random forest algorithm. In [10] based on the probability of decision support, the heart disease is predicted. As a result the author concluded that decision tree performs well and sometimes the accuracy is similar in Bayesian classification. An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Tree, Random Forest and support vector machine and logistic algorithm . The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

Statistical and classification techniques were utilized to develop the multiparametric feature . Besides, they have assessed the linear and the non-linear properties for three recumbent positions, to be precise the supine, left lateral and right lateral position. Numerous experiments were conducted by them on linear and nonlinear characteristics indices to assess several classifiers (Decision Tree) and SVM (Support Vector Machine). SVM surmounted the other classifiers.

2.4 HEART DISEASE (Explanation of problem)

The term Heart sickness alludes to illness of heart & vessel framework inside it. Heart illness is a wide term that incorporates different sorts of sicknesses influencing diverse segments of the heart. Heart signifies "cardio." Therefore, all heart sicknesses fit in with the class of cardiovascular ailments.

2.4.1 SYMPTOMS

The most common symptom of coronary artery disease is angina, or chest pain. Angina can be described as a discomfort, heaviness, pressure, aching, burning, fullness, squeezing, or painful feeling in your chest. It can be mistaken for indigestion or heartburn. Angina may also be felt in the shoulders, arms, neck, throat, jaw, or back.

Other symptoms of coronary artery disease include:

- ✓ Shortness of breath
- ✓ Palpitations (irregular heart beats, or a "flip-flop" feeling in your chest)
- ✓ A faster heartbeat
- ✓ Weakness or dizziness
- ✓ Nausea
- ✓ Sweating
- ✓ lightheadedness and dizzy sensations
- ✓ high levels of fatigue
- ✓ blue tint to the skin

2.4.2 MANIFESTATIONS

Manifestations of a heart assault can include:

- ✓ Discomfort, weight, largeness, or agony in the midsection, arm, or beneath the breastbone.
- ✓ Discomfort emanating to the back, jaw, throat, or arm.
- ✓ Fullness, heartburn, or stifling feeling (may feel like indigestion).
- ✓ Sweating, queasiness, heaving, or unsteadiness.
- ✓ Extreme shortcoming, nervousness, or shortness of breath.
- ✓ Rapid or not regular heart beats

2.4.3 TYPES OF HEART DISEASES

There are many types of heart disease that affect different parts of the organ and occur in different ways.

1. Coronary artery disease (CAD): is the most common type of heart disease. In CAD, the arteries carrying blood to the heart muscle (the coronary arteries) become lined with plaque, which contains materials such as cholesterol and fat. This plaque buildup (called atherosclerosis) causes the arteries to narrow, allowing less oxygen to reach the heart muscle than it needs to work properly. When the heart muscle does not receive enough oxygen, chest pain (angina) or heart attack can occur.

2. Arrhythmia: is an irregular or abnormal heartbeat. This can be a slow heart beat (bradycardia), a fast heartbeat (tachycardia), or an irregular heartbeat. Some of the most common arrhythmias include atrial fibrillation (when the atria or upper heart chambers contract irregularly), premature ventricular contractions (extra beats that originate from the lower heart chambers, or ventricles), and bradyarrhythmia's (slow heart rhythm caused by disease of the heart's conduction system).

3. Heart failure (congestive heart failure, or CHF): occurs when the heart is not able to pump sufficient oxygen-rich blood to meet the needs of the rest of the body. This may be due to lack of force of the heart to pump or as a result of the heart not being able to fill with enough blood. Some people have both problems.

4. Heart valve disease: occurs when one or more of the four valves in the heart are not working properly. Heart valves help to ensure that the blood being pumped through the heart keeps flowing forward. Disease of the heart valves (e.g., stenosis, mitral valve prolapse) makes it difficult.

5. Heart muscle disease (cardiomyopathy): causes the heart to become enlarged or the walls of the heart to become thick. This causes the heart to be less able to pump blood throughout the body and often results in heart failure.

6. Congenital heart disease: is a type of birth defect that causes problems with the heart at birth and occurs in about one out of every 100 live births. Some of the most common types of congenital heart disease include:

atrial septal defects (ASD) and ventricular septal defects (VSD), which occur when the walls that separate the right and left chambers of the hearts are not completely closed.

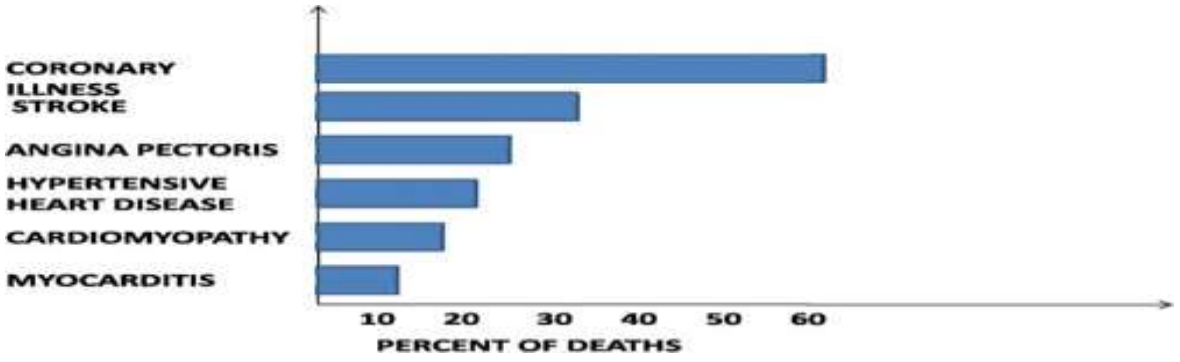


Figure 2.1. Different causes of Heart Disease

Chapter III – Methodology

3.1 DATA DESCRIPTION

We performed computer simulation on one dataset. Dataset is a Heart dataset.Dataset contains 303 samples and 14 input features. Namely

1. **age** (#)
2. **sex** : 1= Male, 0= Female (*Binary*)
3. (**cp**)chest pain type (4 values -*Ordinal*):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic
4. (**trestbps**) resting blood pressure (#)
5. (**chol**) serum cholesterol in mg/dl (#)
6. (**fbs**)fasting blood sugar > 120 mg/dl(*Binary*)(1 = true; 0 = false)
7. (**restecg**) resting electrocardiography results(values 0,1,2)
8. (**thalach**) maximum heart rate achieved (#)
9. (**exang**) exercise induced angina (*binary*) (1 = yes; 0 = no)
10. (**oldpeak**) = ST depression induced by exercise relative to rest (#)
11. (**slope**) of the peak exercise ST segment (*Ordinal*) (Value 1: up sloping , Value 2: flat , Value 3: down sloping)
12. (**ca**) number of major vessels (0–3, *Ordinal*) colored by fluoroscopy
13. (**thal**) maximum heart rate achieved — (*Ordinal*): 3 = normal; 6 = fixed defect; 7 = reversible defect

Note: Our data has 3 types of data:

Continuous (#): which is quantitative data that can be measured

Ordinal Data: Categorical data that has a order to it (0,1,2,3, etc)

Binary Data: data whose unit can take on only two possible states (0 &1)

3.2 DECISION TREE

This paper has emphasized specifically on decision tree classifiers for heart beat prediction within WEKA. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. **These** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees can handle both numerical data and categorical data. For medical purpose, decision trees determine order in different attributes and decision is taken based on the attribute.

