HEART DISEASE PREDICTION

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Prashanthi Thota- (AP21110010069)

Neeli Meghana- (AP21110010127)

Lakshmi Nikhitha Dodda- (AP21110011270)

Under the Guidance of

Mr Ratna Raju Mekala

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[May, 2024]

Certificate

Date: 2-Sep-24

This is to certify that the work present in this Project entitled "**Heart Disease Prediction**" has been carried out by **Prashanthi Thota, Neeli Meghana Nandigam, Lakshmi Nikhitha Dodda** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

Mr Ratna Raju Mekala

Assistant Professor,

Computer Science and Engineering.

Co-supervisor

Prof. Niraj Upadhayaya

Head of the Department,

Computer Science and Engineering.

Acknowledgements

We extend our sincere gratitude to everyone who contributed to the successful completion of our Heart Disease Prediction project. Throughout this journey, we have been fortunate to receive continuous support, valuable advice, and unwavering encouragement.

A special note of appreciation goes to Mr. Ratna Raju Mekala whose extensive knowledge and support played a crucial role in shaping the course of this machine learning project.

Looking back, being part of this challenging yet fulfilling project has been an honor. We are eager to apply the skills and knowledge gained to further advance our careers in machine learning.

Prashanthi Thota (AP21110010069)

Neeli Meghana Nandigam (AP21110010127)

Lakshmi Nikhitha Dodda (AP21110011270)

Table of Contents

1. Introduction

Heart disease is one of the most prevalent and serious health issues, with a high toll on individuals, families, and the entire healthcare system. Despite the remarkable progress in medical science and technology, cardiovascular diseases continue to be the No. 1 cause of morbidity and mortality worldwide. Responding to this imperative public health challenge, our project seeks to harness the power of machine learning techniques in constructing robust predictive models of heart disease. We intend to make better risk predictions, provide early detection, and propose better and personalized treatment for individuals at risk for the development of heart disease, using rich information contained in comprehensive datasets involving clinical, demographic, and lifestyle factors.

1.1 Background

Heart disease comprises a wide range of conditions affecting the cardiovascular system, including coronary artery disease, heart failure, and arrhythmias. Heart disease is a multifaceted interaction of genetic predisposition, environmental influences, and lifestyle factors. Despite decades of research and many clinical interventions, the prevalence of heart disease continues to rise, posing considerable challenges to healthcare systems around the world. Understanding the multifactorial etiology and pathophysiology of heart disease is paramount to devising effective prevention and management strategies.

1.2 Purpose

The objective of our project is to develop cutting edge predictive models that will help in the identification of people who have a higher risk of developing heart diseases. This is done with the hope and application of the strong analytical capability of machine learning algorithms in finding hidden patterns, subtle correlations, and underlying mechanisms in the development of heart diseases. We hope that our work will empower healthcare providers with actionable insights in a timely manner, enabling proactive intervention, personalized treatment, and optimization of patient outcomes.

1.3 Dataset Overview

Our dataset includes essential demographic and clinical variables crucial for predicting heart disease risk. These variables consist of age, gender, chest pain type, resting blood pressure, serum cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved during exercise, presence of exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and

thalassemia type. Each variable offers valuable insights into the patient's cardiovascular health status, facilitating the development of accurate predictive models for early detection and intervention of heart disease.

1.4 Significance

The significance of our project will be extraordinary in the fact that it could change the face of cardiovascular disease management, using advanced data analytics and predictive models. Using machine learning predictive power, we will have the potential to reveal new knowledge about how risk factors interact in intricate ways with each other, genetic predispositions, and environmental factors that result in susceptibility to heart diseases.

1.5 Objectives

1.5.1 Identify Novel Risk Factors: Conduct in-depth exploratory analysis of the dataset to identify novel risk factors and biomarkers associated with heart disease onset and progression.

1.5.2 Develop Robust Predictive Models: Employ state-of-the-art machine learning algorithms to develop robust and scalable predictive models capable of accurately stratifying individuals based on their risk of developing heart disease.

1.5.3 Enhance Risk Stratification Strategies: Enhance existing risk stratification methodologies by integrating advanced predictive modelling techniques, thereby enabling more accurate identification of high-risk individuals and facilitating targeted preventive interventions.

1.5.4 Empower Healthcare Providers: Provide healthcare providers with intuitive and user-friendly decision support tools, enabling them to make informed clinical decisions, optimize resource allocation, and deliver personalized care tailored to each patient's unique risk profile.

