Predicting Heart Disease Risk Using a Random Forest Classifier: A Machine Learning Approach

Nur Aisyah Athirah binti Hishamuddin

Abstract

Heart disease remains a leading cause of mortality in Malaysia thus preventing and treating it early on are essential to lowering the condition's linked death rates. This project aims to develop a machine learning-based diagnostic system for predicting heart disease risk using a Random Forest Classifier. The dataset, sourced from the CDC and preprocessed for quality, includes demographic, clinical, and lifestyle indicators. Various techniques, such as random imputation and KNN, were employed to handle missing values, while outliers were addressed using IQR and Z-score methods. Feature selection was performed using LassoCV, and the selected features were used to train the Random Forest model. The model demonstrated excellent performance with an accuracy of 0.9740, precision of 0.9798, recall of 0.9679, F1 score of 0.9738, and ROC AUC of 0.9960. However, the presence of false positives and false negatives indicates areas for improvement. Implementing this model in a real-life healthcare setting via a user-friendly platform can assist doctors in early diagnosis and personalized care, while policymakers can utilize aggregated data for informed decision-making.

Keywords : Heart Disease Prediction, Machine Learning, Random Forest Classification

INTRODUCTION

Heart disease refers to a condition affecting the heart or blood vessels.[1] It is caused by the build-up of cholesterol and other substances in the blood vessels, leading to their narrowing and hardening. This can reduce blood flow to the heart, potentially causing chest pain and, in severe cases, triggering a heart attack.[2] According to the Statistics on Causes of Death in Malaysia for 2023, heart disease continues to be the leading cause of death in the country.[3] Early detection and prevention are crucial for lowering the death rate linked with heart disease.

The aim of this project is to create a machine learning-based diagnostic system for heart disease. The primary focus is to develop a model that ensures high accuracy, sensitivity, and a low false positive rate. This is crucial for accurately identifying those who are at risk and reducing needless worry and medical procedures. A thorough analysis that incorporates important factors influencing the risk of heart disease is made possible by the use of a large dataset. The project also seeks to offer insightful information to patients, politicians, and healthcare professionals. Enabling early interventions and customised healthcare strategies is the aim, as it has the potential to save lives and lower healthcare expenses.

Ultimately, the goal is to offer a reliable tool for predicting heart disease, with the aim of making a substantial impact on the healthcare sector. This initiative seeks to enhance patient outcomes and preventive care practices.

METHODOLOGY

The upcoming sections will provide insight into the project's methodology. For more detailed implementation details, the code can be accessed on GitHub (https://github.com/Aisyah-Athirah/heart-disease).

The dataset utilized in this project is a CSV file named "heart_2022_with_nans.csv". The dataset was sourced from Kaggle.[4] It comprises essential personal indicators associated with heart disease, obtained from the Centres for Disease Control and Prevention (CDC)[5]. A variety of demographic, clinical, and lifestyle characteristics are included in the dataset, such as age, height, sex, smoking status, and other possible heart disease-related indicators. The dataset includes numerical parameters like age and height, as well as qualitative criteria like the 'HadDiabetes' and 'HadAngina' columns with Yes or No replies.

For starter of the execution, essential libraries were imported, incorporating NumPy for numerical computations, Matplotlib and Seaborn for data visualization, and Scikit-learn for machine learning tasks. These tools play a crucial role in model development, data exploration, pre-processing, and performance assessment.

The initial crucial phase of the methodology focuses on extensive data pre-processing to ensure the dataset is prepared for modelling. Exploratory Data Analysis (EDA) and data preprocessing began with an overview of the dataset to obtain a concise summary of the data, including the number of non-null entries and data types for each column. The dataset's dimensions were inspected by viewing the number of rows and columns, identify and remove any duplicate records. To ensure the dataset's quality, columns with less than 1% missing values were identified and retained. Subsequently, several columns deemed unnecessary for the analysis were dropped.

The cleaned dataset was then copied into a new DataFrame named df1 to preserve the original data for reference. This step concluded the initial phase of data preprocessing, establishing a refined dataset ready for subsequent analysis.

The process of handling missing values in the dataset involved multiple steps. First, the random_impute function was defined to replace missing values with random selections from the non-null entries. This was applied to PneumoVaxEver, AlcoholDrinkers, AgeCategory, and SleepHours, resulting in no null values in these columns. For categorical variables like CovidPos, SmokerStatus, MentalHealthDays, RaceEthnicityCategory, and HighRiskLastYear, bar charts were plotted to visualize their distributions. Missing values in these columns were filled using the mode, ensuring no remaining nulls. Continuous variables HeightInMeters, WeightInKilograms, and BMI were handled differently. Histograms and normal distribution curves were plotted to visualize these variables. Random imputation was used for HeightInMeters, successfully removing null values. K-Nearest Neighbors (KNN) imputation was applied to WeightInKilograms and BMI, effectively eliminating null values. This comprehensive approach ensured a complete dataset ready for further analysis.