A Decision Tree is used to learn a classification function which concludes the value of a dependent attribute (variable) given the values of the independent (input) attributes. This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [4]. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called pruning [5].

The three decision tree algorithms namely J48 algorithm, logistic model tree algorithm and Random Forest decision tree algorithm are used for comparison. The proposed methodology involves reduced error pruning, confident factor and seed parameters to be considered in the diagnosis of heart disease patients. Reduced error pruning has shown to drastically improve decision tree performance. These three decision tree algorithms are then tested to identify which combination will provide the best performance in diagnosing heart disease patients.
Training => Algorithm => Model => Testing => Evaluation

Advantages:

1) Understandable prediction rules are created from the training data.
2) Builds the fastest tree.
3) Builds a short tree.
4) Only need to test enough attributes until all data is classified.

Disadvantages:

1) Data may be over fitted or over classified.
2) Only one attribute at a time is tested for making decision

```
Predicted values:

[1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1
 0 1 1 1 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 1 1 0 1 0
 0 1]

Confusion Matrix:
 [[28 12]
 [11 25]]

Accuracy : 69.737
```

Figure 3.1. Prediction using Decision Tree

3.3 LOGISTIC MODEL TREE ALGORITHM

A logistic model tree basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with k values, the node has k child nodes, and instances are sorted down one of the k branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise.

More formally, a logistic model tree consists of a tree structure that is made up of a set of inner or non-terminal nodes N and a set of leaves or terminal nodes T. Let S = D1 ×⋯× Dm denote the whole instance space, spanned by all attributes V = {v1,...,vm} that are present in the data. Then the tree structure gives a disjoint subdivision of S into regions St , and every region is represented by a leaf in the tree:

$$S = \bigcup_{t \in T} S_t, \quad S_t \cap S_{t'} = \emptyset \ \ for \ \ t \neq t'$$

Figure 3.2

Unlike ordinary decision trees, the leaves t ∈ T have an associated logistic regression function ft instead of just a class label. The regression function ft takes into account an arbitrary subset Vt ⊂ V of all attributes present in the data, and models the class membership probabilities as

$$Pr(G = j | X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}}$$

where

$$F_j(x) = \alpha_0^j + \sum_{v \in V_t} \alpha_v^j \cdot v,$$

or, equivalently,

$$F_j(x) = \alpha_0^j + \sum_{k=1}^{m} \alpha_{v_k}^j \cdot v_k$$

if $\alpha_{v_k}^j = 0$ for $v_k \notin V_t$. The model represented by the whole logistic model tree is then given by

$$f(x) = \sum_{t \in T} f_t(x) \cdot I(x \in S_t)$$

Figure 3.3

where I(x ∈ St) is 1 if x ∈ St and 0 otherwise.

Note that both standalone logistic regression and ordinary decision trees are special cases of logistic model trees, the first is a logistic model tree pruned back to the root, the second a tree in which Vt = ∅ for all t ∈ T.

Ideally, we want our algorithm to adapt to the dataset in question: for small datasets where a simple linear model offers the best bias-variance trade off, the logistic model „tree" should just consist of a single logistic regression model, i.e. be pruned back to the root. For other datasets, a more elaborate tree structure is adequate.

To construct a logistic model tree by developing a standard classification tree, building logistic regression models for all node, pruning a percentage of the sub-trees utilizing a pruning model, and combining the logistic models along a way into a solitary model in some manner is performed.

Tree developing begins by building a logistic model at the root utilizing the LogitBoost algorithm. The quantity of cycles (and basic relapse capacities fmj to add to Fj) is resolved utilizing 10 fold cross-validation. In this process the information is part into preparing and test set 10 times, for each preparation set LogitBoost is rush to a greatest number of cycles and the lapse rates on the test set are logged for each cycle and summed up over the distinctive folds. The quantity of emphases that has the least whole of blunders is utilized to prepare the LogiBoost algorithm on all the information. This gives the logistic regression model at the base of the tree.

```
Predicted values:
[1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 6 30]]

Accuracy : 84.211
```

Figure 3.4 Prediction using Logistic Regression

3.4 RANDOM FOREST ALGORITHM

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the **more trees in the forest** the more robust the forest looks like. In the same way in the random forest classifier, the **higher the number** of trees in the forest gives **the high accuracy** results.

There are three methodologies for Random Forest, for example, Forest-RI(Random Input choice) and Forest-RC (Random blend) and blended of Forest-RI and Forest-RC.

The Random Forest procedure has some desirable qualities, for example

a) It is not difficult to utilize, basic and effortlessly parallelized.
b) It doesn't oblige models or parameters tochoose aside from the quantity of indicators to pick at arbitrary at every node.
c) It runs effectively on extensive databases; it is moderately strong to anomalies and commotion.
d) It can deal with a huge number of information variables without variable deletion; it gives evaluations of what variables are important in classification.
e) It has a successful system for assessing missing information and keeps up accuracy when a vast extent of the data are missing, it has methods for adjusting error in class populace unequal data sets.

Advantages:

1) The same **random forest algorithm** or the random forest classifier can use for both classification and the regression task.
2) Random forest classifier will **handle the missing** values.
3) When we have more trees in the forest, random forest classifier won‟t **overfit** the model.
4) Can model the random forest classifier for **categorical values** also.

Disadvantages:

1) Quite slow to create predictions once trained. More accurate ensembles require more trees, which means using the model becomes slower.
2) Results of learning are incomprehensible. Compared to a single decision tree, or to a set of rules, they don't give you a lot of insight.

```
Predicted values:

[1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 1 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 5 31]]

Accuracy : 85.526
```

Figure 3.5 Prediction using Random Forest

3.5 SUPPORT VECTOR MACHINE

Support vector machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which creates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting on the training data.

Overfitting is caused by incorporating the random errors or noise in the model. Therefore the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keeps the model from over- fitting. The complexity of the model has to be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. A SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most[15].

```
Predicted values:
[1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 6 30]]

Accuracy : 84.211
```

Figure 3.6 Prediction using SVM

3.6 Sample Codes

3.6.1 Logistic Regression

Logistic Regression

```
In [22]: lr=LogisticRegression()
         lr.fit(X_train,y_train)
         y_predict=lr.predict(X_test)
         print(f"Accuracy of Test Dataset: {lr.score(X_test,y_test):0.3f}")
         print(f"Accuracy of Train Dataset: {lr.score(X_train,y_train):0.3f}")
         warnings.simplefilter('ignore')

         Accuracy of Test Dataset: 0.842
         Accuracy of Train Dataset: 0.849
```

Vale Prediction for Test dataset for Logistic Regression

```
In [23]: print("Predicted values:")
         print(y_predict)
         cal_accuracy(y_test, y_predict)

         Predicted values:
         [1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 1 0 0 0 1 0 0 1
          0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 1 0 0 1
          1 0]

         Confusion Matrix:
          [[34  6]
          [ 6 30]]

         Accuracy : 84.211
```

