Cardiovascular Risk Analysis using Machine Learning

A PROJECT REPORT

submitted by

P. Jagruth Reddy, AM.EN.U4ECE19141
K.V. Yokesh Kumar, AM.EN.U4ECE19123
P. Mohana Vamsi, AM.EN.U4ECE19142
Koushik Reddy, AM.EN.U4ECE19129
C. Revanth Kumar, AM.EN.U4ECE19159

under the guidance of

Dr. K.L. Nisha

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI (INDIA)
November - 2022

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

AMRITAPURI (INDIA)

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **”Cardiovascular Risk Analysis using Machine Learning”** submitted by **by P. Jagruth Reddy AM.EN.U4ECE19141,K.V. Yokesh Kumar AM.EN.U4ECE19123, P. Mohana Vamsi AM.EN.U4ECE19142, Koushik Reddy AM.EN.U4ECE19129, C. Revanth Kumar, AM.EN.U4ECE19159** in partial fulfillment of the requirements for the award of the Degree Bachelor of Technology in Electronics and Communication Engineering is a bonafide record of the work carried out by them under my guidance and supervision at Amrita School of Engineering, Amritapuri.

Signature of Supervisor:
Dr. K.L Nisha
Assistant Professor
Department of ECE

Signature of Examiner with Name

Date:13/11/2022

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

AMRITAPURI - 690 542

DEPARTMENT OF ECE

DECLARATION

We,

P. Jagruth Reddy AM.EN.U4ECE19141,
K.V. Yokesh Kumar AM.EN.U4ECE19123,
P. Mohana Vamsi AM.EN.U4ECE19142,
Koushik Reddy AM.EN.U4ECE19129,
C. Revanth Kumar AM.EN.U4ECE19159,

hereby declare that this project report entitled **"Cardiovascular Risk Analysis using Machine Learning",** is the record of the original work done by me under the guidance of **Dr. K.L Nisha,** Assistant Professor , Department of ECE, Amrita School of Engineering, Amritapuri. To the best of our knowledge this work has not formed the basis for the award of any degree/diploma/ associateship/fellowship/or a similar award to any candidate in any University.

Place: Kollam, Kerala

Signature of the Students

P. Jagruth Reddy

K.V. Yokesh Kumar

P. Mohana Vamsi

Koushik Reddy

C. Revanth Kumar

Date: 13/11/2022

Contents

Acknowledgement

We would like to express our special gratitude to Amrita Vishwa Vidyapeetham,Amritapuri and our mentor **Dr. K.L. Nisha** ma'am who gave us the opportunity to work on this wonderful project on Predictive Analysis From Patient Health Records Using Machine Learning. We learnt so many new concepts for which we are really thankful to ma'am.

Abstract

The field of healthcare is moving forward rapidly and one of the important development in that is predictive analysis. Patient health records can be used to predict and diagnose patients. This is very significant and has many use cases. It reduces the workload on healthcare workers as it speeds up the diagnosis. It can even tell us how likely a patient might develop a disease. Cardiology is one such field where predictive analysis is used. The aim of this project is to use the health records of patients and predict whether they are at the risk of having Cardio Vascular Diseases and tuning the base algorithm with feature selection and ensemble techniques for better accuracy. We were able to acquire an accuracy of 88.89 percent.

Chapter 1

Introduction

1.1 Introduction

Cardiovascular Disease or CVD is a class of diseases that involves the heart or blood vessels. They can be identified and predicted with the help of certain attributes of a patient such blood pressure, cholesterol, restECG results, etc. A large majority of CVD cases can be prevented hence it is important to diagnose the at risk patients. This project is an attempt at helping the healthcare workers predict and detect CVDs with the help of minimal information and time. The dataset being used to train and test the model is the Heart Disease Cleveland UCI Dataset. It has 13 attributes ranging from chest pain to number of major blood vessels from the health records of 298 patients.

1.2 Literature review

1.2.1 A Comprehensive survey on Heart Disease Prediction using Machine Intelligence : A Review

Aim: According to the World Health Organization's most recent figures, one-third of the world's population will die from cardiovascular illnesses, including coronary heart disease, heart attacks, and vascular disease. Applying the best machine learning model to target early detection and precise prediction of heart illness is essential given the emergence of AI trends in order to reduce mortality rates and provide the best clinical decision support for treating cardiac patients. This is the driving force for this essay. The Cleveland data-set and the Z-Alizadeh Sani dataset are two well-known heart disease data-sets that were used to construct and evaluate the prediction models presented in this study.

