# **Project**

# **Heart Disease Prediction**



**Project Summary**

The purpose of this project is to predict outcomes of possible heart diseases,

This project is about the prediction of heart disease based on a set of parameters using data retrieved with the API for <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>.

If there is Diagnosis of heart disease then it is considered as 0 i.e. no disease and 1 indicates that the patient may have a heart disease. Our goal is to classify heart disease and find the model with highest accuracy scores along with the minimum value of False negatives as it will be dangerous to predict a patient with disease as no disease.

For this purpose, a classification model is used. After trying and testing with various classification algorithms like Classification trees, Neural networks, Bagging and Boosting techniques along with grid search. It is observed that XG Boost is an appropriate model for this case because the output is more generalized and gives the highest accuracy i.e 91.46% and minimum misclassification i.e. 8.54%.

**Introduction**

Over the last few decades, extreme changes in people's lifestyles have shifted human societies away from farmed foods and active lifestyles and toward fast foods and inactive lifestyles. Heart disease, often known as cardiovascular disease (CVD), is a condition that affects people all over the world and affects their hearts and blood vessels. It is the leading cause of death and inefficiency in the United States and several European countries. Heart diseases (CVDs) are the leading cause of death worldwide, killing an estimated 17.9 million people each year, accounting for 31% of all deaths. CVDs are a common cause of heart failure, and this dataset contains 11 variables that can be used to predict heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to one or more risk factors such as hypertension, diabetes, hyperlipidemia, or prior disease) require early detection and management, which a machine learning model can deliver.

**Main Chapter**

**Steps of data mining process:**

1. **Develop Understanding**

The main purpose of our project is to determine whether a patient has heart disease so that doctors can take preventive actions to decrease the impact or delay the onset of life-threatening symptoms.

Our analysis aims to shed light on the numerous indicators that can signal the occurrence of CVDs. We believe that this will serve as a tool for tracking heart disease risk and enabling people to take preventative actions through lifestyle changes.

1. **Obtain Data for Analysis**

The heart.csv file is based on data retrieved from the Kaggle dataset mentioned above. This dataset was created by combining four different datasets of Cleveland, Hungary, Switzerland, and Long Beach for research purpose.

**Data attribute information**

| **Feature** | **Description** |
| --- | --- |
| Age | Age of the patient |
| Sex | Gender of patient (1 = male; 0 = female) |
| chest pain type | 1. Typical Angina 2. Atypical Angina 3. Non-Anginal Pain 4. Asymptomatic |
| trestbps | Resting blood pressure (in mm Hg on admission to the hospital) |
| chol | Serum cholesterol in mg/dl |
| fbs | Fasting blood sugar > 120 mg/dl -(1 = yes; 0 = no) |
| restecg | Resting electrocardiographic results (values 0,1,2)  0: Normal 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina -(1 = yes; 0 = no) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | The slope of the peak exercise ST segment  0: upsloping 1: flat 2: downsloping |
| ca | Number of major vessels (0-3) colored by fluoroscopy |
| thal | Results of nuclear stress test  0 - unknown  1 - normal  2 - fixed defect  3 - reversible defect |
| target | Diagnosis of heart disease - 0 = no disease and 1 = disease |

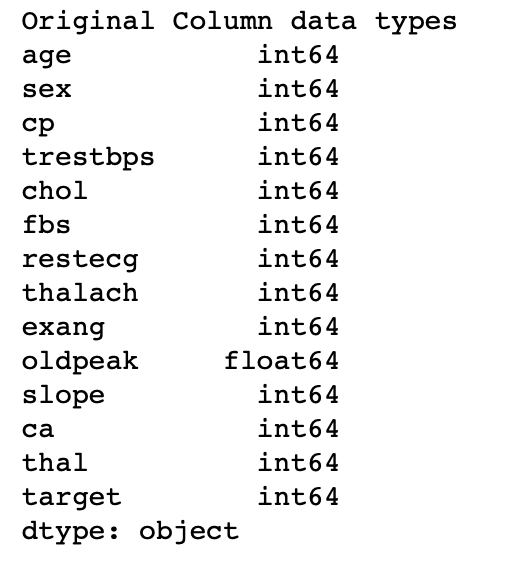
**4. Explore, Clean and Preprocess Data; Reduce Data Dimension:**

**Data overview:** The data set contains 1025 entries and a total of 14 attributes.There are no missing values and all columns are integer and binary type except oldpeak which is of float data type.