Fig 3.6.1 Logistic Regression code

3.6.2 Support Vector Machine

Support Vector Machine

```
In [24]: from sklearn import svm
         svm_linear = svm.SVC(kernel='linear')
         svm_linear.fit(X_train,y_train)
         warnings.simplefilter('ignore')
         print(f"Accuracy of Test Dataset: {svm_linear.score(X_test,y_test):0.3f}")
         print(f"Accuracy of Train Dataset: {svm_linear.score(X_train,y_train):0.3f}")

         Accuracy of Test Dataset: 0.855
         Accuracy of Train Dataset: 0.858
```

Vale Prediction for Test dataset for SVM

```
In [31]: print("Predicted values:")
         print(y_predict)
         cal_accuracy(y_test, y_predict)

         Predicted values:
         [1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 1 0 0 0 1 0 0 1
          0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 1 0 0 1
          1 0]

         Confusion Matrix:
          [[34  6]
          [ 6 30]]

         Accuracy : 84.211
```

Fig 3.6.2 Support Vector Machine code

3.6.3 Decision Tree

Decision Tree

In [32]:
```
from sklearn.tree import DecisionTreeClassifier
gini = DecisionTreeClassifier(criterion = "gini", random_state =100,max_depth=3, min_samples_leaf=5)
gini.fit(X_train, y_train)
warnings.simplefilter('ignore')
print(f"Accuracy of Test Dataset: {gini.score(X_test,y_test):0.3f}")
print(f"Accuracy of Train Dataset: {gini.score(X_train,y_train):0.3f}")
```

```
Accuracy of Test Dataset: 0.697
Accuracy of Train Dataset: 0.831
```

Vale Prediction for Test dataset for Decision Tree

In [33]:
```
y_predict=gini.predict(X_test)
print("Predicted values:\n")
print(y_predict)
cal_accuracy(y_test, y_predict)
```

```
Predicted values:

[1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 0 0 1 1 0 1
 0 1 1 1 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 1 1 1 0 1 0
 0 1]

Confusion Matrix:
 [[28 12]
 [11 25]]

Accuracy : 69.737
```

Fig 3.6.3 Decision Tree code

3.6.4 Random Forest

Random Forest

In [34]:
```
from sklearn.ensemble import RandomForestClassifier

forest=RandomForestClassifier(n_estimators=100)
forest.fit(X_train,y_train)

warnings.simplefilter('ignore')
print(f"Accuracy of Test Dataset: {forest.score(X_test,y_test):0.3f}")
print(f"Accuracy of Train Dataset: {forest.score(X_train,y_train):0.3f}")
```

```
Accuracy of Test Dataset: 0.855
Accuracy of Train Dataset: 1.000
```

Over Fitting Issue

Vale Prediction for Test dataset for Rondom Forest

In [35]:
```
y_predict=forest.predict(X_test)
print("Predicted values:\n")
print(y_predict)
cal_accuracy(y_test, y_predict)
```

```
Predicted values:

[1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 1 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 1 0 1 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 5 31]]

Accuracy : 85.526
```

Fig 3.6.4 Random Forest code

Chapter IV – Results, Interpretation of results, inferences from the results and analysis

4.1 Results

Out[37]:

	Model	Traning Accuracy	Test Accuracy
1	SVM	0.857778	0.855263
3	Random Forest	1.000000	0.855263
0	Logistics Regression	0.848889	0.842105
2	Decision Tree	0.831111	0.697368

Figure 4.1. Performance Comparisons of all algorithms

```
Predicted values:
[1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 6 30]]

Accuracy : 84.211
```

Figure 4.2. Prediction using SVM

```
Predicted values:

[1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 1 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 5 31]]

Accuracy : 85.526
```

Figure 4.3. Prediction using Random Forest

```
Predicted values:
[1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 1
 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 1
 1 0]

Confusion Matrix:
 [[34  6]
 [ 6 30]]

Accuracy : 84.211
```

Figure 4.4. Prediction using Logistic Regression

```
Predicted values:

[1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1
 0 1 1 1 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 1 1 0 1 0
 0 1]

Confusion Matrix:
 [[28 12]
 [11 25]]

Accuracy : 69.737
```

Figure 4.5. Prediction using Decision Tree

4.2 Interpretation of results: Therefore the best algorithm for the taken dataset is “Logistic Regression” from Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM) as it has the makes less errors and accuracy is also not very behind the others.

