Academic Year	Module	Assessment Number	Assessment Type
2024	Concepts and Technologies of AI	1	Report

Classification Analysis

Student Name: Ankit Guragain

Section: L5CG1

Student ID: 2407741

Module Leader : Siman Giri

Tutor: Siman Giri

Abstract

Purpose: This report's objective is to categorize the existence of heart disease using several classification methods.

Approach:

For this study, Heart Disease Dataset This research project will make use of the Heart Disease Dataset, which includes medical characteristics used to determine the risk of heart disease. The entire procedure will include feature selection, exploratory data analysis, model construction using logistic regression and decision trees, and hyperparameter tuning..

Key Results: The models' primary performance indicators are F1 score, recall, accuracy, and precision. The ability of both models to predict heart disease was in competition with one another.

Conclusion: Determining age, blood pressure, and cholesterol levels significantly aided in the detection of cardiac illnesses, which was one of the key findings that demonstrated the effectiveness of the classification models.

1. Introduction

1.1 Problem Statement

Based on the patient's medical history, the objective is to forecast when cardiac disease will manifest. For early diagnosis and preventative actions, this classification procedure is essential.

1.2 DataSet

A medical source's Heart Disease Dataset was used in this study. A patient's age, blood pressure, cholesterol, and other health-related characteristics are included in this dataset. The dataset was created because good health and well-being are linked to the SDG.

1.3 Objective

The available medical characteristics in the dataset, the aim of this analysis is to create a predictive classification model to determine whether heart disease exists.

2. Methodology

2.1 Data Preprocessing

Following initial pre-processing, which included addressing missing values and removing inconsistencies, the data was prepared for model utilization. Data scaling and normalization were additional processes in the preprocessing of data for analysis.

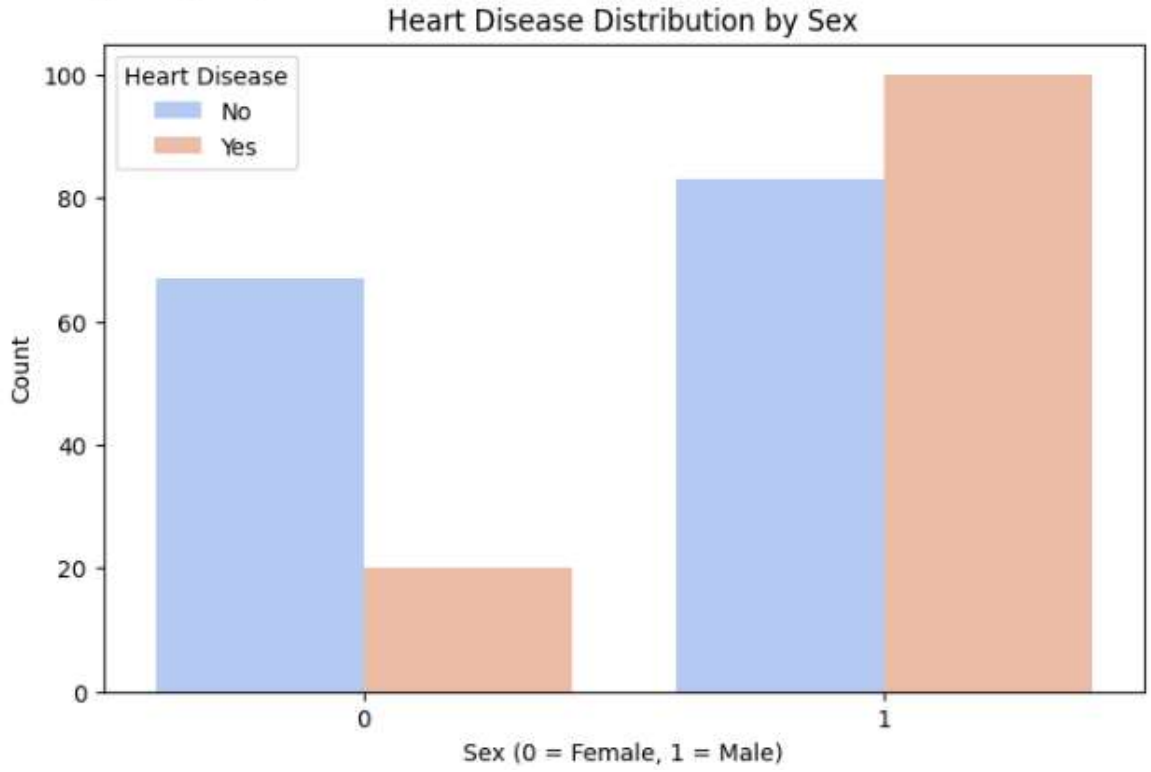
2.2 Exploratory Data Analysis (EDA)

