3. Modelling

We have chosen three supervised-learning algorithms namely Logistic Regression, Support Vector Machines and Decision Trees to build the heart disease prediction model. All the three models were built using normal (unnormalized) dataset and the normalized dataset, the confusion matrix and accuracy was compared in the end to choose the best predictive model.

3.1 Logistic Regression

The first algorithm that we used for heart disease prediction was Logistic Regression.

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression with Unnormalized Data | 87.91 % |
| Logistic Regression with Normalized Data | 84.61 % |

Table 3.1: Accuracy comparison of Logistic Regression with Unnormalized and Normalized Data

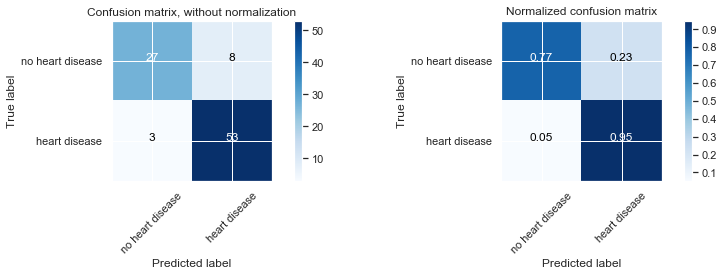


Figure 3.1: Confusion Matrix for Logistic Regression with Unnormalized Dataset

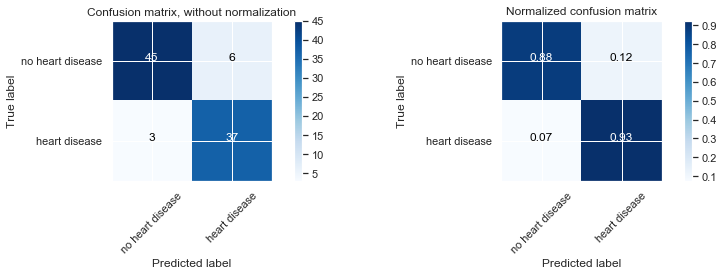


Figure 3.2: Confusion Matrix for Logistic Regression with Normalized Dataset

The main criteria that we used to evaluate the correctness of our model was the number of times the model predicted someone did not have a heart disease, but they actually had it as we are well aware that only looking at accuracy as a measure of success for the model can often be misleading. So, in healthcare where lives of people can be put into jeopardy by the slightest miscalculation, the lower this number (percentage), the more useful and better the model.

As it can be observed from the confusion matrix that in 5% cases the Logistic Regression Model with Unnormalized data predicts that someone did not have a heart disease when they actually did. This is has happened in 7 % of the cases for the Logistic Regression Model with Normalized Data. So, it is clearly visible that Logistic Regression model with Unnormalized Data set beats the Logistic Regression model with Normalized Dataset.

3.2 Support Vector Machines (SVM)

Second algorithm that was used for heart disease prediction was the Support Vector Machines (SVM). GridSearchCV was used for the selection of the various hyper parameter like C, gamma and kernel. GridSearchCV uses cross validation (3- fold by default) and runs SVM on these folds of the dataset until it arrives with the combination of kernel, C and gamma with the best accuracy. GridSearchCV iterated through set of hyper parameters that were manually given to it in form of a dictionary and found the best possible combination. The best hyper parameters for the SVM model with the unnormalized data set were C = 1, kernel = linear, gamma = auto and the best hyper parameters for the SVM model with the normalized data set were also C = 1, kernel = linear, gamma = auto. The results for both the SVM models have been summarized below.

|  |  |
| --- | --- |
| Model | Accuracy |
| SVM with Unnormalized Data | 90.10 % |
| SVM with Normalized Data | 79.12 % |