2. Problem Statement

The problem talks about the most accurate prediction and classification of heart disease based on these comprehensive sets of clinical and demographic features. The era in which cardiovascular diseases remain a major killer worldwide calls for precise risk assessment and early detection. The important attributes of this dataset include: age, sex, type of chest pain (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), the electrocardiographic result at rest (restecg), maximum heart rate achieved during exercise (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and thallium heart scan results (thal).

Through the analysis of these factors, the objective is to develop robust machine learning models capable of accurately predicting the presence or absence of heart disease in individuals. This predictive capability is crucial for healthcare professionals in optimizing patient care, devising personalized treatment plans, and implementing preventive measures. By leveraging advanced computational techniques, the goal is to enhance diagnostic accuracy, facilitate timely intervention, and ultimately improve patient outcomes. Additionally, the development of such models holds the promise of streamlining healthcare processes, reducing healthcare costs, and alleviating the burden on healthcare systems worldwide.

3. Proposed Approach

3.1 Data Preparation:

3.1.1 Data Loading:

Data loading is the initial step in our project, where we import the heart disease dataset, which contains various clinical and demographic features. This dataset serves as the foundation for our predictive modeling task.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1

3.1.2 Handling Missing Values:

Upon loading the dataset, it's crucial to check for missing values that might exist within the data. Missing values can adversely affect the performance of machine learning models if not appropriately handled. We employ strategies such as imputation or removal to address missing values, ensuring the integrity of the dataset.

3.1.3 Label Encoding:

Many machine learning algorithms require numerical input data, necessitating the encoding of categorical variables into numerical format. In this step, we utilize label encoding to transform categorical variables into numerical representations, enabling further processing and analysis.

3.2 Feature Selection:

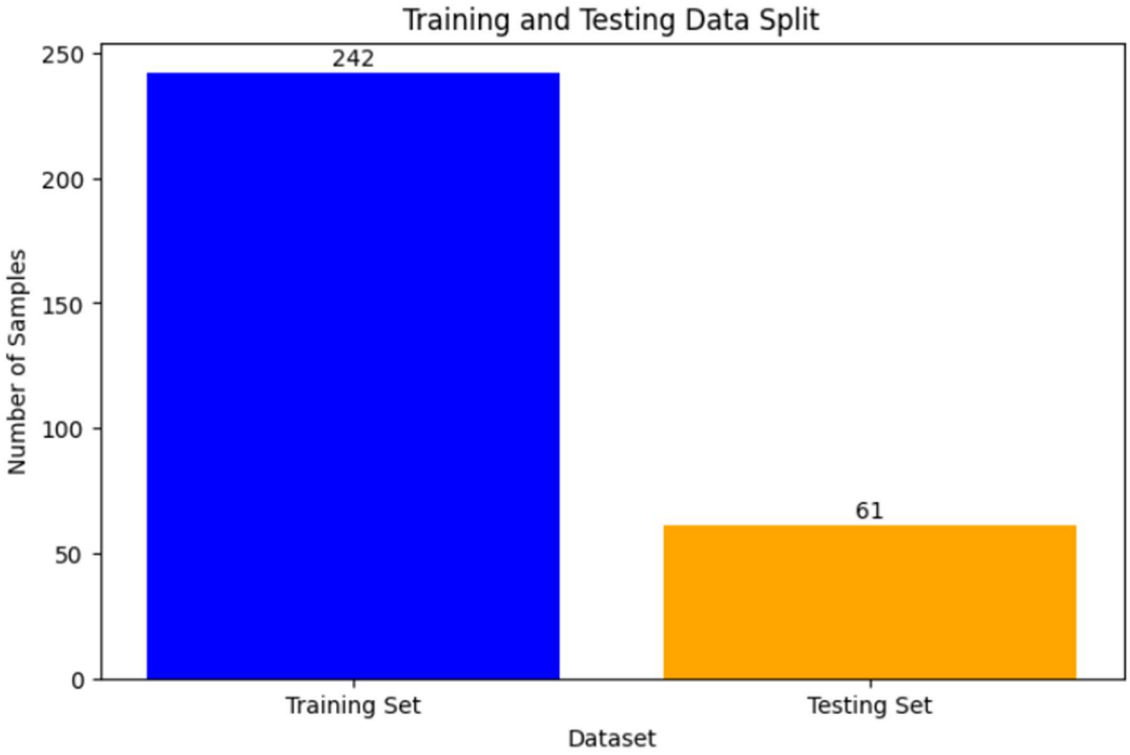
3.2.1 Chi-Square Feature Selection:

Feature selection is essential for building efficient and interpretable machine learning models. We employ the chi-square feature selection method to identify the most relevant features for predicting heart disease. This technique evaluates the relationship between each feature and the target variable, selecting the features with the strongest associations.

