Heart Disease Prediction

Introduction

Heart disease is one of the leading causes of death globally, with millions of individuals affected each year. Early detection of heart disease can significantly reduce mortality rates by allowing timely intervention and treatment. However, diagnosing heart disease can be challenging due to the variety of factors and symptoms associated with it. Traditionally, doctors rely on clinical data and diagnostic tests to predict whether a patient is at risk. In recent years, advancements in machine learning have offered an opportunity to automate and enhance the diagnostic process by predicting heart disease based on patient data.

The **Heart Disease Prediction Project** aims to build a machine learning model that can accurately predict the likelihood of a patient having heart disease based on a range of clinical features, including age, cholesterol levels, chest pain type, and others. The goal of this project is to assist healthcare professionals in making more informed decisions by providing an automated tool that can flag patients at risk of heart disease.

Objectives

The main objectives of this project are:

1. **To analyze and preprocess a dataset containing patient medical records** and related features for heart disease diagnosis.
2. **To build multiple machine learning models** and compare their performance in terms of accuracy, precision, recall, and other relevant metrics.
3. **To select the best-performing model** and evaluate it thoroughly to ensure that it meets the clinical needs of the stakeholders.
4. **To deploy the model through an interactive web interface** that allows healthcare professionals to input patient data and receive real-time predictions.

The project involves exploring the dataset through Exploratory Data Analysis (EDA), selecting and engineering relevant features, and training and testing several machine learning models. Ultimately, the selected model will be deployed in a user-friendly application, allowing medical staff to make quick and informed predictions.

The significance of this project lies in its potential to **reduce the burden on healthcare professionals**, minimize diagnostic errors, and provide patients with a **more accurate and timely diagnosis**, thus improving outcomes and saving lives. By leveraging data-driven methods, this project can enhance decision-making in healthcare and contribute to the broader adoption of machine learning in medical diagnostics.

Stakeholder and Their Problem

The stakeholder for this project is a healthcare organization aiming to improve the early detection and prevention of heart disease. This is critical because cardiovascular diseases remain one of the leading causes of death globally. Early diagnosis and prediction of heart disease can significantly improve patient outcomes and reduce healthcare costs by enabling timely interventions. The organization is seeking a machine learning model that can accurately predict whether a patient has heart disease based on a range of medical attributes and indicators, such as cholesterol levels, blood pressure, and other risk factors. The model needs to be explainable and interpretable so that medical professionals can trust and understand its predictions.

The Problem to Solve

The main problem the healthcare organization is trying to solve is how to predict the likelihood of heart disease in patients based on various medical features. The goal is to build a predictive model that can analyze a patient's medical data and provide an indication of whether or not they are at risk of heart disease. With this model, the organization aims to streamline the diagnosis process and assist healthcare providers in identifying high-risk patients earlier, thereby enabling more personalized and effective treatments.

Predicting heart disease is a challenging task because it involves interpreting multiple features that may interact in complex ways. Medical data often contains noise and variability, which makes building a robust model difficult. The stakeholders are particularly concerned with minimizing both false positives and false negatives, as misclassification in either direction could lead to unnecessary treatments or missed opportunities for early intervention.

Dataset Information

The dataset used in this project is the **heart disease dataset**, a well-known benchmark dataset for predicting cardiovascular conditions. The dataset consists of more than 10 thousand samples with 14 attributes, including age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate, and more. The target variable is a binary indicator where "1" represents the presence of heart disease and "0" indicates its absence.

The dataset was obtained from the Kaggle repository and can be found here:

Dataset Link: https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci

Exploratory Data Analysis (EDA)

Before diving into model building, conducting **Exploratory Data Analysis (EDA)** is essential to understand the dataset's structure and characteristics. EDA helps to identify patterns, spot anomalies, and guide feature selection and engineering.

Key Steps in EDA:

1. **Data Summary**:

2. The first step was to examine the structure of the dataset. The dataset contains 303 rows and 14 columns, with one target variable (target) and 13 predictor variables. The target variable is binary, indicating the presence (1) or absence (0) of heart disease.

