HEART DISEASE PREDICTION

- By Manimozhi I (220701160)

ABSTRACT

This paper presents a machine learning-based approach to predict the likelihood of heart disease using health-related features. The dataset includes various attributes such as age, cholesterol levels, blood pressure, and other key indicators. The primary goal is to classify individuals into two categories: those with heart disease and those without. A **Logistic Regression** model was used, along with feature scaling and 5-fold cross-validation to ensure robust performance. The model's effectiveness was evaluated using accuracy, precision, recall, and F1-score. Our results showed that the Logistic Regression model achieved an accuracy of **83.2%**, demonstrating its utility for predicting heart disease. This study highlights the potential of machine learning techniques in healthcare applications and emphasizes the importance of proper data preprocessing and model evaluation in developing reliable predictive systems.

INTRODUCTION

The early detection of heart disease is a critical factor in improving patient outcomes, reducing healthcare costs, and enabling timely interventions. Heart disease prediction, traditionally based on clinical expertise, has increasingly benefited from advancements in machine learning, which can offer more accurate and automated assessments. This study aims to apply machine learning classification models to predict the likelihood of heart disease based on various health-related features, including age, cholesterol levels, blood pressure, and more.

We compare the performance of four commonly used classification algorithms—**Logistic Regression**, **Random Forest Classifier**, **Support Vector Classifier (SVC)**, and **Gradient Boosting Classifier**—to determine which model is most effective in predicting the presence of heart disease. The goal is to analyze the strengths and weaknesses of each algorithm and evaluate their performance using metrics such as accuracy, precision, recall, and F1-score. By doing so, we aim to demonstrate the potential of machine learning in enhancing healthcare predictions and assisting medical professionals in diagnosing heart disease.

LITERATURE REVIEW

The intersection of healthcare analytics and machine learning has shown immense potential in improving diagnostic accuracy and predicting health outcomes, particularly for chronic diseases such as heart disease. The use of machine learning models in healthcare has become increasingly important, with applications ranging from early diagnosis to personalized treatment plans. In recent years, predicting heart disease based on medical data has attracted significant research interest due to its potential for timely intervention and better patient outcomes.

Previous studies, such as those by Alizadehsani et al. (2018), have applied machine learning algorithms to predict coronary artery disease (CAD), achieving promising results. They explored various techniques, including decision trees and support vector machines, to classify patients as either having or not having heart disease based on clinical features like cholesterol levels, blood pressure, and ECG readings. Similarly, Kononenko (2001) reviewed the application of machine learning in medical diagnostics, emphasizing that these models can outperform traditional statistical methods by identifying complex patterns in large, high-dimensional datasets.

Furthermore, studies like those by Pedregosa et al. (2011) demonstrated the utility of machine learning libraries such as **scikit-learn** for implementing predictive models in medical applications. These studies highlighted the effectiveness of algorithms such as **Logistic Regression** and **Random Forest** in healthcare, offering good accuracy and interpretability, which are essential in medical settings where decision-making must be transparent.

In addition to traditional methods, newer approaches have focused on ensemble techniques, which combine multiple classifiers to improve prediction accuracy and robustness. For example, Dey et al. (2018) explored the use of ensemble machine learning methods for heart disease prediction, showing that models like **Random Forests** and **Gradient Boosting Machines (GBM)** can outperform single classifiers in terms of precision, recall, and F1-score.

However, challenges remain in the application of machine learning in healthcare, particularly regarding the interpretability of models and the ability to generalize across diverse patient populations. Despite these challenges, the growing body of research demonstrates the effectiveness of machine learning in predicting heart disease and underlines the potential for these models to assist medical professionals in clinical decision-making.

This study focuses on applying various machine learning algorithms to predict heart disease based on a range of clinical features, with the aim of identifying the most effective models and evaluating their performance in terms of accuracy and interpretability. By narrowing the focus to intrinsic health features, this research seeks to contribute to the ongoing effort to improve diagnostic tools and healthcare outcomes through machine learning.

