Heart Failure Prediction

Bikkavolu Prasanthi- B20CS010, Tammireddi Sai Unnathi- B20AI045

INTRODUCTION

According to the World Health Organization, 12 million people die each year from heart disease throughout the world. For a few years, the global burden of cardiovascular disease has been quickly growing. Many studies have been carried out in an attempt to define the most important variables in heart disease and to precisely forecast the total risk. Heart disease is also referred to as a "silent killer" since it causes mortality without causing noticeable symptoms. This study tries to predict future heart illness by evaluating patient data and using machine-learning algorithms to classify whether they have heart disease or not.

METHODOLOGY

Overview

There are various classification algorithms present out of which we shall implement the following

- XGBoost Classifier
- Gradient Boost Classifier
- Support Vector Classifier
- Logistic Regression
- Random Forest Classifier

I also made use of GridSearchCV for hyperparameter tuning along with cross-validation. And Sequential Forward Selection was used for feature selection.

Hyperparameter Tuning with Cross-Validation: *GridSearchCV*

Feature Selection: *Sequential Forward Selection*

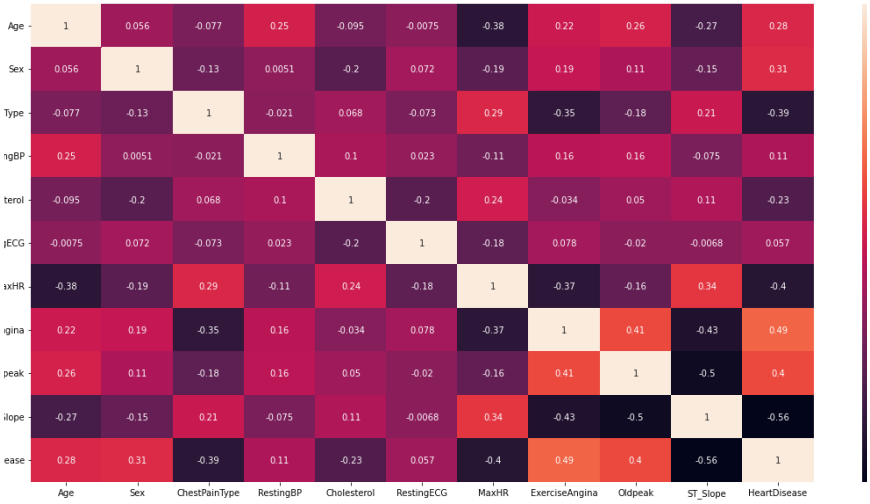
Data Description and Description Statistics are done.

Data Pre-Processing

The following columns are included in the data set: 'Age', 'Sex',' ChestPainType', ' RestingBP', 'Cholesterol', 'FastingBS', 'RestingECG', 'MaxHR, 'ExerciseAngina', 'Oldpeak', 'ST_Slope', 'HeartDisease'.

There were no missing values in the dataset. 'Sex', 'ChestPainType', 'FastingBS', 'RestingECG', 'ExerciseAngina','ST_Slope', 'HeartDisease' were initially categorical.So they have been labelled encoded. Normalization has been done. The target feature is separated The data has been split into 80% train data and 20% test data.

Fig 1.1: Correlation Heatmap for the dataset



Exploratory data analysis

From Fig 1.1, we are able to say HeartDisease depends on ExerciseAngina and Oldpeak.The three share a good correlation relatively. Apart from those, MaxHR and ST_Slope have a somewhat good correlation. Pair plots are also plotted.

Hyperparameter tuning

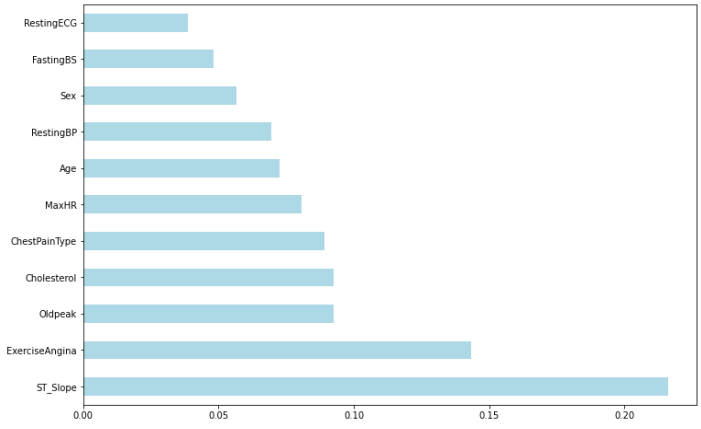
Hyperparameter tuning along with Cross-Validation has been done using GridSearchCV on all 5 models. max_depth,n_estimators,min_child_weight have been varied for XGBoost. C, gamma, and kernel have been varied for SVC.