Summary statistics were generated to get an initial sense of the dataset, including mean, median, minimum, and maximum values for each feature. For example:

   a. **Age**: Ranged from 29 to 77, with an average age of 54 years.
   b. **Cholesterol**: Had a wide range of values, from 126 to 564 mg/dL, suggesting varying cholesterol levels among patients.
   c. **Max Heart Rate**: Ranged from 71 to 202 bpm, indicating a broad spectrum of cardiovascular fitness and response to exercise.

3. **Missing Data**:

Upon checking for missing values, no significant missing data was found in the dataset. This meant that no imputation or data-cleaning procedures were necessary, streamlining the model-building process.

4. **Target Variable Distribution**:

The target variable (presence of heart disease) was checked for class balance. The dataset was relatively balanced, with 165 cases of heart disease and 138 cases without heart disease, ensuring that the models would not be biased toward predicting the majority class.

5. **Feature Distributions**:

The distribution of key features such as age, cholesterol, and blood pressure was visualized using histograms and density plots. For instance:

   a. **Age**: A bell-curve distribution was observed, with the majority of patients falling between 45 and 65 years.
   b. **Resting Blood Pressure**: Most patients had resting blood pressure between 120- and 140-mm Hg, which is on the borderline of normal and high.

c. **Cholesterol**: Cholesterol levels were right-skewed, indicating that a few patients had unusually high cholesterol levels.

6. **Feature Correlations**:

A correlation matrix was generated to identify relationships between the features and the target variable. Key insights included:

a. **Chest Pain Type (cp)** had a strong positive correlation with heart disease.

b. **Max Heart Rate Achieved (thalach)** showed a negative correlation with heart disease, meaning lower maximum heart rates were more associated with heart disease.

c. **ST Depression (oldpeak)** had a positive correlation with heart disease, indicating that greater depression in the ST segment was associated with a higher likelihood of heart disease.

7. **Visualizing Relationships**:

Scatter plots and box plots were used to explore the relationships between key features and the target variable. For example:

a. Patients with heart disease tended to have lower maximum heart rates and higher ST depression values.

b. Higher cholesterol levels were not a consistent indicator of heart disease, suggesting cholesterol alone might not be a strong predictor.

8. **Outliers**:

Outliers were identified using box plots, particularly in features like cholesterol and resting blood pressure. These outliers were retained for model training, as they could represent critical cases that the model needs to learn from.

Feature Introduction

In this project, 13 features were used to predict heart disease. Below is an introduction to some of the key features that significantly contributed to the predictive models:

1. **Age**:

Age is a crucial factor in heart disease risk, as older individuals tend to have a higher likelihood of cardiovascular problems. This feature is a continuous variable and was scaled during preprocessing.

2. **Sex**:

Sex is a binary categorical variable with values "Male" and "Female." Males are generally at higher risk for heart disease compared to females, which was reflected in the models' results. One-hot encoding was applied to this feature.

3. **Chest Pain Type (cp)**:

This categorical variable represents the type of chest pain the patient experiences. Chest pain is a key indicator of heart issues and is categorized into four types:

   a. 0: Typical angina
   b. 1: Atypical angina
   c. 2: Non-anginal pain
   d. 3: Asymptomatic Chest pain type had a strong correlation with heart disease in the EDA.

4. **Resting Blood Pressure (trestbps)**:

This feature records the patient's blood pressure while at rest. Elevated blood pressure is a known risk factor for heart disease. This continuous variable was scaled to ensure all features were on the same scale.

5. **Cholesterol (chol)**:

Cholesterol levels are a key factor in cardiovascular health, with higher levels typically indicating an increased risk for heart disease. This feature was retained as a continuous variable and scaled.

6. **Maximum Heart Rate Achieved (thalach)**:

This feature represents the highest heart rate the patient achieved during exercise testing. A lower maximum heart rate can indicate reduced cardiovascular capacity and was found to be an important predictor in the models.

