machine learning

**Project Report on Heart Disease Prediction**

Table of Contents

[Data Description 2](#_Toc152608476)

[**Introduction** 2](#_Toc152608477)

[**Heart Disease Dataset** 2](#_Toc152608478)

[**Why Heart Disease Dataset?** 5](#_Toc152608479)

[Data Analysis 6](#_Toc152608480)

[**Correlation Analysis:** 6](#_Toc152608481)

[**Distribution of Target Variable:** 7](#_Toc152608482)

[Data Exploration 8](#_Toc152608483)

[Experimental method 10](#_Toc152608484)

[**Execution Process:** 10](#_Toc152608485)

[results and Analysis 12](#_Toc152608486)

[Bibliography 15](#_Toc152608487)

# **Data Description**

## **Introduction**

Heart disease is a major global health concern with its prevalence rising quickly. The World Health Organization (WHO) states that a significant percentage of deaths globally are caused by cardiovascular diseases. For early intervention and prevention, it is essential to comprehend and anticipate the onset of cardiac disease. The goal of this research is to create a strong predictive system for the early identification of heart disease by utilizing machine learning techniques.

Predicting heart disease is difficult since it affects several different organs and has sophisticated interdependencies. Although there are established diagnostic techniques, it is still difficult to reliably forecast heart disease in its early stages. By combining insights from several machine learning methods, the initiative aims to maximize the potential of data science and machine learning to improve early diagnosis.

The primary objective of the project is to predict the probability of heart disease using three different machine learning techniques: Logistic regression, Ensemble Methods like AdaBoost, Gradient Boost, Random Forest, Extra Forest. The goal of the study is to create accurate predictive models by utilizing the extensive features of the Heart Disease Dataset.

Additionally, this initiative uses machine learning algorithms to provide a successful early detection strategy for cardiac disease.

## **Heart Disease Dataset**

Dataset Link: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data>

The Heart Disease Dataset is a dataset that is used to predict the risk of heart disease based on several types of health-related factors. The structure of this dataset allows for analysis and prediction of an individual's likelihood of having heart disease or not.

There are 1025 records in the dataset, and each record has 14 attributes. It focuses on health metrics particularly, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, oldpeak, the slope of the peak exercise ST segment, number of major vessels colored by flourosopy, thal, and a target variable called "target," which indicates whether heart disease is present (1) or absent (0). The dataset was acquired from “Kaggle” which holds a good reputation in providing real-world datasets.

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Serial Number** | **Name of the Attribute** | **Description of the Attribute** |
| 1 | age | Age of patient |
| 2 | sex | Gender of Patient |
| 3 | cp | Chest Pain Type (4 values) |
| 4 | trestbps | Resting Blood Pressure |
| 5 | chol | Serum Cholestrol in mg/dl |
| 6 | fbs | Fasting Blood Sugar > 120 mg/dl |
| 7 | restecg | Resting Electrocardiographic Results |
| 8 | thalach | Maximum heart beat |
| 9 | exang | Exercise Induced Angina |
| 10 | oldpeak | Oldpeak |
| 11 | slope | The slope of the peak exercise |
| 12 | ca | Number of major vessels (0-3) colored by flourosopy |
| 13 | thal | Thal: 0 = normal; 1: fixed defect, 2 = reversable defect |
| 14 | target | Outcome: 1= Presence of Disease, 0= Absence of disease |

Data Type: It indicates a combination of scalar and categorical data representing patients' health information. It consists of 8 scalar predictive variables and 1 categorical output (target).

Volume: The dataset meets the criteria for being moderately big, meaning that it may be used for machine learning modelling without requiring an excessive amount of processing with a file size of 39 KB and 1025 instances.

Variables: The attributes in the dataset are primarily numerical and represent parameters related to health. The "target" variable, on the other hand, highlights the existence or absence of heart disease and functions as a categorical indicator.

File Format: It is available in the popular CSV (Comma-Separated Values) format, which makes it easy to work with and integrate data into Python environments particularly when using tools like Pandas.

Usability: The dataset requires minimum preprocessing for machine learning analysis because it is extremely well-structured. Its CSV structure and cleanliness make it easier to use in machine learning processes.

Veracity: This dataset is considered dependable in the field of machine learning because it does not contain any errors or missing values. Like real-world health data, though, it may include some outliers that call for cautious handling.

A screen shot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

Time Relevance: Because the dataset does not have time stamps, it is historical health-related data that may always be used for analysis, regardless of how old it may be.

Value: Priceless for building prediction models to anticipate the incidence of heart disease by utilizing a wide variety of health indicators to promote early detection and risk assessment.

## **Why Heart Disease Dataset?**

Accessibility: This dataset is publicly available which makes it easily accessible to researchers and developers, encouraging their use in a variety of projects.

Interpretability: Combining numerical and categorical variables makes machine learning algorithm evaluation easier and promotes thorough analysis.

Real World Application: May find use in healthcare settings, allowing for the early detection and diagnosis of cardiac disease.