4.3 Tableau representation

4.3.1 bar chat of the data

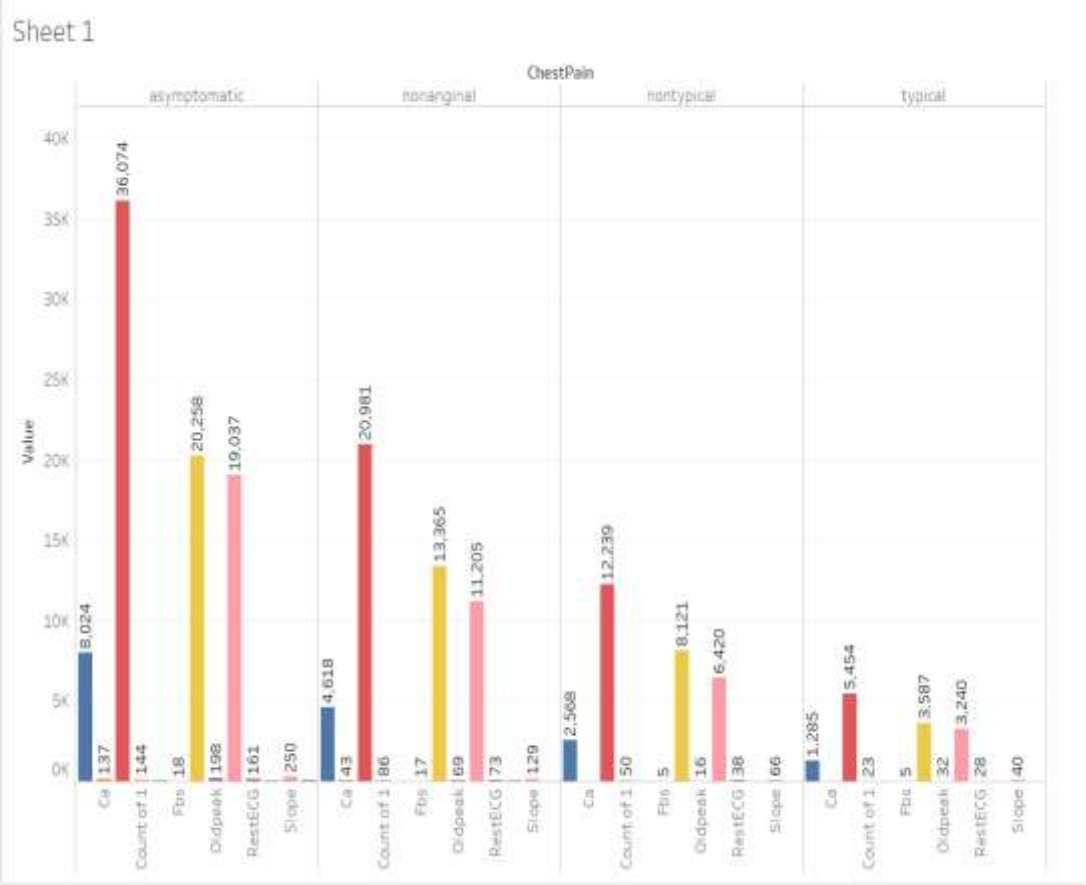


Figure 4.6 bar chart in tableau of the data

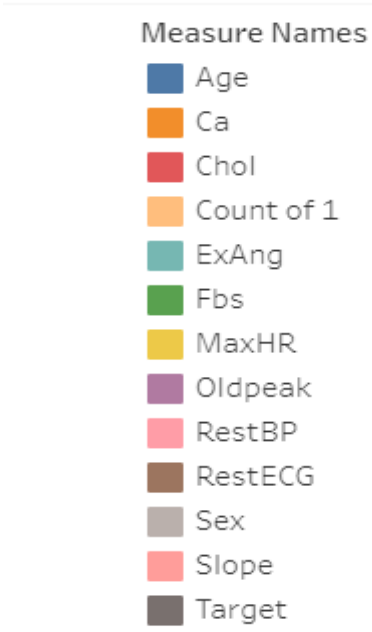


Figure 4.7 Representation of color indication in the (fig 4.6) bar chart

Explanation: This chart represents the given column's in (Fig 4.7) bar chart with respect to "ChestPain" column.

4.3.2 coloumn chat of the data

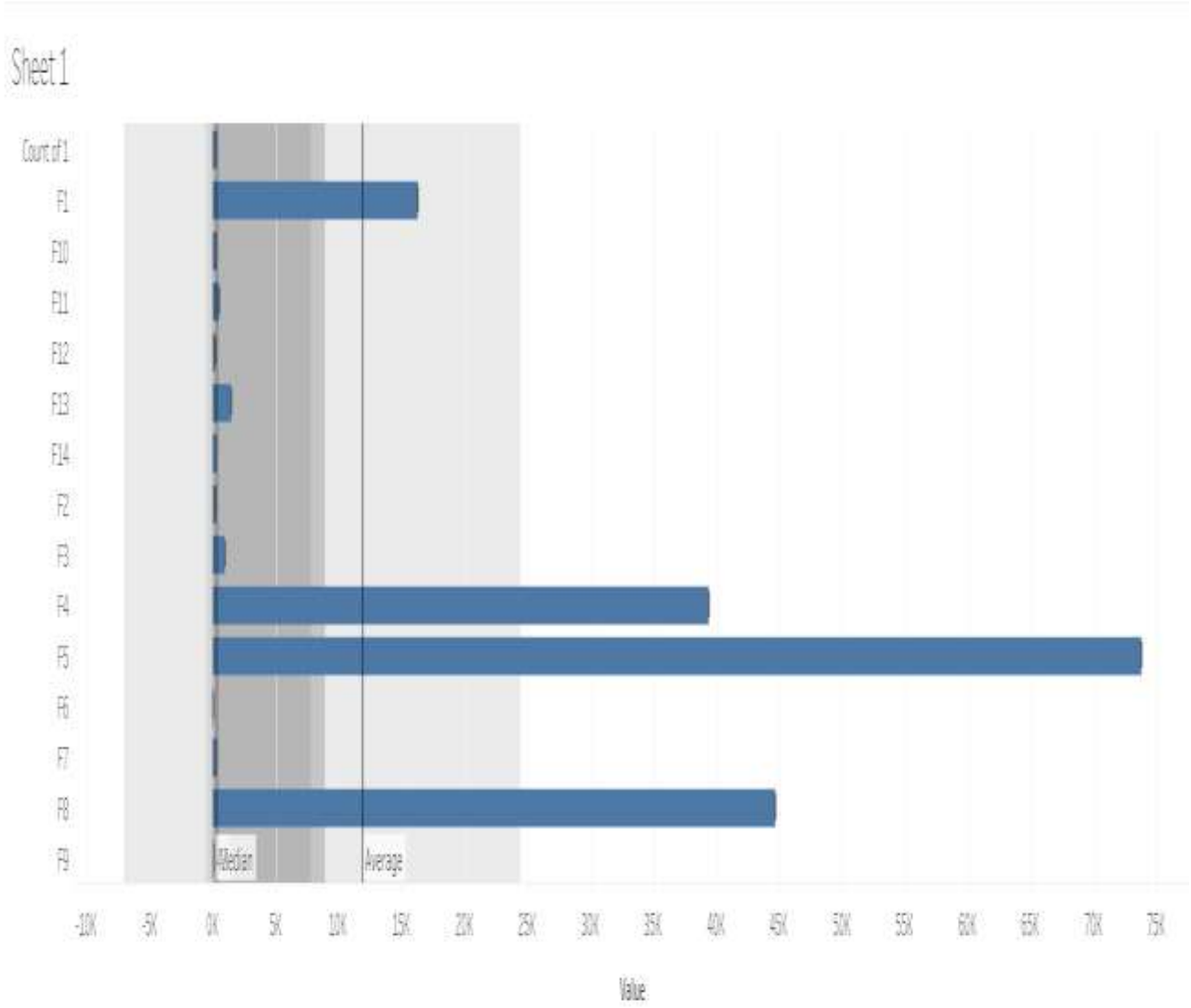


Figure 4.8 Column chart in tableau of the data

Explanation: This chart represents the Mean and median value with respect to the F1 to F9 column.