Methods: Between 2005 and 2020, papers from the Google Scholar, Scopus, Web of Science, Research Gate, and PubMed search engines were used to conduct this study. Heart illness, prediction, coronary disease, healthcare, heart datasets, and machine learning were the primary search terms.

Results: This study examines the drawbacks of several methods for predicting cardiac disorders. It lists the benefits and drawbacks of various research approaches as well as the standards by which each evaluated article was validated.

1.2.2 Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators: A Review

Ten machine learning (ML) classifiers from various categories, including Bayes, functions, lazy, meta, rules, and trees, were trained in this study to predict the risk of heart disease accurately using both the Cleveland heart data-set's full set of attributes and the best attribute sets determined by three attribute evaluators. The 10-fold cross-validation testing technique was used to assess the performance of the employed algorithms.The procedure concludes with the tuning of instance-based (IBk) classifier's hyper-parameter k, which stands for the number of nearest neighbors. Using the whole set of attributes, the sequential minimum optimization (SMO) produced an accuracy of 85.148%, while the optimum attribute set derived using the chi-squared attribute evaluator yielded the greatest accuracy result of 86.468%

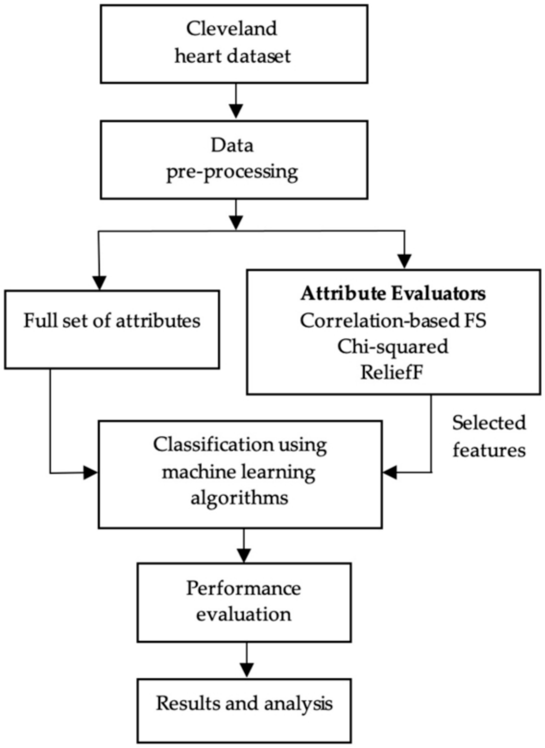


Figure 1.1: Process Flow

1.2.3 Enhanced Heart Disease Prediction Based on Machine Learning and chi-squared Statistical Optimal Feature Selection Model : A Review

In order to enhance heart disease diagnosis, this research suggests a novel heart disease classification model based on the support vector machine (SVM) method. The chi-squared statistical optimal feature selection method was employed to improve prediction accuracy. The effectiveness of the recommended model was then verified by contrasting it with conventional models using a variety of performance metrics. The suggested model's accuracy rose from 85.29% to 89.7%. The componential burden was also cut in half. This outcome shows that the proposed algorithm fared better at predicting cardiac disease than other cutting-edge techniques.

A correlation-based feature selection technique known as the chi-squared test establishes the relationship between the characteristics and the anticipated class. To ascertain which characteristics are dependent on the anticipated attribute, each non-negative feature (Xi) computes chi-square statistics. The characteristic is more dependent on the projected class when the chi-square score is greater.

1.2.4 Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques : A Review

Among the most effective machine learning methods for prediction is classification. While some categorization algorithms offer adequate prediction accuracy, others only show marginal accuracy. This study looks at a technique called ensemble classification, which combines many classifiers to increase the precision of weak classifiers. The Cleveland UCI Heart Disease dataset related to heart disease was used in this study. To ascertain how the ensemble technique may be used to increase prediction accuracy in heart disease, a comparative analytical approach was used. In order to demonstrate the algorithm's usefulness in early illness prediction, this research focuses not only on improving the accuracy of weak classification algorithms but also on how to apply the method using a medical dataset. The study's findings show that ensemble approaches, such bagging and boosting, are useful for increasing the predictive accuracy of weak classifiers and perform admirably in predicting the risk of developing heart disease. Ensemble classification helped weak classifiers achieve an accuracy improvement of up to 7%. Implementing feature selection improved the method' performance even further, and the results revealed a notable rise in prediction accuracy.

