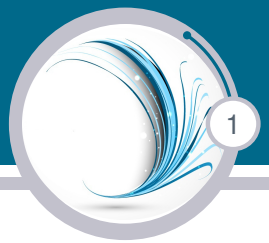
PREDICTION OF HEART DISEASE USING MACHINE LEARNING

Presented by: Biplab Mandal
Course: DSCI-6612-01
Under the guidance of Prof.Vahid Behzadan

May 3, 2021

Content



Overview of the Project

- Heart disease is the leading cause of death in the United States, causing about 1 in 4 deaths. The term “heart disease” refers to several types of heart conditions. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack.

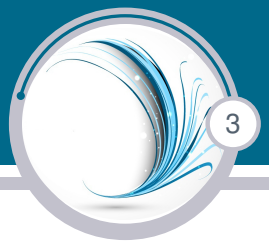
Overview of the Project

- Heart disease is the leading cause of death in the United States, causing about 1 in 4 deaths. The term “heart disease” refers to several types of heart conditions. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack.
- Experienced physician able to predict the cardiovascular disease around 67%, and most of the cases the physicians overestimate the prediction.

Overview of the Project

- Heart disease is the leading cause of death in the United States, causing about 1 in 4 deaths. The term “heart disease” refers to several types of heart conditions. In the United States, the most common type of heart disease is coronary artery disease (CAD), which can lead to heart attack.
- Experienced physician able to predict the cardiovascular disease around 67%, and most of the cases the physicians overestimate the prediction.
- Machine Learning models predict around 73% and the precision value is also around 74%.

Overview of the Dataset and Variables



https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

- The dataset has 13 features
- The dataset has 70000 samples(cardio positive:35021, cardio negative:34979)
- Correlation between feature variables
- Divided the dataset into train and test

Decision Tree Classifier

Using the Decision Tree model, we did not have a good accuracy score. The evaluation was done using Confusion matrix.

- Accuracy: 0.6346428571428572
- Precision: 0.6361829025844931
- Recall: 0.6369970140764966
- F1 score: 0.6365896980461813

Optimizing hyperparameter of a Decision Tree model using Grid search:
Optimized max depth:8
Optimized number of components:6

- We got the cross value score around 70%. So using Grid search the accuracy of the model increased and that is very effective.

Random Forest Classifier

Using Random forest classifier model we got better results than the decision tree classifier. For the evaluation we used confusion matrix.

- Accuracy: 0.7019428571428571
- Precision: 0.7213094192901041
- Recall: 0.6576329331046312
- F1 score: 0.6880009570522789

XGBoosting Classifier

Using XGBoosting Classifier we got better result than the Random forest and Decision tree. We got model score around 73%.

XGBoosting Optimaization

Using Grid search cv we were able to increase the value of the model score to 74%.

Logistic Regression

Here also we used confusion matrix to evaluate the accuracy of the model. And here also we got a very good accuracy around 72%. Also the precision value of the model is also 74%, which is very good.

Accuracy: 0.7198857142857142

Precision: 0.7400974634512058

Recall: 0.6773013150371641

F1 Score: 0.7073083353236208

We tried to enhance the accuracy of the model, but unfortunately the accuracy and precision both reduced to 50%.

Evaluation Method and Conclusion

Comparison of models

By comparing all four models, we got that the accuracy score for all of them had more than 70%, and the precision also above 70%. Cross value score for XGBoosting is 74%, which is definitely very good. Among Decision tree, Random forest and logistic regression, logistic regression has the best accuracy score and precision value.

Future Scope

For the future research, we have to enhance the accuracy and precision value above 80%, which will definitely create a big impact on diagnosis of the cardiovascular disease.

Thank you!