Learning rate,max_depth, and n_estimators have been varied for Gradient Boost.
max_depth,min_sample_leaf, and n_Estimators have been varied for Random Forest. For Logistic Classification, C, solver have been varied.

Feature Selection

Sequential feature selection algorithms are a type of wrapper technique that successively adds and removes features from a dataset. This approach is known as naïve sequential feature selection because it assesses each feature independently and picks M features from N features based on individual scores. Because it does not account for feature reliance, it only works extremely seldom. In SFS, variant features are gradually added to an empty set of features until the criteria are not reduced by the inclusion of more characteristics.

	feature_idx	cv_scores	avg_score
1	(10,)	[0.8027210884353742, 0.8639455782312925, 0.829...	0.813335
2	(4, 10)	[0.8367346938775511, 0.8571428571428571, 0.843...	0.831069
3	(1, 4, 10)	[0.8571428571428571, 0.8299319727891157, 0.857...	0.836539
4	(1, 4, 5, 10)	[0.8639455782312925, 0.8299319727891157, 0.857...	0.841981
5	(1, 4, 5, 8, 10)	[0.8435374149659864, 0.8639455782312925, 0.863...	0.843304
6	(1, 4, 5, 8, 9, 10)	[0.8707482993197279, 0.8707482993197279, 0.863...	0.856947
7	(1, 2, 4, 5, 8, 9, 10)	[0.8639455782312925, 0.8843537414965986, 0.877...	0.865129
8	(1, 2, 3, 4, 5, 8, 9, 10)	[0.891156462585034, 0.8843537414965986, 0.8911...	0.877355
9	(1, 2, 3, 4, 5, 6, 8, 9, 10)	[0.8843537414965986, 0.891156462585034, 0.8979...	0.880095
10	(0, 1, 2, 3, 4, 5, 6, 8, 9, 10)	[0.8571428571428571, 0.8707482993197279, 0.891...	0.867832
11	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	[0.8299319727891157, 0.8707482993197279, 0.897...	0.861029



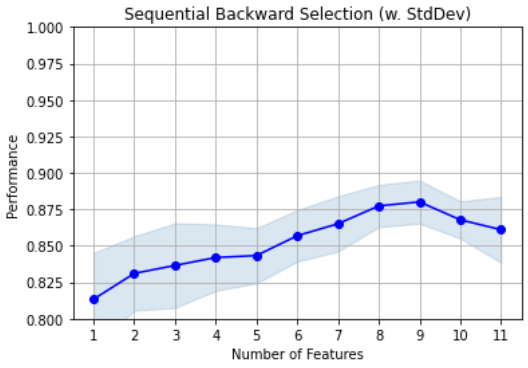


Fig 1.2: metric data frame of SBS, Feature importance plot, No. of features through SBS

The sequential forward search has been performed with a number of desired features as 11. The cross-validation scores of no. of features ranging from 1-11 have been obtained. From the feature importance and SFS results, *RestingECG* has been dropped as its importance is low and without the avg scores are high relatively.

Implementation of Classification algorithms

- ***XGBoost Classifier:***

XGBoost is a distributed gradient boosting library that is optimised for efficiency, flexibility, and portability. It uses the Gradient Boosting framework to create machine learning algorithms. XGBoost is a parallel tree boosting algorithm that addresses a wide range of data science issues quickly and accurately.

3 types of XGBoost *Classifiers* were made:

  - *XGBoost Classifier*
  - *XGBoost Classifier with GridSearchCV*
  - *XGBoost Classifier with GridSearchCV and SFS*

- ***Gradient Boost Classifier:***

Gradient boosting is a machine learning approach for classification and Classification problems. It returns a prediction model in the form of an ensemble of weak prediction models, most often decision trees. Gradient-boosted trees is the consequence of a decision tree that is a poor learner; it generally outperforms random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as other boosting methods, but it differs in that it allows optimization of any differentiable loss function.