1.2.5 An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour.

Many studies have been conducted in this field utilising machine learning to forecast various illnesses. The accuracy of single classifiers is not necessarily greater. This study is essentially based on ensemble learning mechanism, another ground-breaking advancement in machine learning, which combines numerous simple classifiers for a certain model to provide very satisfying results. The suggested model uses a bagging ensemble technique with Logistic Regression, Gaussian Naive Bayes, and K Nearest Neighbor as the basic classifier to predict the occurrence of heart disease. A cardiovascular disease prediction system employing the bagging approach is the suggested model. The accuracy rates attained using three distinct classifiers—Logistic Regression, Gaussian Naive Bayes, and K closest neighbor—for bagging are 82.8%, 82.5%, and 83.2%, respectively. The data set used is same as the one used in above studies.

Chapter 2

Implementation

2.1 Introduction

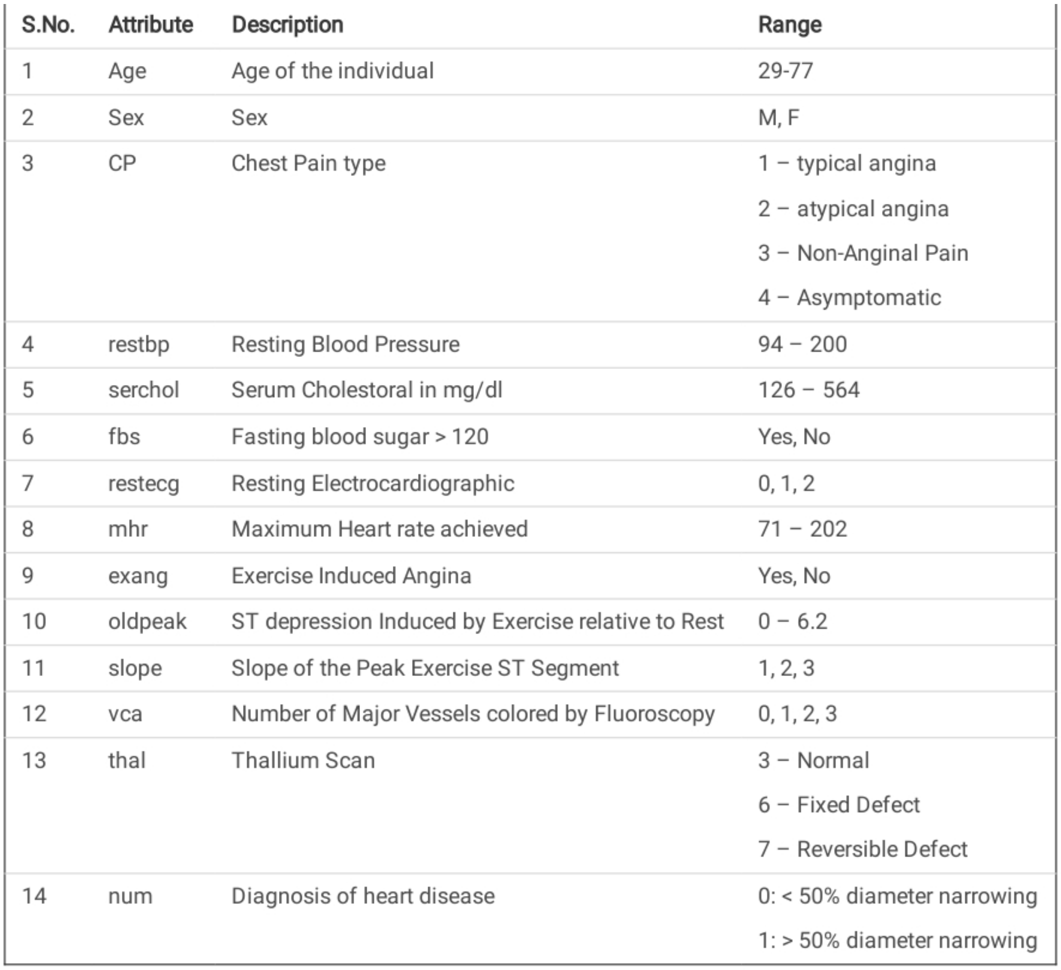
2.1.1 Introduction

The process of implementation as usual started with the pre-processing of the dataset, visualizing and assessing the data and looking for discrepancies. After selecting a few models we trained them with the pre-processed data, followed by adding feature selection, bagging and boosting one-by-one, gradually improving the accuracy of the models while being as efficient as possible. The ML models we used are - Support Vector Machine, Random Forest, Decision Trees, Logistic Regression, and Naive Bayes.