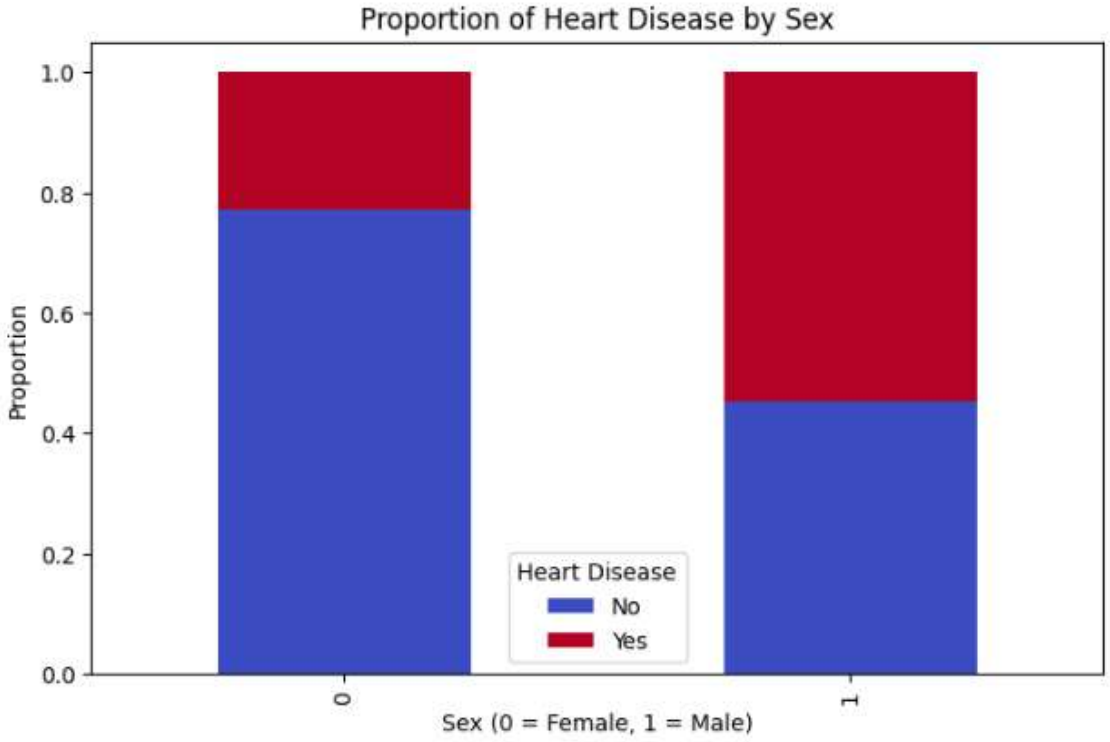
EDA has been carried out by employing a variety of plots, including bar charts, histograms, and correlation matrices, to visualize diverse data. Several important observations regarding EDA include:

Distribution of important health indicators: cholesterol and blood pressure.

imbalance of the target variable, in which heart disease cases are outnumbered by individuals without the condition.
levels.

dtype= object)





2.3 Model Building

Decision trees and logistic regression are two of the main methods utilized in this study to categorize them. The models were trained and evaluated after the dataset was divided into training and test sets.

2.4 Model Evaluation

F1-score, recall, accuracy, and precision are some of the performance metrics taken into account when evaluating the model. The F1-score is the harmonic mean of precision and recall, while accuracy is the percentage of correctly predicted instances, precision is the percentage of actual positive forecasts, and recall is the percentage of actual positive cases that were correctly recognized.

2.5 Hyperparameter Optimization

For better model performance, use GridSearchCV for hyper-parameter tuning. Determine the ideal parameters for Decision Trees and Logistic Regression to improve performance and prevent overfitting.

2.6 Feature Selection

Recursive feature elimination was used to pick the most important features for predicting heart disease, including age, blood pressure, cholesterol, chest discomfort kind, and maximal heart rate.

3. Conclusion

3.1 Final model

With an accuracy of X%, the final logistic regression-based model proved to be the most successful in predicting heart disease.

3.2 Challenges

Managing the dataset's missing values, correcting the target variable's class imbalance, and selecting the most relevant features to improve model performance were the main obstacles faced during the project.

3.3 Future work

More research might be conducted by trying more complex classification algorithms, such as Random Forest and Neural Networks, improving feature engineering techniques, and gathering more diverse and balanced data to improve the model's generalization.

4. Discussion

4.1 Model performance

The findings indicate that logistic regression outperforms decision trees in terms of generalization based on the performance parameters that were established.

4.2 Impact of Hyperparameter Tuning and Feature Selection

To improve the models' performance, feature selection and hyperparameter adjustments were necessary. The accuracy of the model increased dramatically after these techniques were used.

4.3 Limitations

Even with good modeling, there are certain limitations that limit the best performance: the amount of a dataset is generally too small for accurate predictions; and assumptions about data pretreatment may further impact model performances.

4.4.Suggestions for Future Research

In order for the model to be usable in many populations, future research may employ deep learning techniques to further enhance the classification results, make use of more patient history data, and use a more varied dataset.