3 types of Gradient Boost *Classifiers* were made:

- *Gradient Boost Classifier*
- *Gradient Boost Classifier with GridSearchCv*
- *Gradient Boost Classifier with GridSearchCv and SFS*

- ***Support Vector Classifier:***

A support vector machine (SVM) is an administered AI model that involves order calculations for two-bunch arrangement issues. Subsequent to giving an SVM model arrangement of marked preparing information for every classification, they're ready to classify new text.

Support vectors are data points that are closer to the hyperplane and have an influence on the hyperplane's position and orientation. We maximize the classifier's margin by using these support vectors. The hyperplane's position will vary if the support vectors are deleted.

3 types of support vector classifiers were made:

- *Support Vector Classifier*
- *Support Vector Classifier with GridSearchCv*
- *Support Vector Classifier with GridSearchCv and SFS*

- ***Logistic Regression:***

A supervised Regression approach is a logistic Regression. It's a probabilistic analytic method that predicts outcomes. It estimates probabilities using an underlying logistic function to evaluate the connection between the dependent variable and one or more independent variables (risk factors) (sigmoid function). Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using SFS.

3 types of Logistic Regression were made:

- *Logistic Regression*
- *Logistic Regression with GridSearchCv*
- *Logistic Regression with GridSearchCv with SFS*

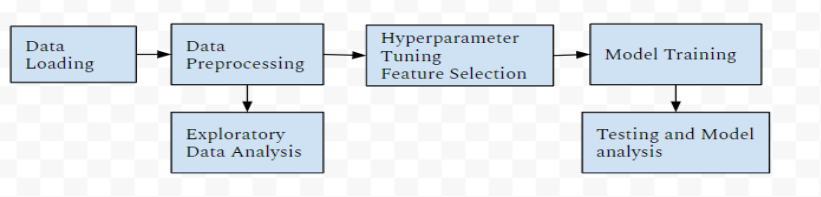
- ***Random Forest Classifier:***

Random forests, also known as random decision forests, are an ensemble learning approach for classification, regression, and other problems that work by training a large number of decision trees. The mean or average forecast of the individual trees is returned for Classification tasks.

3 types of Random Forest Classifiers are made:

- *Random forest*
- *Random forest with GridSearchCv*
- *Random Forest with GridSearchCv and SFS*

PROJECT PIPELINE



EVALUATION OF MODELS

The models implemented were evaluated using techniques like - **classification report, confusion matrix, recall, precision, accuracy score, roc AUC score, f1 score, and roc curve.** *Classification reports are taken for each and every model.*