2.2 Data-set Descreption

There are 298 observations in the dataset, along with 13 features and 1 target attribute. The outcomes of some non-invasive diagnostic tests like exercise electrocardiogram, thallium scintigraphy and fluoroscopy of coronary calcification are among the 13 characteristics, together with other pertinent patient data. The target variable comprises the outcome of the invasive coronary angiography, which indicates whether the patient has coronary artery disease or not. Labels 0 and 1-4 denote the existence of CHD, respectively. The majority of studies utilising this dataset have focused solely on attempting to differentiate between presence (values 1, 2, 3, 4) and absence (value 0).

Robert Detrano, M.D., Ph.D. of the Cleveland Clinic Foundation gathered the information in the dataset.



S.No.	Attribute	Description	Range
1	Age	Age of the individual	29-77
2	Sex	Sex	M, F
3	CP	Chest Pain type	1 – typical angina 2 – atypical angina 3 – Non-Anginal Pain 4 – Asymptomatic
4	restbp	Resting Blood Pressure	94 – 200
5	serchol	Serum Cholestoral in mg/dl	126 – 564
6	fbs	Fasting blood sugar > 120	Yes, No
7	restecg	Resting Electrocardiographic	0, 1, 2
8	mhr	Maximum Heart rate achieved	71 – 202
9	exang	Exercise Induced Angina	Yes, No
10	oldpeak	ST depression Induced by Exercise relative to Rest	0 – 6.2
11	slope	Slope of the Peak Exercise ST Segment	1, 2, 3
12	vca	Number of Major Vessels colored by Fluoroscopy	0, 1, 2, 3
13	thal	Thallium Scan	3 – Normal 6 – Fixed Defect 7 – Reversible Defect
14	num	Diagnosis of heart disease	0: < 50% diameter narrowing 1: > 50% diameter narrowing

Figure 2.1: Features in the data-set[2]

2.3 Block Diagram and Method

The process flow is depicted in Figure 2.2:

2.3.1 Classifiers Used

Naive Bayes

The Bayes Theorem is the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilised for a variety of classification problems.It doesn't