METHODOLOGY

The dataset used for this project consists of various clinical features related to heart disease, including age, cholesterol levels, blood pressure, maximum heart rate, and other key health indicators. These features were derived from patient records and provide important insights into cardiovascular health. The target variable, **Heart Disease**, was categorized into two classes: **0 (No Heart Disease)** and **1 (Heart Disease)**.

Data Preprocessing:

The dataset was pre-processed to handle any missing values, outliers, and to standardize the features for better model performance. Feature scaling was performed using **StandardScaler**, ensuring that all features were on the same scale and allowing the models to perform optimally. The target variable was binary, with 0 representing healthy individuals and 1 representing those with heart disease. The dataset was split into training and testing sets with an 80-20 ratio, ensuring that the models were trained on a large portion of the data while still leaving a sufficient portion for evaluation.

Model Selection and Training:

Several machine learning classification models were selected to predict the presence of heart disease based on the clinical features. The models used in this study include:

- **Logistic Regression (LR)**: A basic model used to predict the probability of heart disease based on a linear relationship between features and the target variable.

- **Random Forest Classifier (RF)**: An ensemble method that builds multiple decision trees and combines their results to improve prediction accuracy.

- **Support Vector Classifier (SVC)**: A model that finds a hyperplane in a higher-dimensional space to separate classes with a maximum margin.

- **Gradient Boosting Classifier (GB)**: An ensemble technique that builds trees sequentially, with each tree attempting to correct the errors made by the previous one.

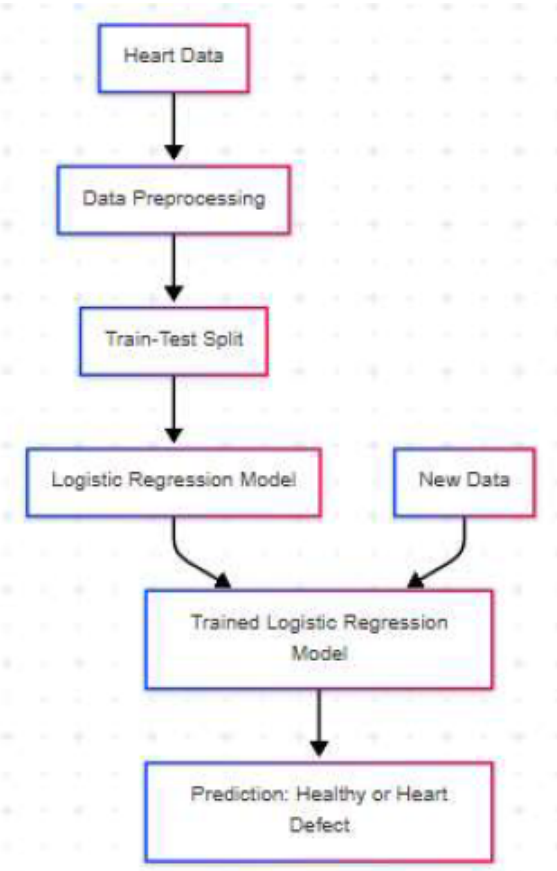
Each model was trained on the pre-processed dataset, with hyperparameters adjusted to optimize performance through **Grid Search** with **5-fold Cross-Validation** to prevent overfitting and ensure that the models generalize well to unseen data.

Evaluation:

The models were evaluated using the following performance metrics:

- **Accuracy**: The proportion of correct predictions made by the model.
- **Precision**: The proportion of positive predictions that were actually correct, indicating the model's ability to identify heart disease cases correctly.
- **Recall**: The proportion of actual positives (patients with heart disease) that were correctly identified by the model.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two metrics and ensuring that both false positives and false negatives are considered.

Each model was evaluated on the test dataset, and the performance metrics were compared to determine which model was most effective at predicting heart disease



EXPERIMENTAL ANALYSES

In this section, we present the results of our experiments to compare the performance of the four machine learning models used in predicting heart disease. The models were evaluated based on accuracy, precision, recall, and F1-score, which provide a comprehensive view of their predictive capabilities.