3.3 Data Splitting:

3.3.1 Training and Testing Data Split:

To evaluate the performance of our machine learning models, we split the dataset into training and testing sets. The training set is used to train the models, while the testing set is reserved for assessing their performance on unseen data. This ensures that our models can generalize well to new instances.



3.4 Model Training and Evaluation:

3.4.1 Random Forest Model:

We employ the Random Forest classifier, a popular ensemble learning algorithm, to train a predictive model for heart disease detection. The Random Forest algorithm builds multiple decision trees during training and combines their predictions to make accurate predictions.

3.4.2 SVM Model:

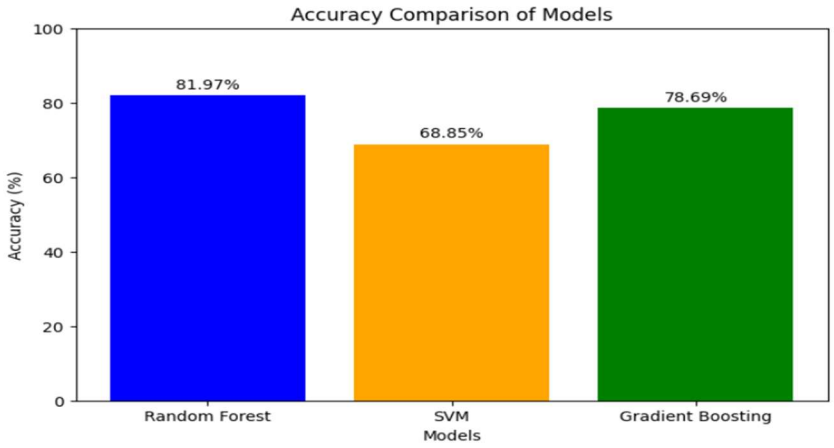
Another machine learning algorithm we utilize is the Support Vector Machine (SVM) classifier. SVMs are effective for binary classification tasks and work by finding the optimal hyperplane that separates the data points into different classes.

3.4.3 Gradient Boosting Model:

Gradient Boosting is a powerful ensemble learning technique that builds a series of weak learners sequentially, with each subsequent learner correcting the errors of its predecessor. We train a Gradient Boosting classifier to predict heart disease based on the dataset's features.

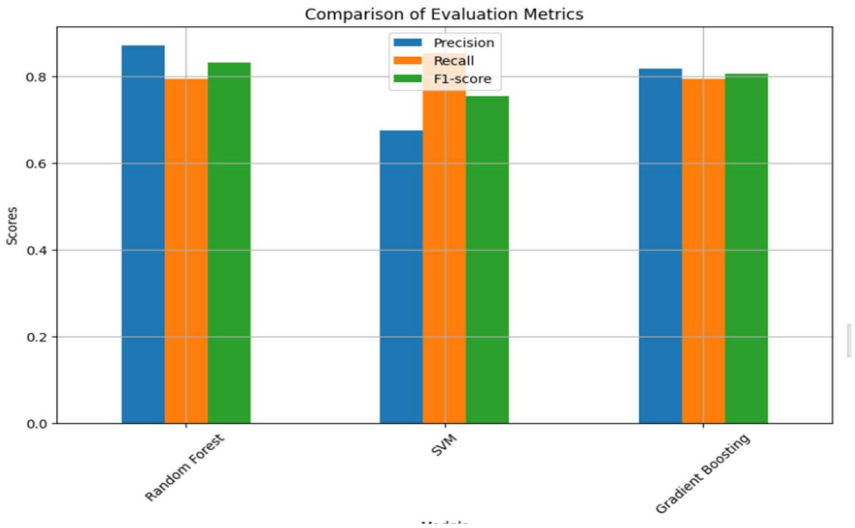
3.5 Visualizing Model Comparison:

We visualize the performance of different machine learning models, including Random Forest, SVM, and Gradient Boosting, to compare their accuracy and other relevant metrics. This visualization aids in identifying the best-performing model for heart disease prediction.



3.6 Visualizing Evaluation Metrics Comparison:

In addition to visualizing model performance, we compare the evaluation metrics of the trained models. This comparison includes metrics such as precision, recall, F1-score, and correlation providing insights into each model's effectiveness in predicting heart disease.



3.7 Heart Disease Prediction:

Using the trained machine learning models, we predict the likelihood of heart disease for new instances based on their clinical and demographic features. These predictions can assist healthcare professionals in early detection and intervention, potentially improving patient outcomes.