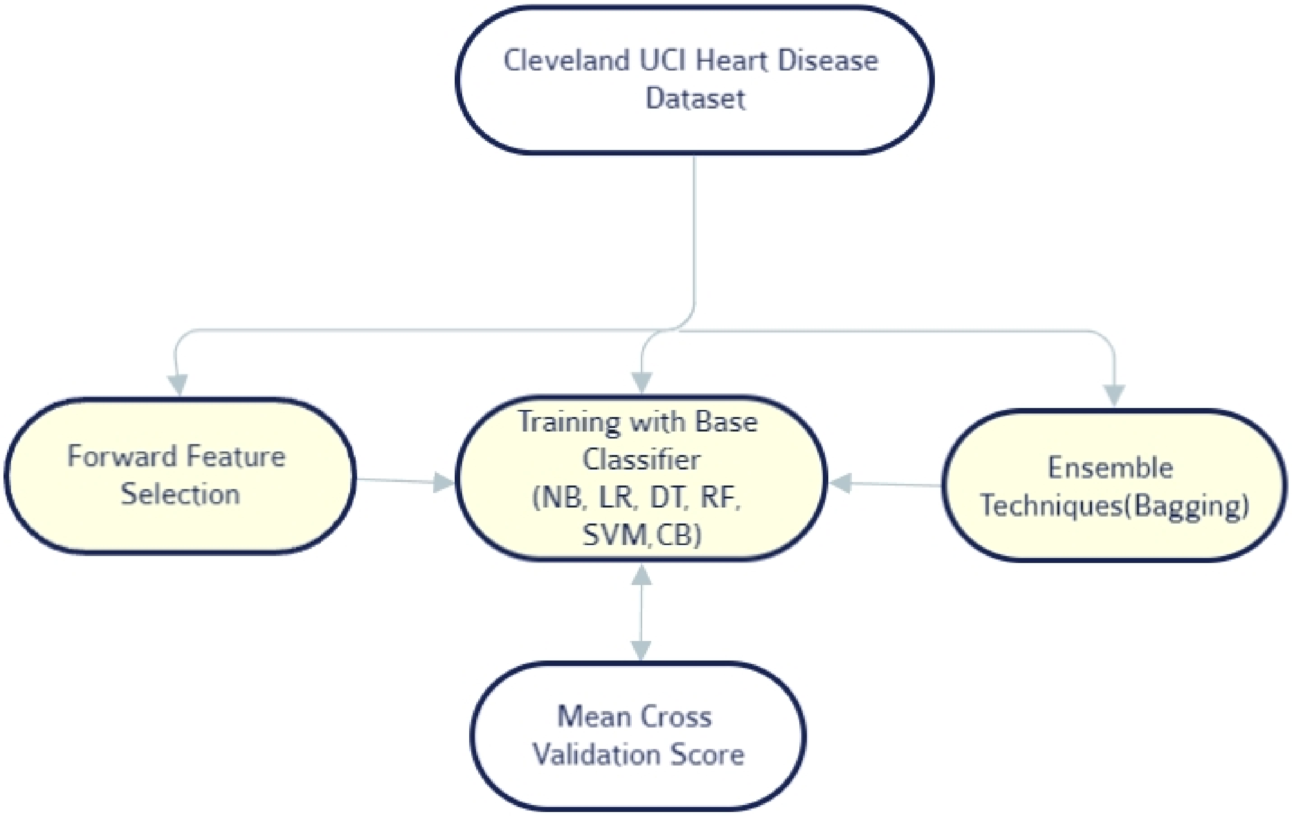

Cleveland UCI Heart Disease Dataset
Forward Feature Selection
Training with Base Classifier (NB, LR, DT, RF, SVM,CB)
Ensemble Techniques(Bagging)
Mean Cross Validation Score

Figure 2.2: Process flow Block Diagram

require as much training data. It handles both continuous and discrete data.

Logistic Regression

It is used to calculate or predict the probability of a binary (yes/no) event occurring. In this context, it would mean classifying or finding the probability of a person having the cardiovascular disease or not.

Decision Tree

It is a non-parametric supervised learning approach.It is organised hierarchically structure consisting of a root node, branches, internal nodes, and leaf nodes.Each leaf node (terminal node) stores a class label, and each internal node indicates a test on an attribute. Each branch reflects an outcome of the test.

Random Forest

A classification system made up of several decision trees is called the random forest. It attempts to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree by using bagging and feature randomization while generating each individual tree.

Support Vector Machine

The sorted data are produced as a map by an SVM, with the margins between the two being as far away as feasible. The objective is to select a hyperplane that has the largest feasible margin between it and any point in the training set, increasing the likelihood that fresh data will be properly categorised.

CatBoost

Sometimes known as Category Boosting, is similar to XGBoost in that it is based on decision trees and gradient boosting, but it performs much better! CatBoost performs particularly effectively on data with categorical variables.With relatively little data, it can get an excellent outcome.

2.3.2 Forward Feature Selection

An iterative process called forward selection starts with the model having no features. The feature that best enhances our model is added in each iteration until the performance of the model is not improved by the addition of a new variable.

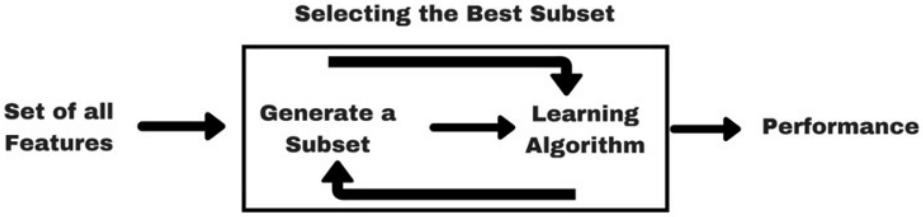


Figure 2.3: Forward selection

2.3.3 Bagging

Another name for bagging is bootstrap aggregation. With replacement, bagging randomly chooses a few patterns from the training set. With few omissions ,repetitions and replacements new training set is formed. Bagging involves retrieving bootstrap samples from the data and training the classifier with each sample. Each classifier's votes are added together, and the classification outcome is decided by majority vote or average. A poor classifier might perform better when bagging is applied, according to research. Bagging lowers the prediction's variance.