7. **ST Depression Induced by Exercise (oldpeak)**:

Oldpeak measures ST segment depression during exercise relative to rest, indicating the level of ischemia (reduced blood flow to the heart). It was one of the most important features in predicting heart disease, with higher values associated with a greater risk.

8. **Exercise Induced Angina (exang)**:

This binary feature indicates whether the patient experienced exercise-induced angina (chest pain). Angina induced by exercise is a strong indicator of heart disease and was given high importance in the models.

9. **Fasting Blood Sugar (fbs)**:

This binary feature checks if the patient’s fasting blood sugar was greater than 120 mg/dL. High blood sugar levels are often associated with diabetes, a significant risk factor for heart disease.

Models Tried and Rationale

Several machine learning models were tested during the project. These models were chosen to explore both classical machine learning approaches and more modern neural network techniques:

1. **Logistic Regression**:

2. Logistic regression was selected as a baseline model. It is a simple yet powerful linear model used for binary classification problems like predicting the presence or absence of heart disease. Logistic regression is interpretable and provides clear insights into how each feature contributes to the prediction. Since the goal is medical prediction, interpretability is important for clinicians, and logistic regression offers that.

3. **Naive Bayes**:

Naive Bayes was chosen due to its simplicity and effectiveness in cases where features are conditionally independent. While the assumption of independence is rarely true in practice, Naive Bayes can still perform well in many applications. Its speed and relatively low computational cost make it attractive for medical diagnosis systems that may require real-time predictions.

4. **Multi-Layer Perceptron (MLP) Neural Network**:

An MLP neural network was used as a more complex, non-linear model. Neural networks can capture complex patterns and relationships in the data, which may not be possible with simpler models like logistic regression. However, they are more computationally intensive and less interpretable. Given that medical datasets often contain non-linear interactions between variables, an MLP was chosen to assess if a more sophisticated model could improve predictive performance.

Feature Selection and Engineering

Feature selection is a crucial aspect of building an effective model, as irrelevant or redundant features can introduce noise and degrade model performance. In this project, the following steps were taken:

1. **Feature Selection**:

All 13 features in the dataset were used in the initial stages of model training. Since the dataset is relatively small, we opted to retain all features to avoid losing potentially valuable information. However, careful attention was paid to features that may have higher predictive power, such as chest pain type, maximum heart rate, and exercise-induced angina.

2. **Feature Scaling**:

The dataset contained features with different scales (e.g., age, cholesterol levels, and blood pressure). Standardization was applied to ensure all features contributed equally to the model's performance. Standard Scaler was used to normalize the continuous features, which is especially important for distance-based models like neural networks.

3. **One-Hot Encoding**:

Categorical variables such as chest pain type, sex, and fasting blood sugar were one-hot encoded to allow the models to properly interpret and learn from these categorical inputs. This helped prevent the models from assuming any ordinal relationship between the categories.

4. **Feature Importance Analysis**:

After training the logistic regression model, feature importance was analyzed to identify the most critical features for predicting heart disease. The logistic model revealed that variables like chest pain type, maximum heart rate, and oldpeak (ST depression induced by exercise) were among the top predictors.

Model Evaluation and Metrics

Multiple evaluation metrics were used to assess the models' performance. Given the critical nature of the task—identifying patients at risk of heart disease—we emphasized the importance of both precision and recall in addition to overall accuracy.

1. **Accuracy**:

Accuracy is the proportion of correct predictions (both true positives and true negatives) out of all predictions. However, accuracy alone can be misleading, especially when dealing with imbalanced datasets like this one.