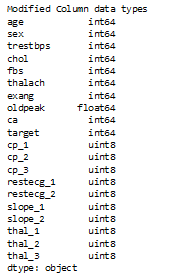


Some of the columns such as cp, restecg, slope, thal are converted to dummy variables to be used in our analysis.

Data types and column names before conversion:

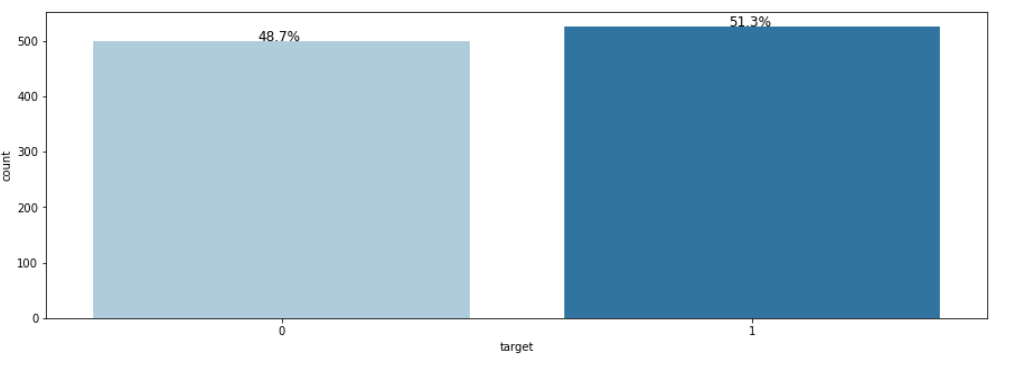


Data types and column names after conversion:



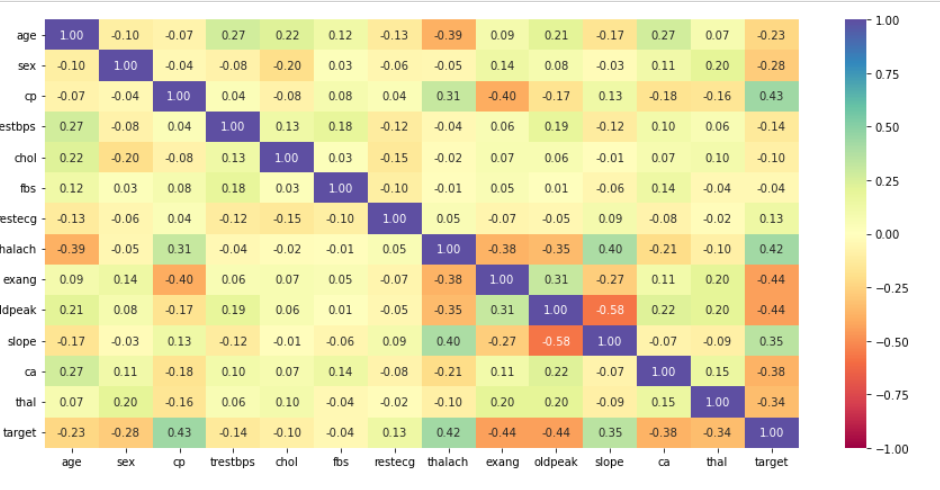
To perform **exploratory data analysis**, we plotted histograms and bar charts to get the understanding of an individual attribute and heatmap to find the relationship between the variables