4.3.3 bar chat of the data

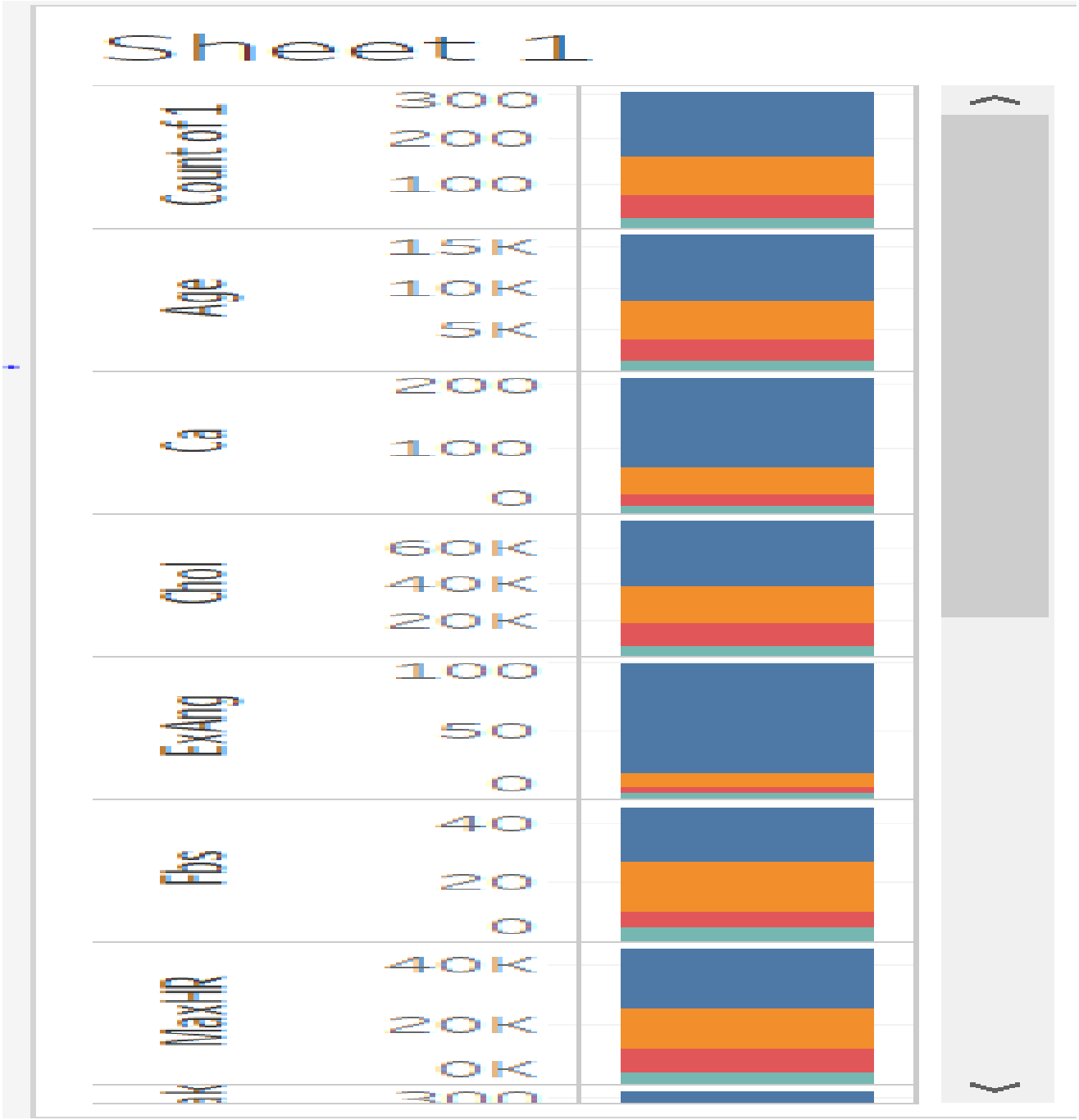
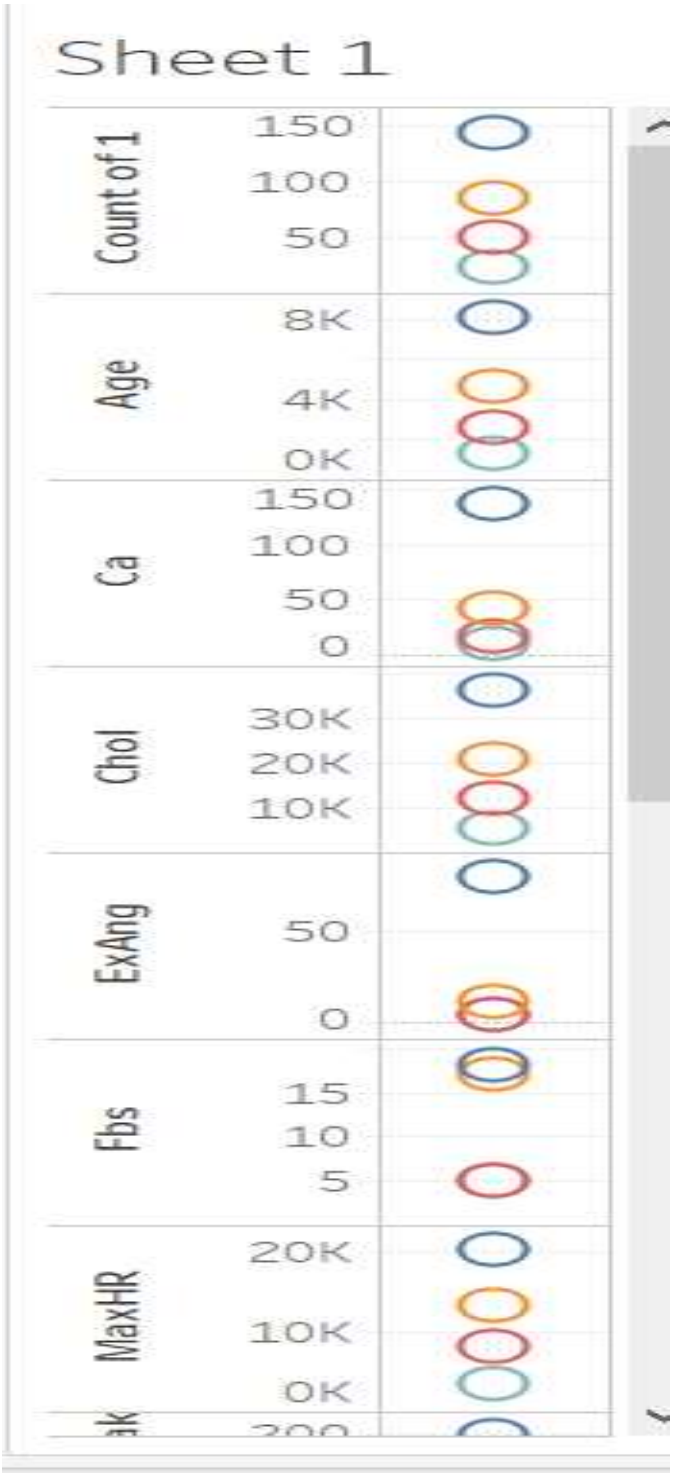


Figure 4.9 Bar chart in tableau of the data

Explanation: This bar chart represents “ChestPain” with different column’s of the dataset given in Fig 4.7.

4.3.4 Boxplot and circle view of the data



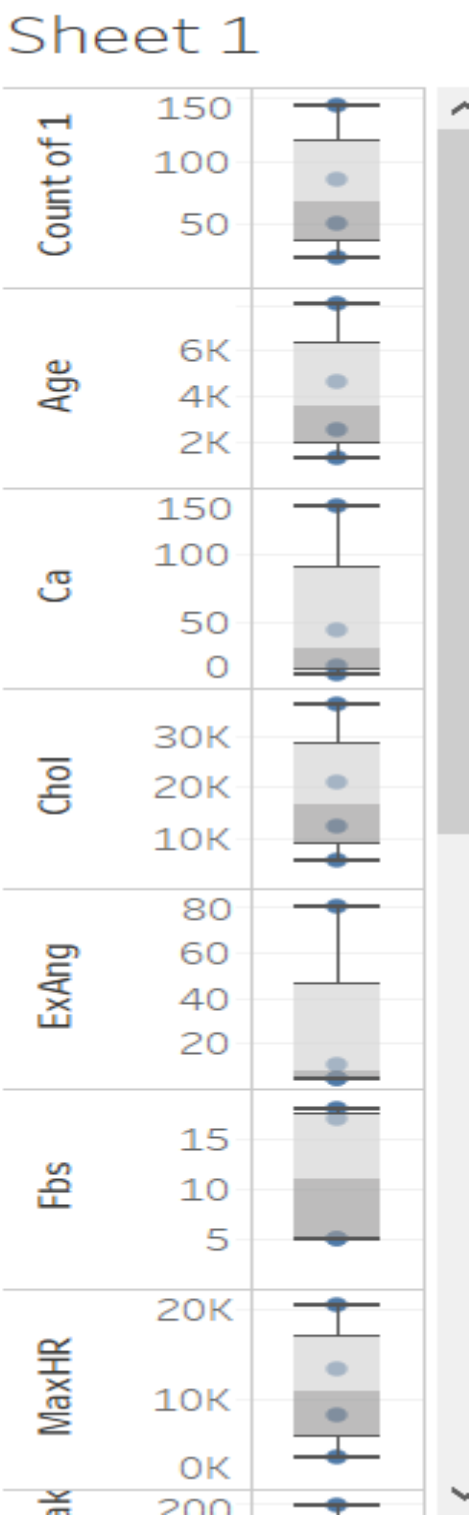


Figure 4.10 Circle view in tableau of the data

Figure 4.11 Box Plot in tableau of the data

Explanation(4.10): This is the Circle view of all the column's with respect to "ChestPain".