The outliers were identified using boxplots and heatmaps, which allowed for the visual detection and extraction of anomalies that could adversely impact model performance. A heatmap was also used to identify missing data across the features, facilitating effective data cleaning. The process of handling outliers in the dataset involved several steps. Outliers were identified using the Interquartile Range (IQR) method for some columns such as BMI, PhysicalHealthDays, MentalHealthDays, and SleepHours. This method calculated the IQR and established lower and upper limits to identify data points outside these bounds as outliers. The identified outliers were then removed from the dataset, and the DataFrame was reindexed to maintain consistency. For the HeightInMeters column, outliers were managed using the Z-score method. Values with Z-scores greater than 3 or less than -3 were capped at the upper and lower limits based on three standard deviations from the mean. After this adjustment, the Z-score

column was removed. The shape of the data frame was checked before and after handling outliers to ensure data integrity. Finally, box plots for each numeric column were generated to visually confirm the removal of outliers, resulting in a clean dataset ready for further analysis.

To address class imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was applied. This method generates synthetic samples to balance the class distribution, thereby improving the robustness of the model. The class distribution was visualized using bar graphs before and after the application of SMOTE to illustrate the effectiveness of the technique.

For feature selection, the MinMaxScaler was initialized with values within the range of 0 to 1 to standardize the data. After scaling, the features and target variable for feature selection were defined, with the target variable being HadHeartAttack, which was mapped to binary values. A LassoCV model, using cross-validation with 5 folds and a range of alpha values, was trained to identify significant features. The model was fitted on the dataset with the target variable. Features with non-zero coefficients in the Lasso model were considered significant. These significant features, along with their corresponding coefficients, were identified. To visualize the importance of these significant features, a bar plot was created, displaying the coefficient values of the significant features. This plot helped in understanding the influence of each feature on the target variable, HadHeartAttack, providing insights into which features were most impactful in predicting heart attacks.

Model training was conducted using a Random Forest classifier. Initially, the features and target variable for model training were defined, with the selected significant features serving as the predictors. The dataset was then split into training and testing sets, with 80% of the data used for training and 20% for testing, ensuring that the data was numeric for compatibility with the model. A Random Forest classifier, configured with 100 estimators and a fixed random state for reproducibility, was initialized. The classifier was trained using the training dataset, fitting the model to the training features and target variable. Once trained, the model was used to make predictions on the test data, providing a basis for evaluating its performance in predicting the target variable. This process helped to ensure that the model was properly trained and capable of making accurate predictions.

Model evaluation was conducted to assess the performance of the Random Forest classifier. The evaluation metrics included accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC).

RESULT

The results section synthesizes the performance outcomes of the model, highlighting its strengths and weaknesses based on the employed evaluation metrics. Emphasis is placed on identifying the model's predictive accuracy and robustness across various metrics.

Table 1 : Score for each evaluation metrics

Evaluation metrics	Score
Accuracy	0.9740
Precision	0.9798
Recall	0.9679
F1-score	0.9738
ROC AUC	0.9960

The performance of the RandomForestClassifier in predicting heart disease was evaluated using various metrics, each providing insights into different aspects of the model's capabilities.

An overall accuracy was 0.9798. Precision measures the proportion of true positive predictions among all positive predictions made by the model. A precision score of 0.9798 means that when the model predicts a positive instance, it is correct

about 97.98% of the time, highlighting its ability to avoid false positives.

The precision score was 0.9798. Precision measures the proportion of true positive predictions among all positive predictions made by the model. A precision score of 0.9798 means that when the model predicts a positive instance, it is correct about 97.98% of the time, highlighting its ability to avoid false positives.

The recall score was 0.9679. Recall measures the proportion of actual positive instances that the model correctly identified. With a recall of 0.9679, the model successfully identified 96.79% of all actual positive instances, indicating its effectiveness in capturing the true positive cases.