Relevance: As heart diseases have such a huge impact, early detection models are very important which makes the creation and training of strong predictive models.

Use: The discipline of machine learning makes extensive use of it that highlights its importance in predicting heart disease and explaining complex correlations between health metrics.

# **Data Analysis**

Mean: Provides a baseline indication of typical measurements by representing the central tendency or average value of the related variable across the collection.

Minimum and Maximum Values: Indicate the range of values found in the dataset, emphasizing the lowest and highest values found for every variable. This helps to clarify the dataset's variability and diversity.

A screenshot of a black and white screen

Description automatically generated

## **Correlation Analysis:**

Age and Cholestrol: About 0.21 is the correlation coefficient (moderate positive correlation). Age and cholesterol levels appear to have a moderately positive link, according to the correlation coefficient. Though the relationship is not very strong, cholesterol levels tend to grow with age. This realization suggests that older people may typically have somewhat elevated cholesterol levels.

Age and fasting blood sugar: Coefficient of 0.15 or such (not very strong positive correlation).

The small positive correlation points to a minor age-related increase in fasting blood sugar. The weak correlation suggests that age may not be the primary factor influencing the fasting blood sugar of the participants in the dataset.

A screenshot of a computer

Description automatically generated

## **Distribution of Target Variable:**

Understanding the balance and class distribution of the heart disease dataset depends critically on the distribution of the target variable, which often indicates the presence or absence of heart disease.

In this case, the target variable typically indicates whether or not a person has heart disease. It is frequently binary (0 for no heart disease, 1 for heart disease) or categorical (identifying various heart disease stages or types). In our case, it is binary.

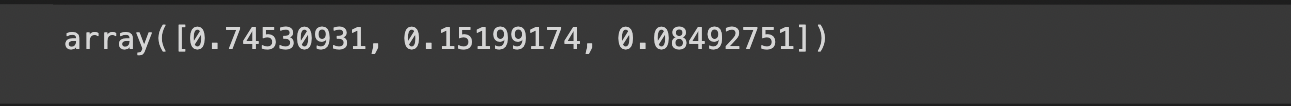
A graph with red and blue bars

Description automatically generated

# **Data Exploration**

**Principal Component Analysis (PCA):** PCA reduces the dimensionality of the dataset while preserving the majority of its contents. Their explained variance establishes the proportional importance of the primary components.

The cumulative explained variance from the major components is computed to obtain 96% of the variance. The amount of dimensionality reduction while maintaining substantial information is shown by the number of principal components required to attain or surpass 96% cumulative variance. In our case, with PCA, there are 3 components necessary to explain the 96% of the variance.



PCA provides information on the variables that most influence the variance of the dataset. It finds feature combinations that account for the variability in the data and may reveal important aspects affecting the prediction of heart disease.

While PCA gives a transformation that maximizes variance, it does not directly reveal the nature of variables. Interpreting their importance is made easier by knowing the connections between the original variables and major components.

**Decision Trees:** Using decision trees, particularly the less deep ones (2–5), provides a clear understanding of how the algorithm divides the data according to feature thresholds.

Easy-to-understand decision trees make it possible to identify the most important characteristics and their threshold points for heart disease prediction.

The most discriminative variables at each node are displayed in the decision tree graphic, which illustrates how the algorithm divides the dataset according to feature values.

Understanding how certain factors work together to predict heart disease is made easier with the use of depth-limited decision trees, which provide insights into feature connections and relationships.

A diagram of a number of data

Description automatically generated with medium confidence

# **Experimental method**

Hypothesis: Predict Heart Disease based on all the attributes contributing towards the heart disease using ensemble methods such as Random Forest, Extra Trees, AdaBoost, Gradient Boost and MLP (neural network) for regression analysis.

## **Execution Process:**

* Partitioning the data into separate training and testing subsets while preserving an 80/20 ratio was a crucial first step in utilizing the heart disease dataset for machine learning. This preserved an independent collection for objective model evaluation while guaranteeing a strong approach to model training on a significant portion of the dataset. By using this split, the models were trained to identify patterns in the data while permitting evaluation on cases that were not seen, which is essential for confirming their relevance in the actual world.
* After that, StandardScaler was used to scale the data. By standardizing the range of features, this crucial normalization process ensured more equitable model training by reducing the dominance of some variables over others. To eliminate biases toward characteristics resulting from different units or scales, scaling the features uniformly helped to improve the effectiveness of the model learning process.
* The models were started with default parameters once the data had been preprocessed, which provided a baseline from which to measure their initial performance. In order to better understand the behaviour of the models before adjusting the hyperparameters, this step was helpful.
* The next step was to use GridSearchCV to carefully adjust the model's hyperparameters. For every regression technique, specific hyperparameter grids were created, allowing for a thorough search for the best possible parameter combinations. Through methodical study, the ideal parameter configurations had to be found to improve model behaviour and forecast accuracy.
* Five-fold cross-validation was used during the hyperparameter tuning process, which is an essential procedure to guarantee robustness and generalizability. By validating the models over many subsets of the training data, this technique reduced the possibility of overfitting and produced a trustworthy evaluation of the models' performance.