**Findings from univariate analysis:**



* For target variable output seems to have equal classes. i.e. in 51.3% of patients no heart disease detected for 48.7% heart disease got detected.
* Similarly, based on the charts from the notebook, we have 30.4% female and 69.6% male patients in the dataset.
* Around 33% patients seem to have exercise angina

**Findings from bivariate analysis:**



The above heatmap shows the relation between each variable with every other variable as well as the target variable, based on this chart we found below insights:

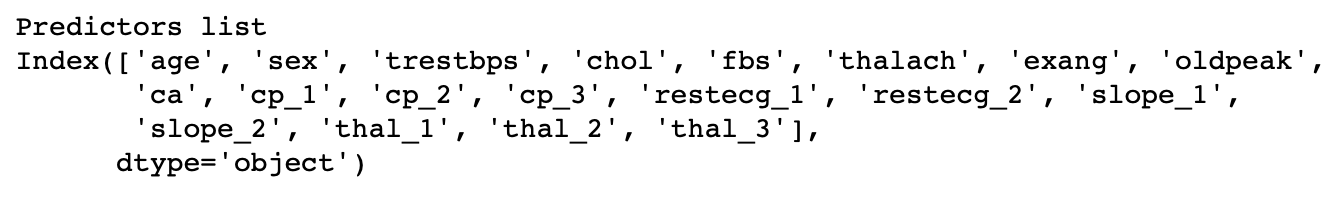
* The target variable is positively correlated with cp, thalach, and slope variable
* The target variable is negatively correlated with exang, oldpeak, ca, thal variable.
* Oldpeak is has a negative correlation with slope variable and the target variable
* Slope has a strong positive relation with thalach

All the predictors are relevant to our outcome variable, and hence we have all the 14 predictors put to use, eliminating the need for data reduction.

**5. Determine the Data Mining Task**

There are two potential outcome variables for this data set that is indicated in the Target column, where ‘0’ indicates no disease and ‘1’ indicates heart disease. Hence, we will use classification algorithms to predict whether a patient has a heart disease or not.

• The predictor variables are listed below:



**6.Partition Data**

• In order to avoid an overfitting situation, we use partitions to develop our data by using train\_test\_split with test-size at 40%. While we will split the data into ‘Training’ and ‘Validation’, the training partition contains 60% of the data to develop the model along with validation contains 40% of data to evaluate the performance on the new dataset.

**7. DM techniques/models:**

There are both numerical and categorical variables in the data set. Target variable (diagnosis of heart disease) is the outcome variable and the other 13 variables are predictor variables. The output variable is a categorical variable. So, Classification trees, random forest trees, boosted trees and neural network algorithms are applied to predict the outcome. Our goal is to use multiple algorithms to categorize the target variable and then choose the best approach based on the accuracy measures from the confusion matrix.

**8. Selected techniques and associated algorithms:**

**Classification Tree Models:**

Tree model where the target variable can take categorical variables are called classification trees. Natural end of the tree growing process is 100% purity in each leaf. This overfits the data, which ends up fitting noise in the data. Overfitting leads to low predictive accuracy of new data. Tree with the best split results in best accuracy.

The two classification tree techniques used in this analysis are

1. Smaller Classification tree using random controlled parameters
2. Smaller Classification tree using Grid Search results

**Smaller Classification tree using random controlled parameters:**

In order to avoid overfitting, the growth of the tree is limited by various control parameters in DecisionTreeClassifier () function.

In DecisionTreeClassifier () we can control tree developments using a variety of control parameters

* Maximum tree depth – number of splits, e.g. max\_depth = 30
* Minimum impurity decreases per split, e.g. min\_impurity\_decrease = 0.01
* Minimum number of a node records for splitting, e.g., min\_samples\_split = 20

Small Classification tree with controlled parameters are shown below.

Timeline

Description automatically generated

The confusion matrices of training and validation partition of a smaller classification tree with control parameters are shown below.