Explanation(4.11): This is the Box Plot of all the column's with respect to "ChestPain".

4.3.5 column chart of the data

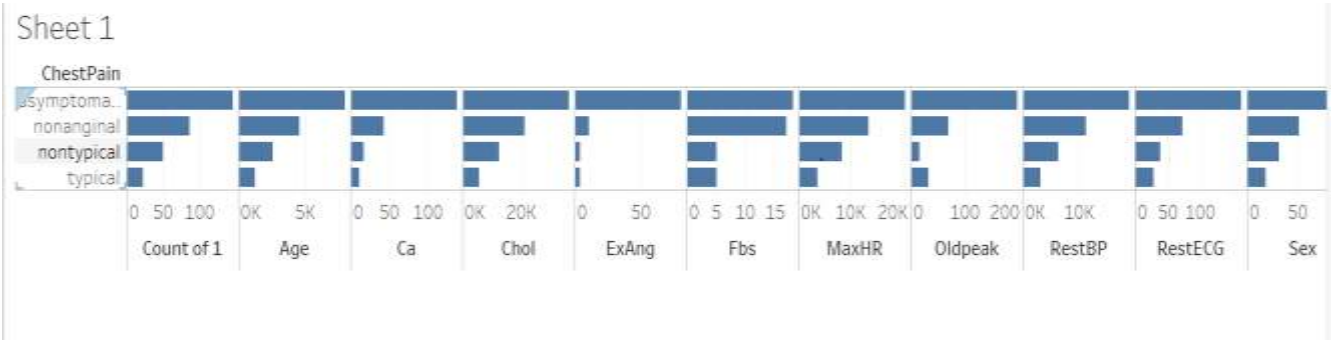


Figure 4.12 2nd column chart in tableau of the data

Explanation(4.12): This is the Historical Bars of all the column's with respect to "ChestPain".

4.3.6 circle view of the data

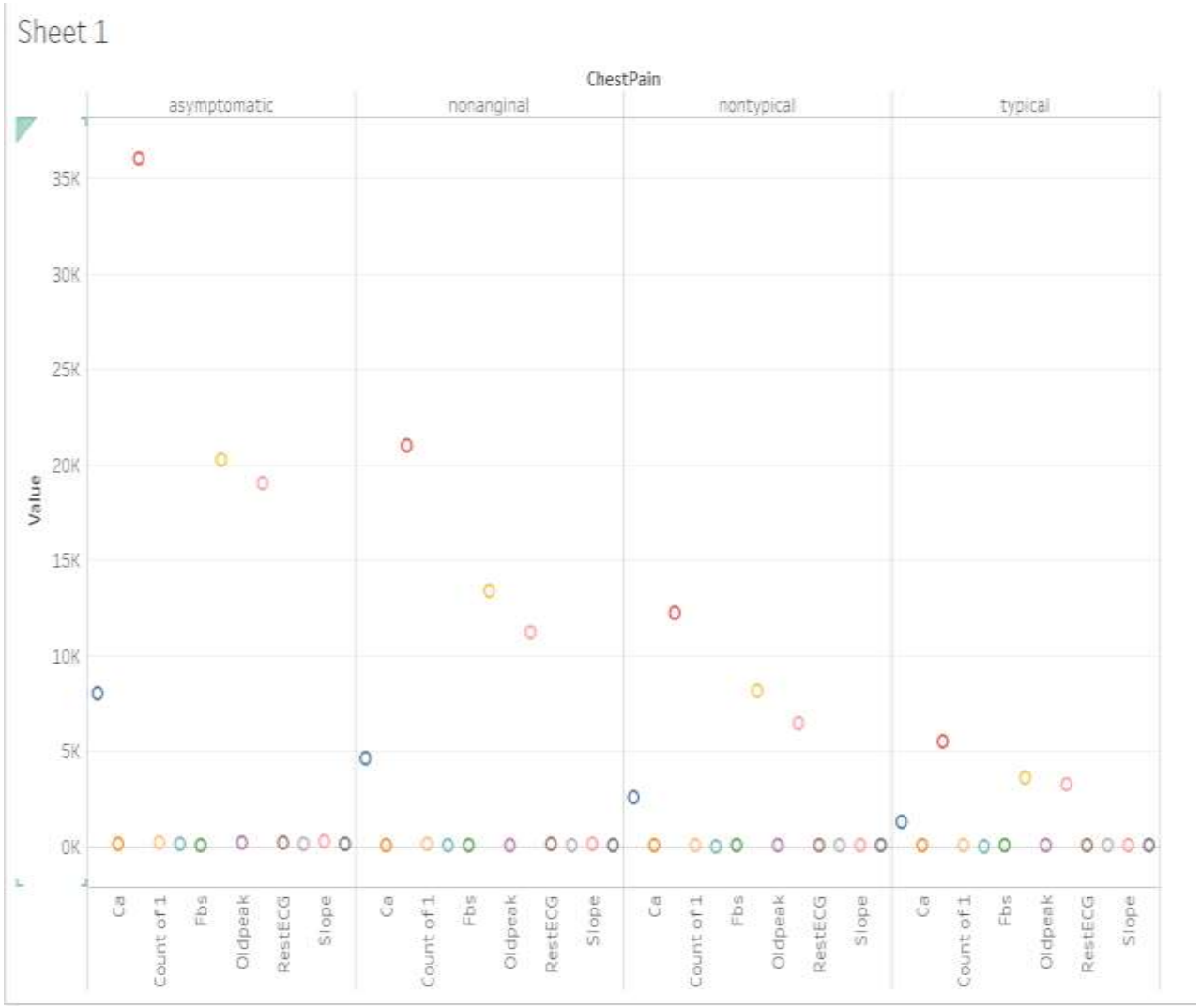


Figure 4.13 side-by-side circle view in tableau of the data

Explanation(4.13): This is the Side-By-Side Circle view of all the column's with respect to "ChestPain".