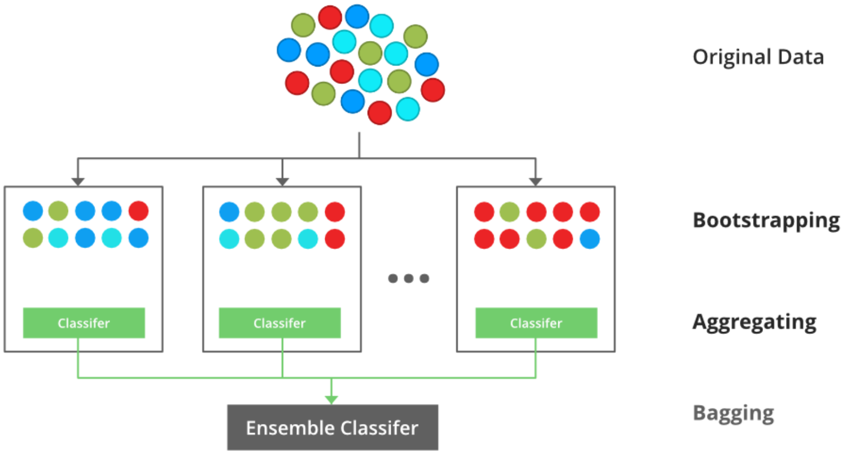


Figure 2.4: Bootstarp Aggregation[7]

2.4 Training the Classifiers

We train the base classifier first and obtain the mean cross-validation scores(MCVS) with cv=10. After this process is completed we move to feature selection and bagging. After the feature selection and bagging setup is done, MCVSs are obtained for each complex model.

2.5 Results

The mean cross-validation scores of the accuracy metric are shown in table 2.1 Bagging improved the scores in classifiers like Decision trees, SVM, and marginally in Naive Bayes classifier. With forward feature selection there is a considerable improvement across all the classifiers except Logistic Regression which didn't show any improvement with either bagging or feature selection.

Features selected for SVM : ['sex', 'cp', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']

Features selected for RF: ['chol', 'fbs', 'restecg', 'oldpeak', 'slope', 'ca', 'thal']

Model	Base	with Bagging	with Feature Selection
SVM	81.48	83.16	83.85
RF	80.78	80.78	82.17
DT	73.72	77.09	77.09
LR	83.86	83.86	82.86
NB	83.11	83.2	85.17
CatBoost	88.89	N/A	N/A

Table 2.1: Mean Cross-validation Scores.

Features selected for DT: ['sex', 'chol', 'restecg', 'exang', 'slope', 'ca', 'thal']