.

# **results and Analysis**

The results of training different regression models on the heart disease dataset demonstrated different machine learning method performances. In terms of predicting accuracy, the Multilayer Perceptron (MLP) neural network and the Extra Trees method stood out among these models. With an astounding R2 Score of 0.9265 and a noticeably decreased Mean Squared Error (MSE) of 0.0184, the MLP model proved to be incredibly powerful. This implies that it has extraordinary predictive power based on the properties of the dataset.

Furthermore, the Extra Trees algorithm demonstrated strong performance, as evidenced by its MSE of 0.0240 and R2 Score of 0.9040, confirming its effectiveness in capturing intricate relationships within the dataset. These models proved to be reliable in predicting heart disease based on the given data, as they not only produced better predicted accuracy but also showed consistency across several evaluation criteria.

In contrast to the best-performing MLP and Extra Trees, the Random Forest and Gradient Boost models did quite well, despite having somewhat higher MSE values. On the other hand, the AdaBoost model performed poorly in terms of predicting accuracy. It showed a lower R2 Score of 0.4897 and a comparatively higher MSE of 0.1274.

These results indicate possible real-world uses for the Extra Trees and MLP models in medical settings. Their strong predictive abilities may prove to be useful tools for medical professionals, helping to forecast heart disease based on patient factors. Using these models to measure risk may help with early intervention or customized treatment plans for those who are more vulnerable, which could enhance patient outcomes and healthcare decision-making. On considering all the things, the findings highlight the potential of these machine learning models as prognostic instruments for cardiovascular health care.

A screenshot of a computer error

Description automatically generated

On decreasing the values of the hyperparameters to very low, I observed:

**Random Forest**: MSE: 0.0497, R2 Score: 0.8009

**Extra Trees**: MSE: 0.0261, R2 Score: 0.8954

**AdaBoost**: MSE: 0.1126, R2 Score: 0.5490

**Gradient Boost**: MSE: 0.0789, R2 Score: 0.6840

**MLP**: MSE: 0.0962, R2 Score: 0.6149

* Most models showed a reduction in predictive performance as the number of hidden layers and depth was decreased, hence reducing the complexity. Although the Random Forest and MLP models both displayed somewhat higher mean square error (MSE) and lower R2 Scores, which suggest a lower predictive accuracy, the effects differed among the other models.
* Interestingly, performance measures showed very little variation for Extra Trees, suggesting that it is resilient to complexity reductions. AdaBoost and Gradient Boost models showed increases in R2 Scores, but at the cost of higher MSE, indicating a bias-variance trade-off. On the other hand, the relative decline in MLP's performance indicates how sensitive it is to changes in architecture, indicating how crucial a good network architecture is to maintain predictive accuracy.

On increasing the values of the hyperparameters to very low, I observed:

**Random Forest**: MSE: 0.0488, R2 Score: 0.8045

**Extra Trees**: MSE: 0.0267, R2 Score: 0.8929

**AdaBoost**: MSE: 0.1291, R2 Score: 0.4829

**Gradient Boost**: MSE: 0.0455, R2 Score: 0.8179

**MLP**: MSE: 0.0241, R2 Score: 0.9035

* We can see from the comparison that performance measures slightly alter as complexity increases. In contrast to its initial results (MSE: 0.0490, R2 Score: 0.8037), the Random Forest model showed a somewhat higher R2 Score (0.8045) and a comparable MSE (0.0488). This small difference could mean that performance stays steady even with more complexity.
* After complexity changes, the Extra Trees model continued to perform fairly close, demonstrating consistency from its initial values (MSE: 0.0240, R2 Score: 0.9040) with an MSE of 0.0267 and an R2 Score of 0.8929.
* Nevertheless, following complexity changes, the AdaBoost model showed a minor rise in MSE (0.1291) and a decrease in R2 Score (0.4829), suggesting that added complexity may have an effect on the model's performance.
* The MSE and R2 Score for Gradient Boost and MLP also somewhat changed as complexity increased, but they continued to function pretty close to their initial values, suggesting that the complexity changes did not significantly alter their behaviour.

# **Bibliography**

"1.11. Ensembles: Gradient Boosting, Random Forests, Bagging, Voting, Stacking." *Scikit-learn*, scikit-learn.org/stable/modules/ensemble.html. Accessed 4 Dec. 2023.

"Heart Disease Dataset." *Kaggle: Your Machine Learning and Data Science Community*, [www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data](http://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data).

"Python Seaborn Tutorial." *GeeksforGeeks*, 26 July 2023, [www.geeksforgeeks.org/python-seaborn-tutorial/](http://www.geeksforgeeks.org/python-seaborn-tutorial/).

*World Health Organization (WHO)*, www.who.int/.