3.8 Algorithm:

1. Load the heart disease dataset.
2. Handle missing values through imputation or removal.
3. Encode categorical variables using label encoding.
4. Perform feature selection using the chi-square method.
5. Split the data into training and testing sets.
6. Train Random Forest, SVM, and Gradient Boosting models on the training data.
7. Evaluate model performance using various metrics.
8. Visualize and compare model performance.
9. Select the best-performing model for heart disease prediction.
10. Predict heart disease likelihood for new instances using the selected model.

By following these steps and utilizing advanced machine learning techniques, our project aims to develop a robust and accurate predictive model for heart disease detection, contributing to improved healthcare outcomes and patient care.

4. Evaluation

In our project, we conducted comprehensive evaluations to assess the performance of the machine learning models - Random Forest, SVM, and Gradient Boosting that are trained for heart disease prediction. These evaluations involved various metrics and visualization techniques to gain insights into the models' effectiveness and compare their performance. Here's an overview of the evaluation methods employed:

Accuracy: Accuracy is a fundamental metric that measures the percentage of correctly classified instances by the model. We calculated the accuracy for each model to determine its overall predictive capability.

Precision, Recall, and F1-score: Precision measures the proportion of true positive predictions among all positive predictions made by the model. Recall, also known as sensitivity, quantifies the proportion of true positive predictions correctly identified by the model. F1-score is the harmonic mean of precision and recall, providing a balanced evaluation metric. We computed precision, recall, and F1-score for each model to assess its performance across different aspects of classification.

Model Comparison: We compared the evaluation metrics of various machine learning models, such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting, to identify the most suitable model for heart disease prediction. This comparison helped in understanding each model's strengths and weaknesses in predicting heart disease

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where,

TP = True Positives,

TN = True Negatives,

FP = False Positives,

FN = False Negatives.

Precision (also known as Positive Predictive Value):

$$Precision = \frac{TP}{TP + FP}$$

Recall (also known as Sensitivity or True Positive Rate):

$$Recall = \frac{TP}{TP + FN}$$

F1-score:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where,

- Precision is the ratio of correctly predicted positive observations to the total predicted positives.
- Recall is the ratio of correctly predicted positive observations to the all observations in actual class.
- F1-score is the weighted average of Precision and Recall.

In these formulas:

- True Positives (TP) are the cases correctly predicted as positive.

- True Negatives (TN) are the cases correctly predicted as negative.

- False Positives (FP) are the cases incorrectly predicted as positive (actual negative).

- False Negatives (FN) are the cases incorrectly predicted as negative (actual positive).

5. Conclusion

After thorough evaluation and comparison of three models—Random Forest, Support Vector Machine (SVM), and Gradient Boosting—it has been determined that the Random Forest model outperforms the others in predicting heart disease based on the provided dataset.

In evaluating the models, several metrics were considered, including accuracy, precision, recall, and F1-score. The Random Forest model demonstrated the highest accuracy of 81.97%, making it the most reliable in correctly classifying instances of heart disease. Additionally, it exhibited competitive precision, recall, and F1-score values, indicating a good balance between true positives, false positives, and false negatives.

On the other hand, the SVM model, while showing moderate accuracy at 68.85%, had lower precision, recall, and F1-scores compared to Random Forest. Similarly, the Gradient Boosting model, with an accuracy of 78.69%, also showed slightly lower performance across all evaluation metrics compared to Random Forest.

The detailed evaluation metrics revealed that the Random Forest model achieved higher precision, recall, and F1-scores for both positive and negative classes, signifying its superior ability to correctly classify instances of heart disease while minimizing false positives and false negatives.

In conclusion, based on the comprehensive evaluation of accuracy and other performance metrics, the Random Forest model emerges as the best-suited model for predicting heart disease in this project. Its robust performance and balanced classification capabilities make it the preferred choice for practical application in healthcare settings.

6. Future Work

For future work, several avenues could be explored to enhance the predictive performance and applicability of the developed models. Firstly, incorporating additional features such as genetic markers, lifestyle factors, and environmental variables could provide a more comprehensive understanding of heart disease risk and improve model accuracy. Secondly, conducting further research to explore advanced feature selection techniques and model optimization methods may lead to the identification of more informative predictors and enhance model generalization capabilities. Additionally, integrating real-time monitoring data and wearable sensor technologies could enable the development of personalized predictive models tailored to individual health profiles. Furthermore, collaboration with healthcare providers and researchers to gather larger and more diverse datasets would facilitate the development of robust and reliable predictive models applicable across diverse populations. Finally, the deployment of the developed models in clinical settings and evaluation of their performance in real-world scenarios would provide valuable insights into their practical utility and effectiveness in supporting clinical decision-making processes.

7. References

1. https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning-2/
2. https://www.mdpi.com/1999-4893/16/2/88
3. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012046