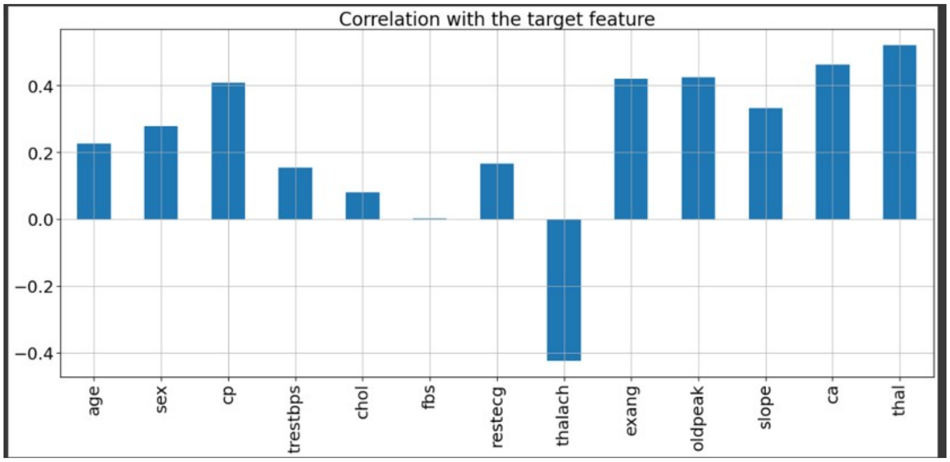


Figure 2.5: Correlation of features with the target

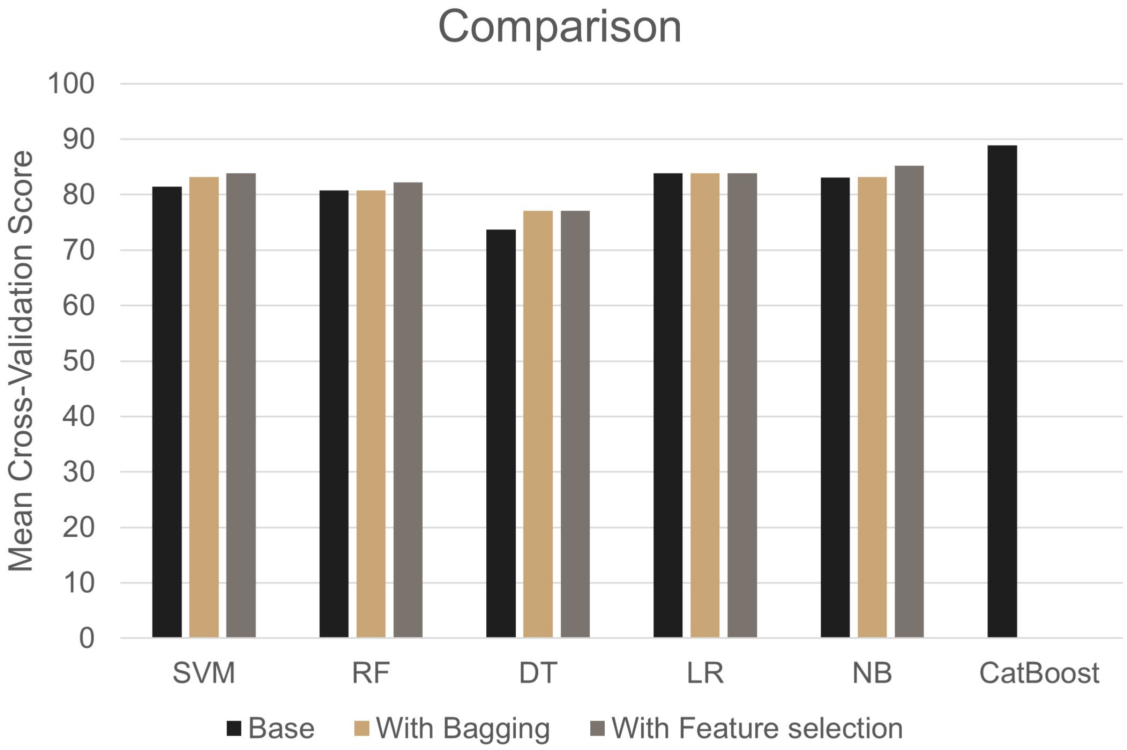


Figure 2.6: Results

Chapter 3

Conclusion and Scope for further research

We have been able to successfully improve the models' accuracy with the help of feature selection and bagging. CatBoost gave a considerably higher accuracy. With feature selection and ensemble techniques we have been able to get the main features that correlated well with disease and improve the accuracy. The aim is to apply this experience and knowledge in our main project which deals with risk analysis of Retinopathy of Prematurity. The dataset is based on the patient health records from a Hospital at Bengaluru.

References

[1] C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, Volume 16, 2019, 100203, ISSN 2352-9148, [https://doi.org/10.1016/j.imu.2019.100203.]

[2] Santhosh Gupta Dogiparthi, Dr. Jayanthi K, Dr. Ajith Ananthakrishna Pillai et al. A Comprehensive survey on Heart Disease Prediction using Machine Intelligence, 06 July 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-680505/v1]

[3] Annwesha Banerjee Majumder et al 2022 J. Phys.: Conf. Ser. 2286 012017[https://iopscience.iop.org/article/10.1088/1742-6596/2286/1/012017]

[4] Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Appl. Sci. 2021, 11, 8352. [https://doi.org/10.3390/app11188352]

[5] Sarra, R.R.; Dinar, A.M.; Mohammed, M.A.; Abdulkareem, K.H. Enhanced Heart Disease Prediction Based on Machine Learning and 2 Statistical Optimal Feature Selection Model. Designs 2022, 6, 87. [https://doi.org/10.3390/designs6050087]

[6] Dataset - https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] Image Illustrations -https://www.geeksforgeeks.org/