4.3.7 different arrangements of circle view of the data

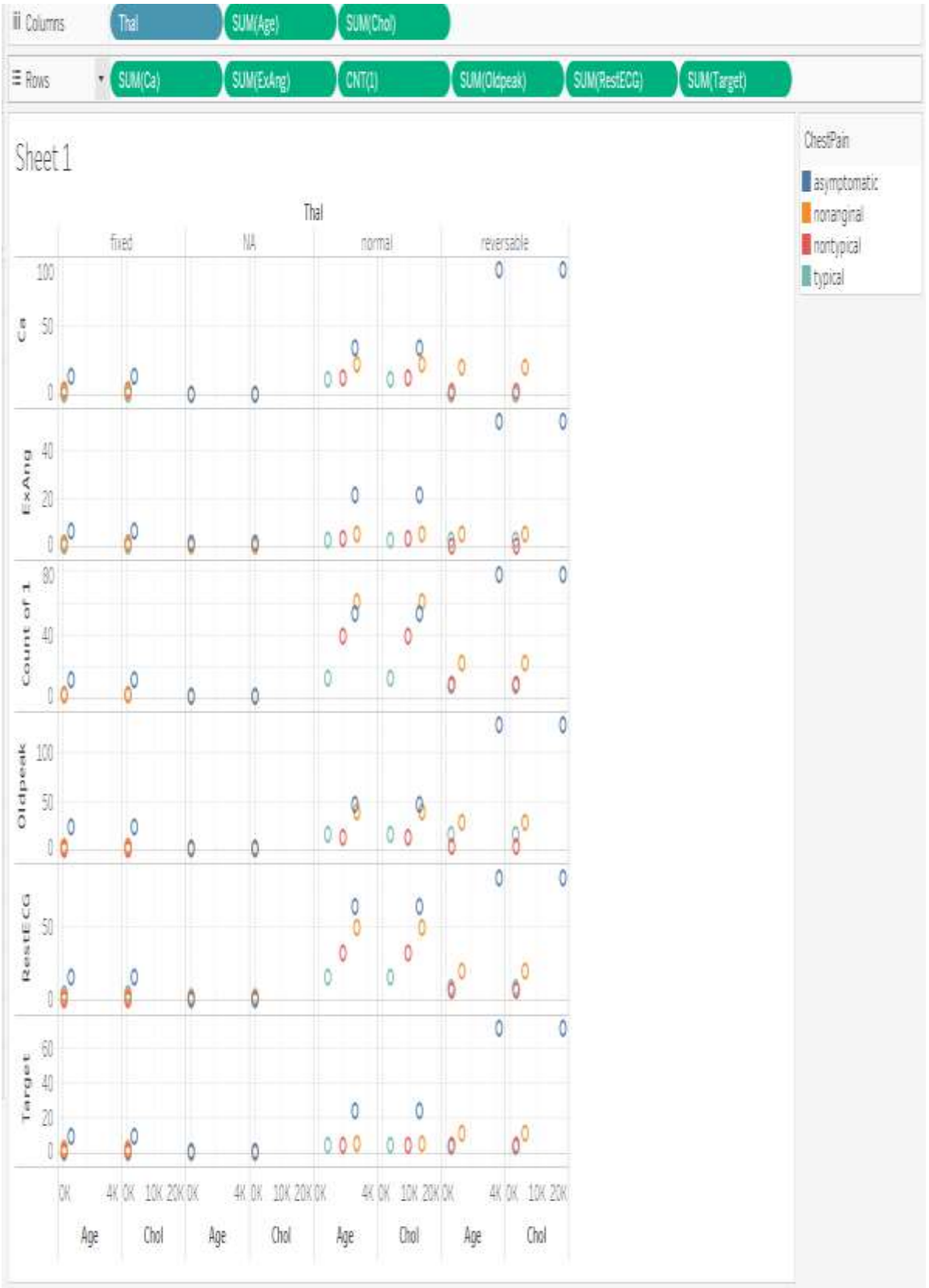


Figure 4.14 Different arrangement of side-by-side in tableau of the data

Explanation(4.14): This is the Side-By-Side Circle view of all the column's with respect to "ChestPain" , "Thalassemia" , "Cholesterol" & "Age" .

4.4 POWER BI Representation

4.4.1 stacked area chart

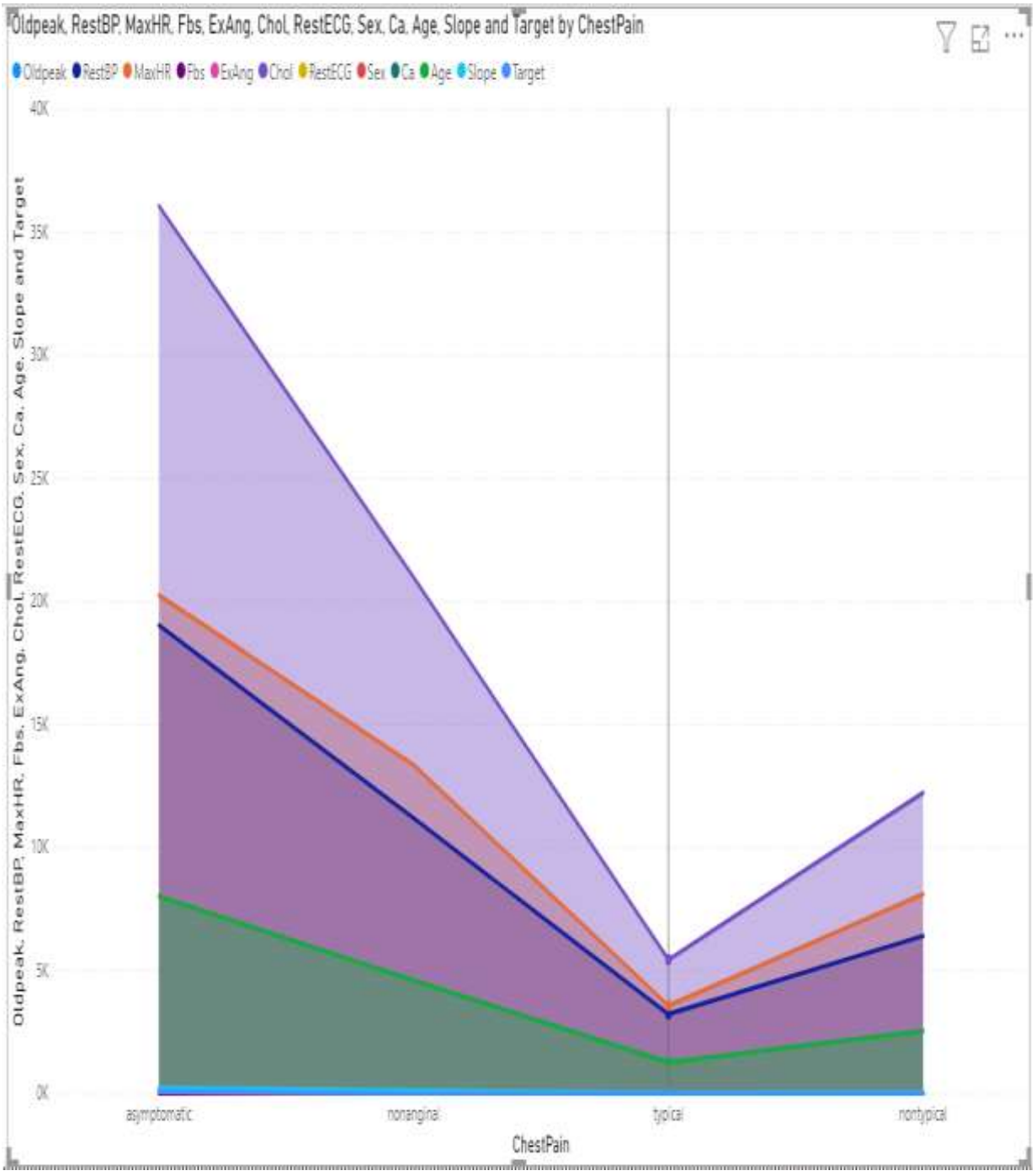


Figure 4.15 Stacked area chart of the data in Power BI

Explanation(4.15): This is the Stack Area Chart of all the column's with respect to "ChestPain" in different software known as PowerBI.