Graphical user interface, text, application, letter, email

Description automatically generated

The confusion matrices for the training and validation partitions have an accuracy of 91.54% for the training partition and 84.15% for the validation partition. Thus, the trained classification tree model fits well for the validation date set. The misclassification rate for the training partition is 1-0.9154= 0.0846 or 8.46%, and for the validation partition 1-0.8415=0.1585 or 15.85%.

**Smaller Classification tree using Grid Search results**:

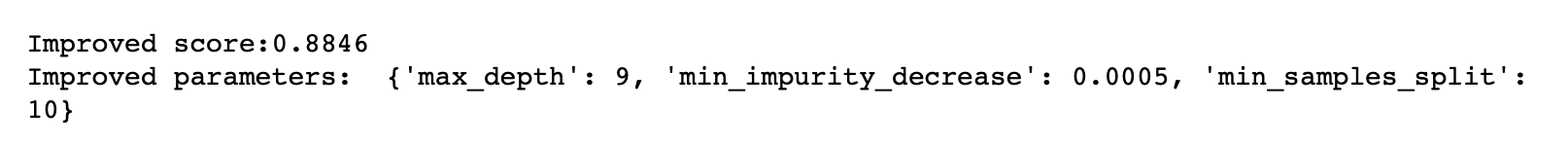
Main problem with control parameters is to identify the optimal control parameters. So, we apply exhaustive grid search over a combination of different parameter values,

Maximum tree depth in range of [2, 20]

Minimum impurity decreases as a combination of [0, 0.0005, 0.001]

Minimum number of records in a node for split in range of [10, 30]

The exhaustive grid search allows to find the combination that leads to the tree with the highest accuracy and smallest error like lowest misclassification level

The control parameters obtained from grid search are shown below.Classification tree with controlled parameters using grid search are shown below.A picture containing timeline

Description automatically generatedThe confusion matrices of training and validation partition of a smaller tree with control parameters using grid search are shown below.

Graphical user interface, text, application, email

Description automatically generated

The confusion matrices for the training and validation partitions show an accuracy of above 90% and thus the trained classification tree model fits well for the validation date set. The misclassification rate for the training partition is 1-0.9593= 0.0407 or 4.07%, and for the validation partition 1-0.9171=0.0829 or 8.29%.

The accuracy for validation partition for the model using grid search is 91.71% which is greater than the model with controlled parameters which has accuracy of 84.15 %. So, the classification tree using grid search results seems to be a better one.

**Random Forest**

The random forest approach is using the averages of multiple estimates which are more reliable than a single estimate. This uses the bootstrap resampling where a sample record is drawn and recorded back in the sample before the next sampling records are drawn.

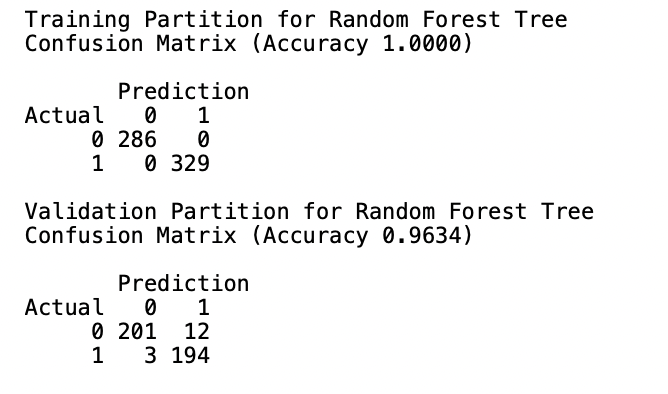
Random trees use a random subset of predictors to fit a classification or regression tree to each sample, and thus obtain a “forest” of trees. With the ensemble method that uses the voting for classification and averaging for predictions, the improved predictions can be obtained by combining the classifications/predictions of individual trees resulting in multiple tree results.

We have developed a random forest with 1000 randomly developed trees. We get the no.of nodes for each tree with rf.estimators\_[0].tree\_.node\_count where 0 denoted the first randomly developed tree and so on.