Table 3.2: Accuracy comparison of SVM with Unnormalized and Normalized Data

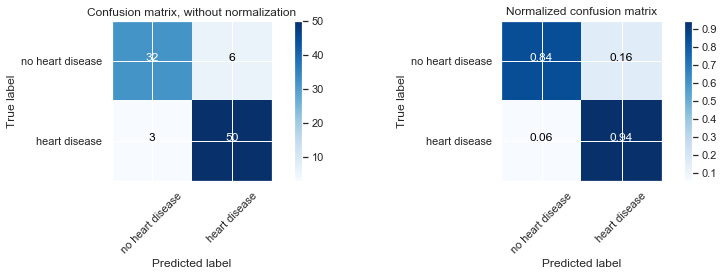


Figure 3.3: Confusion Matrix for SVM with Unnormalized Dataset

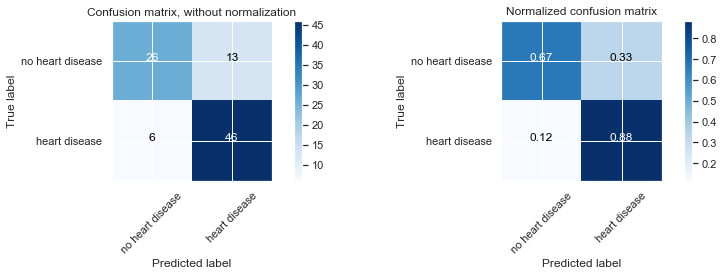


Figure 3.4: Confusion Matrix for SVM with Normalized Dataset

It is visible from the confusion matrix that in 6% cases the SVM Model with Unnormalized data predicts that someone did not have a heart disease when they actually did. This is has happened in 12 % of the cases for the SVM Model with Normalized Data. Keeping this information and the accuracy results in mind we can easily conclude that SVM model that was trained with unnormalized dataset is better at predicting heart diseases than the SVM model trained with the normalized dataset.

3.3 Decision Trees

The third and final algorithm that we tried was decision trees. We tried to tweak the maximum depth of the decision tree with the max\_depth parameter and during the process we found out that max\_depth = 3 was giving the best result and the accuracy was decreasing as we were increasing it. The results for decision trees trained with unnormalized and normalized dataset have been summarized below.

|  |  |
| --- | --- |
| Model | Accuracy |
| Decision Tress with Unnormalized Data | 86.81 % |
| Decision Tress with Normalized Data | 80.21 % |

Table 3.3: Accuracy comparison of Decision Trees with Unnormalized and Normalized Data

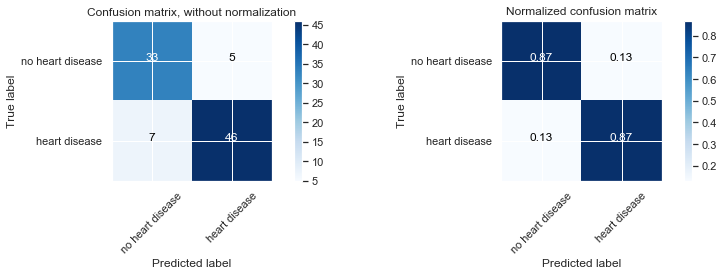


Figure 3.4: Confusion Matrix for Decision Trees with Unnormalized Dataset

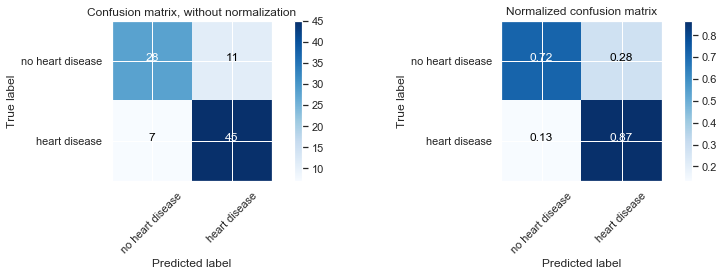


Figure 3.5: Confusion Matrix for Decision Trees with Normalized Dataset

Decision tree that was trained with unnormalized data set had a better accuracy than the decision tree that was trained with Normalized dataset. Moreover, it can be observed from the confusion matrix that in 13% cases the Decision tree Model that was trained with Unnormalized data predicts that someone did not have a heart disease when they actually did, and this is 13 % in the case of the decision tree that was trained with normalized data. So, it can be concluded that although one model has better prediction accuracy but from healthcare perspective both the models perform equally worse because of the 13 % incorrect predictions that can put a patient’s life in danger.

3.4 Conclusion

The best model for predicting the presence of the heart disease is **Logistic Regression model** that was trained with Unnormalized Data. Although the SVM model trained with unnormalized dataset had a higher accuracy (90.10 %), the Logistic Regression model is a preferred choice because the no of cases in which it predicted that someone did not have a heart disease when they actually did was 5 % as compared to 6 %in SVM.

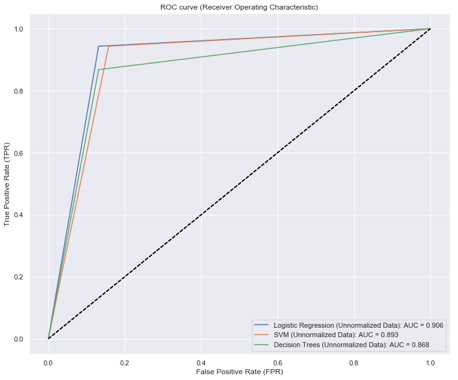


Figure 3.6 ROC Curve