4.4.2 elaborated stacked area chart

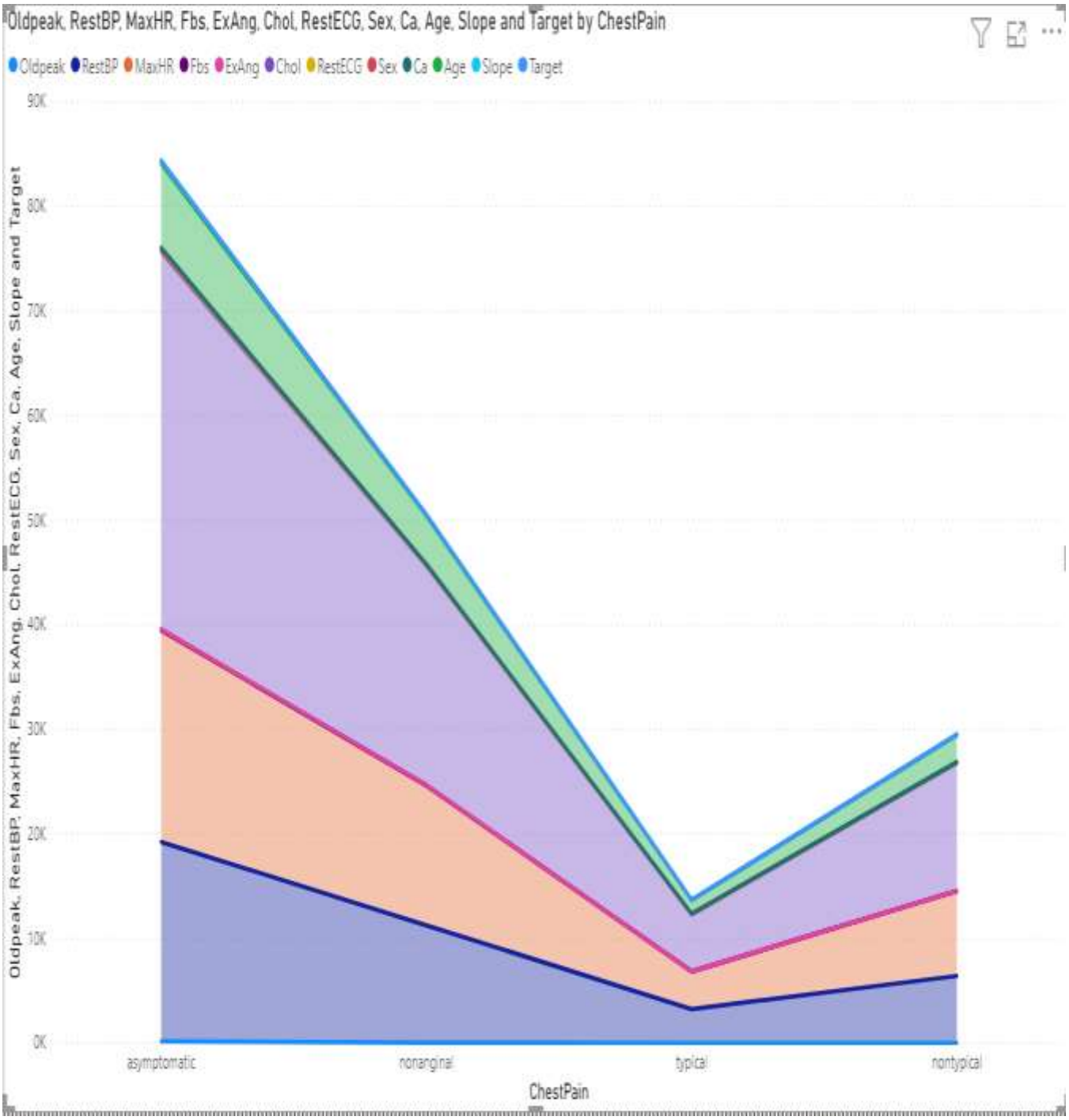


Figure 4.16 Stacked area chart of the data in Power BI

Explanation(4.15): This is the elaborated Stack Area Chart of all the column's with respect to "ChestPain" in different software known as PowerBI.

5. Chapter V – Summary, Conclusions and scope for Further Study

5.1 Summary

Finally, we carried out an experiment to find the predictive performance of different classifiers. We select four popular classifiers considering their qualitative performance for the experiment. Naïve base classifier is the best in performance. In order to compare the classification performance of four machine learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results in table 1, it can be concluded that Logistic Regression is the best as compared to Support Vector Machine, Decision Tree and Random Forest. After analyzing the quantitative data generated from the computer simulations, Moreover their performance is closely competitive showing slight difference. So, more experiments on several other datasets need to be considered to draw a more general conclusion on the comparative performance of the classifiers.

5.2 Conclusions

By analyzing the experimental results, it is concluded that Logistic Regression technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least total time to build. We can clearly see that highest accuracy belongs to Logistic Regression algorithm with reduced error.

In conclusion, as identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

5.3 Scope for Further Study

I have only used 4 basic machine learning algorithms i.e. for the dataset we can apply new machine learning algorithms see the see and compare our results like few of them may be k-nearest neighbors' algorithm, artificial neural networks (ANNs), genetic algorithm (GA) so many other algorithms are available so we can use those and compare among them to get the best algorithm suitable for the dataset.

6. REFERENCES

[1]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200

[2]. B.L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm " M.Akhil jabbar* International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) Volume 30, Number 2 – 2015 pp. 2-5 2013.

[3]. Michael W.Berry et.al,"Lecture notes in data mining",World Scientific pp150-155 (2006)

[4]. S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis : Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[5]. C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[6]. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. 10Y. C. T. BoJin, "Support vector machines with genetic fuzzy feature transformation for biomedical data classification.," Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[7]. N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

[8]. A. E. Hassanien and T. Kim, "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks," J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec.2012.

[9]. Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Professor, Orissa Engineering College, Bhubaneswar, Odisha – India. Volume 27, Number 4 pp. 340–384, Mar. 2015

[10]. UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machinelearningd atabases/ statlog/german/

[11]. Domingos P and Pazzani M. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp105-112, 1996.

[12]. Elkan C. "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, University of California, San Diego, USA, pp190-220, 1997.

[13]. B.L Deekshatulua Priti Chandra "Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. "Empirical Evaluation of Classifiers‟ Performance Using Data Mining Algorithm" International Journal of C omputer Trends and Technology (IJCTT) – Volume 21 Number 3 – Mar 2015 ISSN: 2231-2803