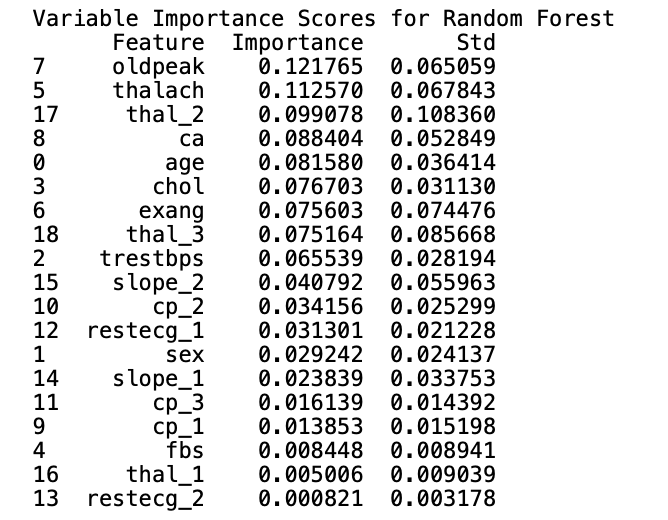
Below is the number of nodes for first random forest tree



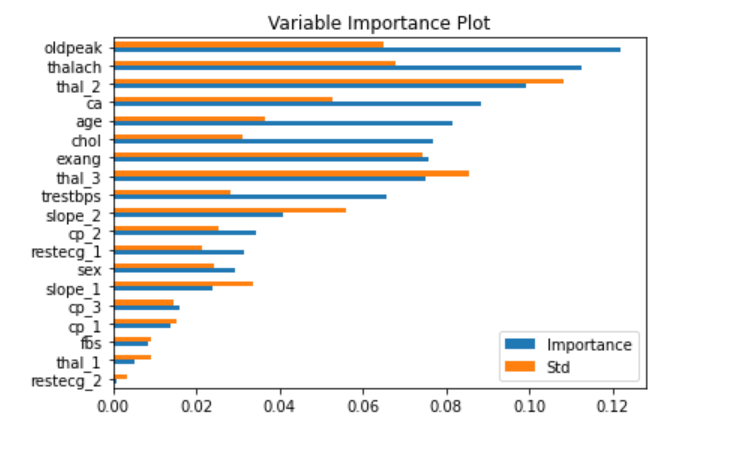
As for checking the possibility of overfitting, we use the confusion matrix for training and validation sets’ accuracies. The confusion matrix is displayed below.



For identifying the importance of each variable on the classification tree, we use the importance of each variable which helps in finding the utilization of each variable in building the classification tree and the standard deviations of the estimators. The below image displays the variable important scores and their standard deviations.



The below plot diagram gives a pictorial representation of variables’ importance to that of their standard deviations.

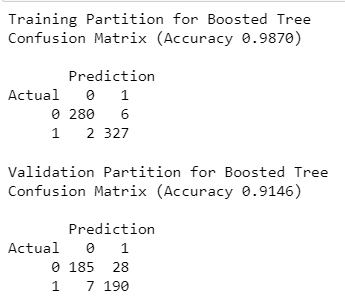
****

**XGBoost Algorithm:**

In XGBoost model we can tune the accuracy by a variety of control parameters

Number of estimators - n\_estimators=150, learning\_rate=0.05, subsample=0.7.

The confusion matrices of training and validation partitionXGBoost algorithm with control parameters are shown below.

****

The confusion matrices for the training and validation partitions have an accuracy of 98.70% and 91.46% respectively.. The misclassification rate for the training partition is 1-0.9829= 0.0171 or 1.71%, and for the validation partition 1-0.0.9854=0.0146 or 1.46%.