The F1-score was 0.9738, which is the harmonic mean of precision and recall. The F1-score provides a single metric that balances both precision and recall. A score of 0.9738 reflects the model's overall effectiveness in making accurate and complete positive predictions

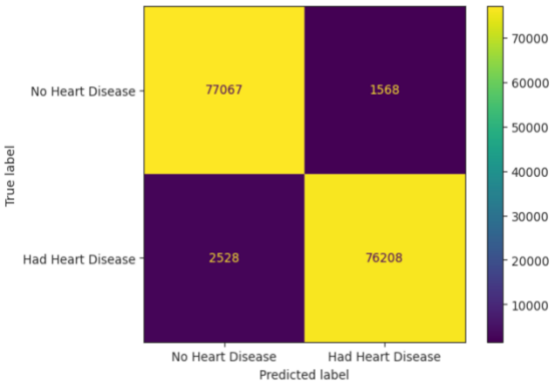

Figure 1: Display for Confusion Matrix

The confusion matrix provided a detailed breakdown of the model's predictions. It showed that the model correctly predicted 77,067 instances of no heart disease and 76,208 instances of heart disease. This high number of correct predictions highlights the model's overall accuracy and reliability. Additionally, the matrix revealed that there were 1,568 false positives, where the model incorrectly predicted heart disease when there was none. This indicates areas where the model could potentially be improved to reduce these incorrect classifications. On the other hand, there were 2,528 false negatives, where the model failed to identify actual cases of heart disease. These false negatives are critical as they represent missed cases of heart disease, which could have significant implications. Overall, the confusion matrix provided a comprehensive view of the model's performance, illustrating both its strengths in accurate predictions and the areas needing improvement to reduce misclassifications. This detailed analysis is essential for understanding the effectiveness and reliability of the model in practical applications.

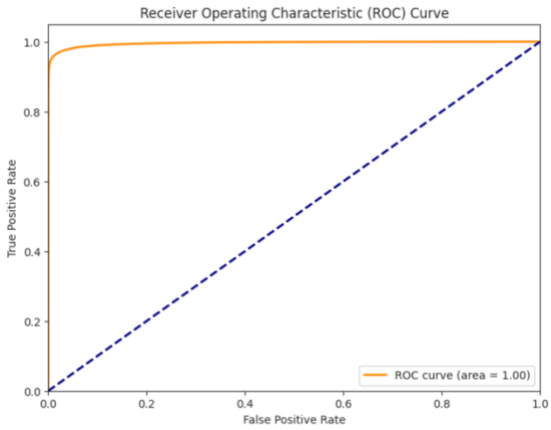

Figure 2: Receiver Operating Characteristic (ROC) Curve

The ROC AUC score of 1.0 demonstrates the model's strong ability to distinguish between patients with and without heart disease. This high score indicates that the model has a good balance of sensitivity (true positive rate) and specificity (true negative rate), making it effective in clinical settings where distinguishing between conditions is vital.

CONCLUSION

The evaluation of the Random Forest Classifier model for predicting heart

disease has demonstrated excellent results, with an overall accuracy of 0.9740. The model exhibited strong performance across key metrics, with a precision of 0.9798, recall of 0.9679, and an F1 score of 0.9738, indicating a robust ability to make correct predictions. The ROC AUC score of 0.9960 further underscores the model's exceptional capability to distinguish between patients with and without heart disease.

However, the analysis also revealed areas for improvement. The presence of 1,568 false positives and 2,528 false negatives in the confusion matrix indicates some shortcomings. False positives can lead to unnecessary anxiety and additional testing, while false negatives represent missed diagnoses, potentially delaying treatment. Future work should focus on enhancing recall and reducing false negatives, possibly by integrating additional features, employing more sophisticated modeling techniques, or improving data quality and quantity.

To implement this model in a real-life healthcare setting, a user-friendly web-based platform or mobile application can be developed. This platform could allow doctors to input patient data and receive diagnostic predictions and treatment recommendations, aiding in early diagnosis and personalized care. Patients could use the app to enter symptoms and medical history, receiving preliminary assessments and guidance. Policymakers could use aggregated, anonymized data from the app to identify trends and make informed healthcare policy decisions.

REFERENCES

[1] National Academies Press (US). (2010). *Ischemic heart disease*. Cardiovascular Disability - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK209964/

[2] Portal MyHEALTH. (2016, August 18). *Coronary artery Disease - PORTAL MyHEALTH*. PORTAL MyHEALTH. http://myhealth.moh.gov.my/en/coronary-artery-disease

[3] Eecpadmin. (2024, May 14). The top causes of death in Malaysia 2023. *EECP Centre Malaysia - Treatment for Heart, Stroke, and Neurodegenerative Diseases*. https://www.eecpcentre.com/news/top-causes-of-death-in-malaysia-2023/

[4] *Indicators of heart Disease (2022 UPDATE)*. (2023, October 12). Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data?select=2022

[5] *Heart disease risk factors*. (2024, May 15). Heart Disease. https://www.cdc.gov/heart-disease/risk-factors/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/risk_factors.html