Fig1.3 The comparison plots of the models through the pipeline

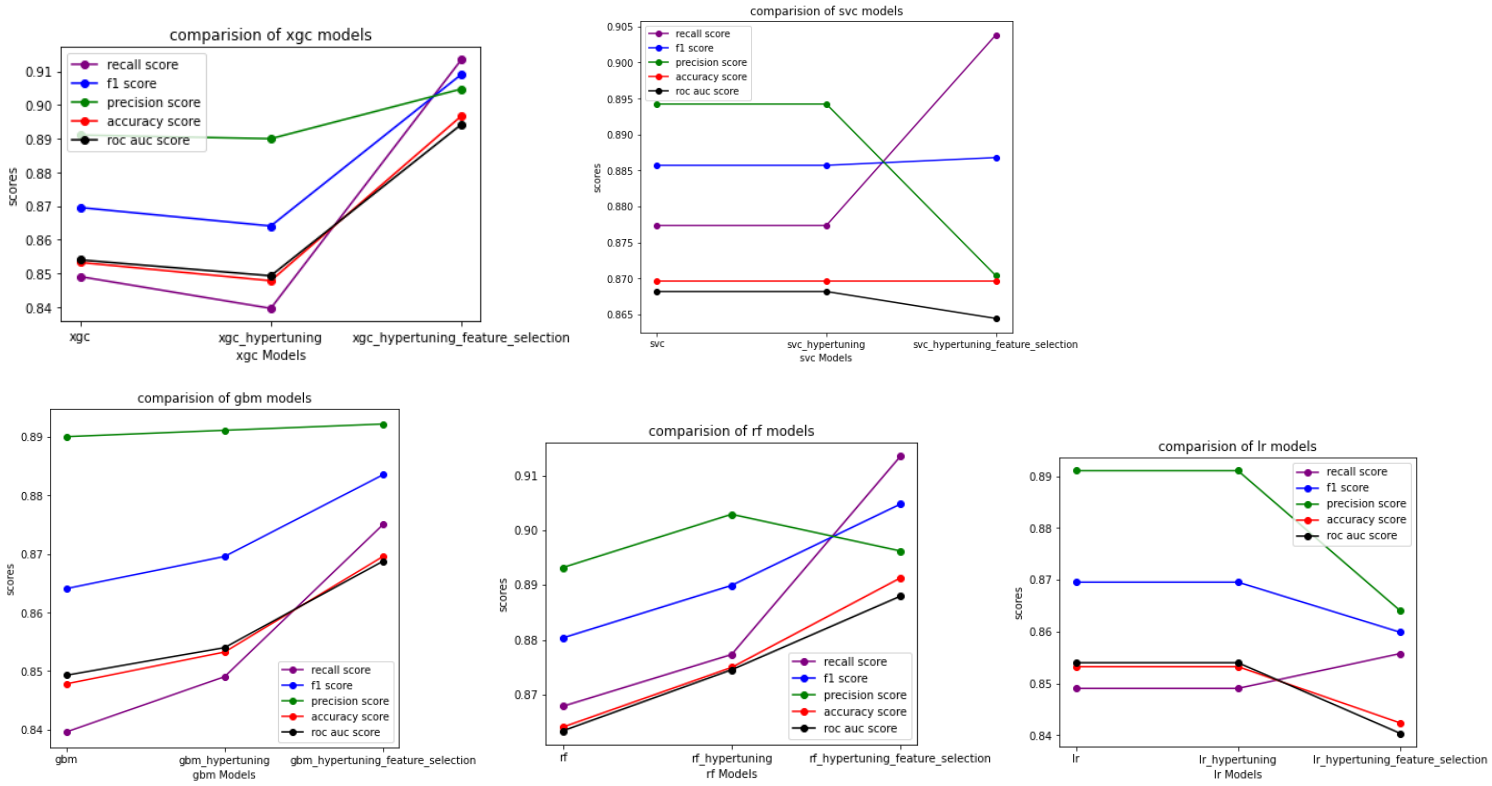


Table 1.1:The accuracy scores of the models through the pipeline

Classifiers	Normal	After Hyperparameter tuning	After Hyperparameter tuning and Feature Selection
XGBoost	0.85326086956521741	0.8478260869565217	0.8967391304347826
Support Vector Classification	0.8695652173913043	0.8695652173913043	0.8695652173913043
Logistic Regression	0.8532608695652174	0.8532608695652174	0.842391304347826
Gradient Boost	0.8478260869565217	0.8532608695652174	0.8695652173913043
Random Forest Classifier	0.8641304347826086	0.875	0.8913043478260869

Fig 1.4: The ROC curves for all models after parameter tuning and feature selection