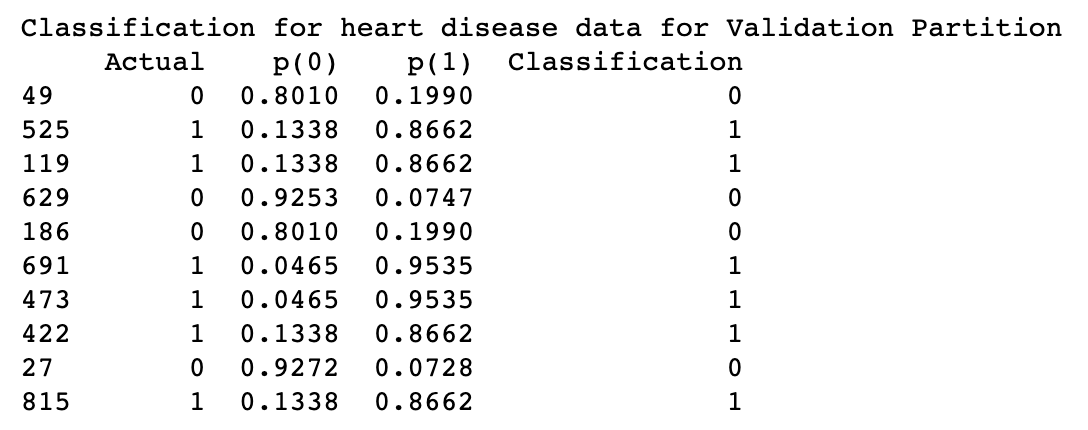
**Neural networks**

Different models of neural networks:

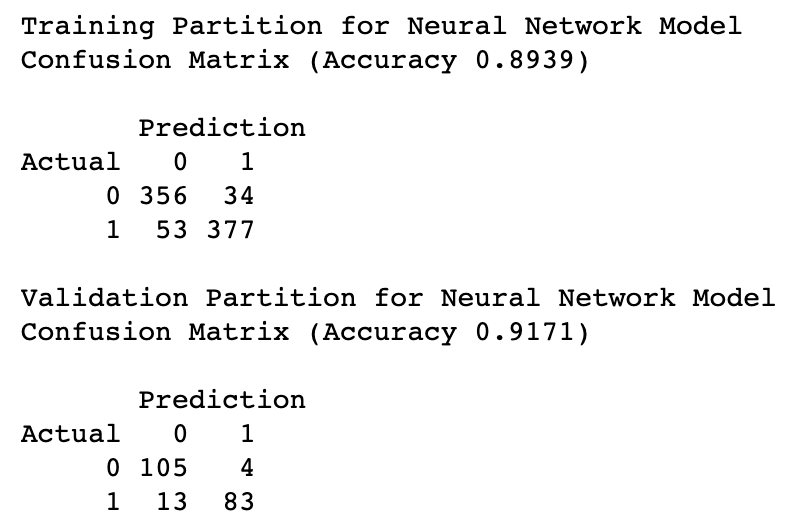
1. Neural Network models using Multiple predictors.
2. Grid Search-based neural network model.

In this data set since the target variable is to predict whether a person might have heart disease or not we used function.

Results of the neural network training for the validation partition. We produce predictions for the validation partition for the records based on the trained data. The set of probabilities for various outcomes, such as prediction 0 for no heart disease and prediction 1 for heart disease. Identifying the likelihood, with p(0) denoting the absence of heart disease and p(1) denoting the presence of heart disease. The categorization results for validation partitions for the first 10 records are shown in the data frame below.



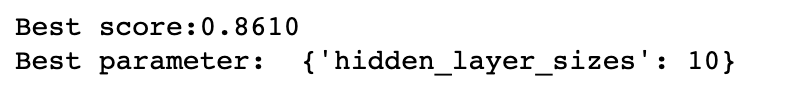
To be more specific, we wanted to evaluate if there was any overfitting in this situation, as well as how accurate the training and validation partitions were. The confusion matrix is displayed below.