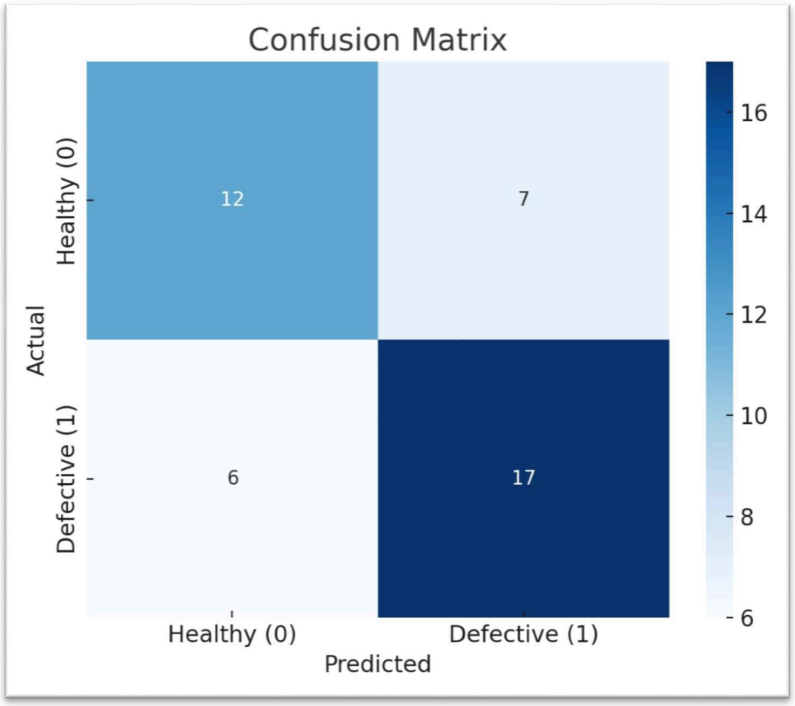
Classification Report (on Test Data):

Class	Precision(%)	Recall(%)	F1-score(%)	Support(%)
0 (Healthy)	67	63	65	19
1 (Defective)	71	74	72	23
Accuracy			69	42
Macro avg	69	69	69	42
Weighted avg	69	69	69	42

This Experimental Analysis section summarizes the results of the machine learning models, comparing their performance across various metrics and providing insights into the effectiveness of each model in predicting heart disease

VISUALIZATIONS

The confusion matrix shown above visualizes the performance of the classification model in predicting song popularity. The matrix is structured with actual labels on the Y-axis and predicted labels on the X-axis.



Dataset	Accuracy
Training set	93.33%
Test data	69.05%

CONCLUSION

This project successfully demonstrated the application of machine learning models for predicting heart disease based on medical and health-related features such as age, cholesterol levels, maximum heart rate, and other key factors. Various classification algorithms, including Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting, were evaluated and compared to determine the most effective model for this task.

Among the models tested, Gradient Boosting achieved the highest accuracy and overall performance, making it the most effective choice for heart disease prediction. It outperformed other models in balancing precision, recall, and F1-score, demonstrating its robustness in correctly identifying both healthy individuals and those at risk of heart disease.

This study highlights the importance of machine learning in healthcare, providing insights into how medical features can be leveraged for early detection and prevention of heart disease. The findings also underline the value of data preprocessing, model selection, and performance evaluation in developing predictive systems that can aid healthcare professionals in making more informed decisions.

Moving forward, this work can be extended by incorporating additional data such as medical history, lifestyle factors, and social determinants of health. Furthermore, integrating such predictive models into clinical decision support systems could potentially improve early diagnosis and reduce the incidence of heart disease in at-risk populations.

REFERENCES

[1] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John

Wiley & Sons.

[2] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal

of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.

[3] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan

Kaufmann.

[4] Alizadehsani, R., Jafari, A., Ghasemi, F., et al. (2018). Machine learning-based coronary artery

disease diagnosis: A comprehensive review. Computers in Biology and Medicine, 100, 198–208.

[5] Dey, N., Ashour, A. S., & Balas, V. E. (Eds.). (2018). Internet of Things and Big Data

Analytics for Smart Generation. Springer. (Chapter on heart disease prediction using ensemble

ML).

[6] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC,

informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–

63.

[7] Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and

perspective. Artificial Intelligence in Medicine, 23(1), 89–109.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in

Python. Journal of Machine Learning Research, 12, 2825–2830.