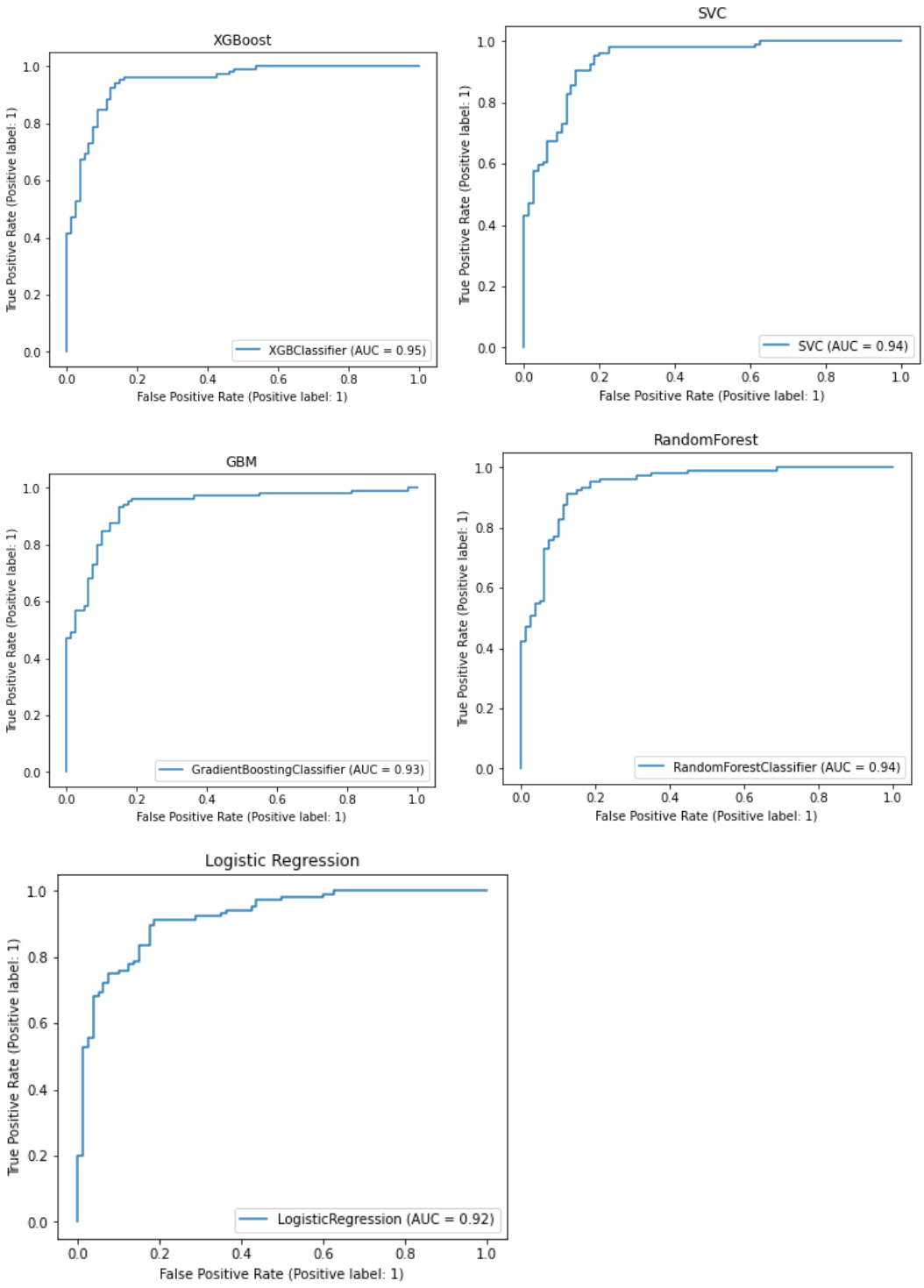
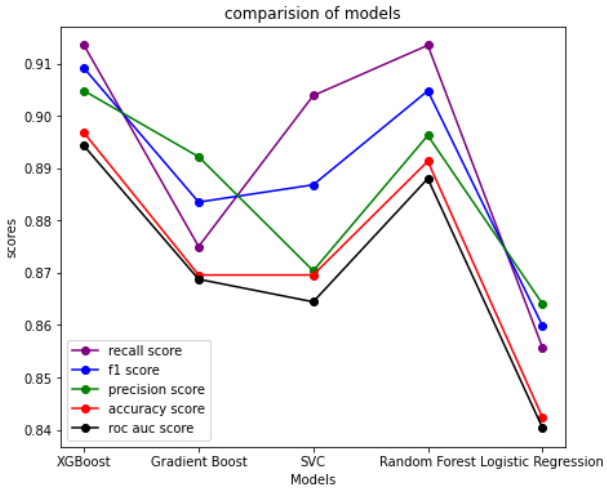


Fig 1.5:Comparison plot of all models after hyperparameter tuning and feature selection.



CONCLUSION

XGBoost has the best accuracy compared to all other models. While XGBoost, Random Forest and Gradient Boost steadily performed well after hyperparameter tuning and feature selection, not much change has been seen in the other two. XG Boost and Logistic regression have a lesser run time compared to Random Forest, SVC and Gradient Boost. Logistic Regression performs the least well among the five while XGBoost performs the best with Random Forest as runner-up. Feature selection enhances the performance of almost all the models. SVC performed similarly all throughout the pipeline.

XGBoost is a good option for unbalanced datasets. XGBoost needs only a very low number of initial hyperparameters (depth of the tree, number of trees) when compared with the Random forest. XG Boost might have chances of overfitting as well and can rely only on when testing data is provided. Therefore for this model ensemble learning techniques perform better than other Classification models out of which XGBoost has the best performance.

CONTRIBUTIONS

- Bikkavolu Prasanthi (B20CS010)- Exploratory data Analysis, Feature Selection, Model Training, Report
- Tammireddi Sai Unnathi (B20AI045)- Data Preprocessing, Hyperparameter Tuning, Model Analysis, Report

REFERENCES

1.A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018

2. https://en.wikipedia.org/wiki/XGBoost

3. https://xgboost.readthedocs.io/en/stable/get_started.html

4.https://en.wikipedia.org/wiki/Gradient_boosting

5. https://en.wikipedia.org/wiki/Random_forest

6. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

7. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html