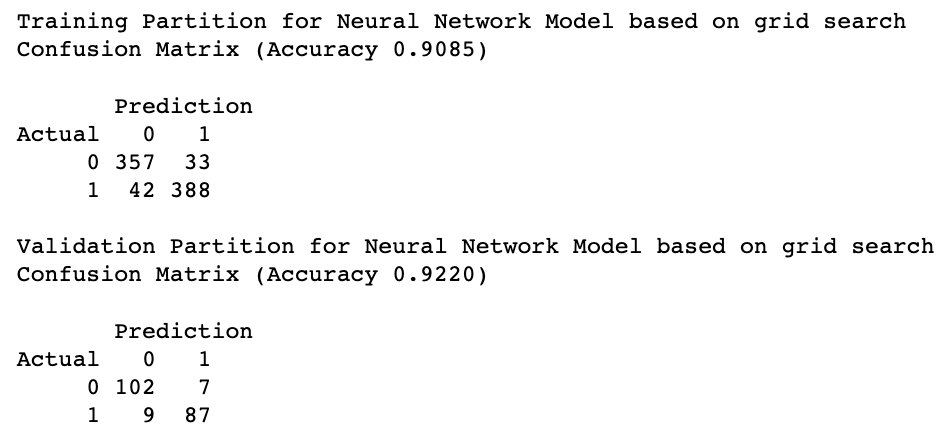
According to the confusion matrix, there is no overfitting, and the training partition's accuracy is 89.39 percent, while the validation partition's accuracy is 91.71 percent, indicating good overall accuracy and higher accuracy than the training partition.

**Neural Network model using Grid Search:**

We started by determining how many nodes we want in this model. Then we used the GridSearchCV() function to fit it into a trained partition using cross validation of 5 partitions, 20% of the original dataset, and the maximum number of feasible jobs. With these, the best score is 86.1 percent, which is the most accuracy we can attain with the optimal number of nodes we have i.e. 10.



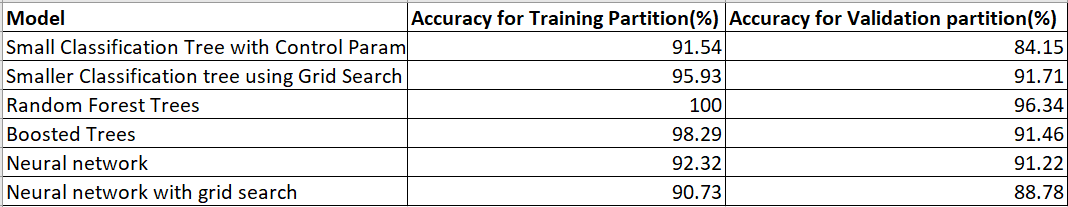
Below is the confusion matrix for Grid Search based neural network model. According to the confusion matrix, there is no overfitting, and the training partition's accuracy is 90.8 percent, while the validation partition's accuracy is 92.2 percent, indicating good overall accuracy and higher accuracy than the training partition.



Neural Network summarizes that using Grid Search we are able to predict heart disease with the accuracy of 92% in comparison to the Neural network using multiple predictors.

1. **Interpretation of the results:**

Below table shows the accuracy scores of different classification algorithms to predict heart disease:



* All the algorithms are working fine in this case however for most of the models like classification and random forest, accuracy score is slightly higher for the training partition than the validation partition.
* In the case of neural networks, the accuracy is good for training and validation partitions.
* For the Boosted trees the scores are approximately similar for both the partitions and the Boosted trees are giving the highest accuracy. Also, XGBoost with an accuracy of around 91.46% shows promising results and overall generalized performance. Hence, we recommend using the XGBoost model for heart disease prediction.

**Appendices:**

Methaila, Aditya, Prince Kansal, Himanshu Arya, and Pankaj Kumar. "Early heart disease prediction using data mining techniques." Computer Science & Information Technology Journal 24 (2014): 53-59.

Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47, no. 10 (2012): 44-48.

Sultana, Marjia, Afrin Haider, and Mohammad Shorif Uddin. "Analysis of data mining techniques for heart disease prediction." In 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT), pp. 1-5. IEEE, 2016.

Shmueli, G., Bruce, P.C., Gedek, P., and Patel, N.R. Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python, Wiley, 2020, ISBN: 9781119549840.