2. **Precision and Recall**:
    a. **Precision**: Precision measures the percentage of true positive predictions out of all positive predictions made by the model. This metric is crucial in healthcare applications, where false positives (misidentifying someone as having heart disease when they don't) can lead to unnecessary treatments.
    b. **Recall**: Recall measures the percentage of actual positives that were correctly identified by the model. In this case, high recall is important because missing a patient with heart disease (a false negative) could lead to serious consequences.
    c. **F1-Score**: The F1-score was also used to balance precision and recall. It provides a single metric that reflects the trade-off between these two.
3. **Confusion Matrix**:

A confusion matrix was generated to visualize the model's performance. It showed the number of true positives, true negatives, false positives, and false negatives, allowing for a deeper understanding of the model’s classification performance.

Future Work and Improvements

If more time and resources were available, the following improvements could be explored:

1. **Hyperparameter Tuning**:

More extensive hyperparameter tuning could be conducted using grid search or random search to improve the performance of the neural network and other models. For example, adjusting the number of layers and neurons in the MLP, or tuning regularization parameters in logistic regression, could yield better results.

2. **Cross-Validation**:

Implementing k-fold cross-validation would provide a more reliable estimate of model performance by ensuring the model is tested on different subsets of the data.

3. **Feature Engineering**:

Additional feature engineering could be performed to derive new features from the existing ones. For example, interaction terms between features like age and cholesterol level could be created to capture more complex relationships.

4. **Ensemble Methods**:

Combining multiple models in an ensemble, such as Random Forests or Gradient Boosting Machines, could potentially improve predictive accuracy by leveraging the strengths of different models.

Recommendation to Stakeholders

Based on the results, the MLP neural network model provided strong predictive performance, particularly with respect to recall, which is crucial in a medical diagnosis context. However, the logistic regression model’s interpretability makes it an attractive option for healthcare professionals who need to understand the reasoning behind the model’s predictions.

Given the trade-off between interpretability and performance, it is recommended that the stakeholders use a combination of both models. The logistic regression model can be used as an initial, interpretable predictor, while the MLP can serve as a secondary, more sophisticated model for cases where the decision is less clear.

Deployment Strategy

The model will be deployed as a web-based application using **Gradio** for the user interface. The interface will allow healthcare providers to input patient data and receive a prediction of whether the patient is at risk of heart disease. The Gradio interface was chosen for its simplicity, ease of integration, and ability to handle real-time inputs and outputs.

The model can be deployed on a cloud service such as Heroku, AWS, or Google Cloud, where it will be accessible to clinicians and integrated into the hospital's existing IT infrastructure.

Conclusion

In conclusion, this project successfully developed a predictive model for heart disease using several machine learning algorithms. Key insights gained from this project include the importance of selecting the right features and evaluation metrics to ensure the model is effective and interpretable in a clinical context. Based on the results of the MLP neural network and logistic regression, the following conclusions can be drawn:

1. **Interpretability vs. Performance**:

2. Logistic regression provided a highly interpretable model, which is crucial for healthcare applications where transparency and understanding of the model's decisions are vital. However, the neural network showed better overall predictive performance, especially in terms of recall, which is critical for reducing false negatives in heart disease diagnosis.

3. **Feature Importance**:

Features such as chest pain type, ST depression, and maximum heart rate were the most influential predictors of heart disease, aligning with medical knowledge about cardiovascular risk factors. This reinforces the model's potential to serve as a reliable diagnostic tool.

4. **Model Recommendation**:

After evaluating the models, it is recommended that the stakeholder adopt the logistic regression model for initial use, due to its interpretability. However, the MLP neural network could be integrated into the diagnostic pipeline as a second-opinion system, offering deeper insights in more complex cases.

5. **Future Work**:

Future work should focus on improving the neural network through hyperparameter tuning and cross-validation. Additional features or medical indicators could be incorporated to further improve the model's predictive accuracy. Another promising direction would be to explore ensemble methods that combine the strengths of multiple models for more robust predictions.

6. **Deployment and Impact**:

The model is deployed using Gradio to provide an interactive, user-friendly interface for healthcare providers. This ensures that the model can be used effectively in clinical settings, assisting healthcare professionals in making better-informed decisions for heart disease prevention and management.

By implementing this predictive model in the healthcare organization’s diagnostic pipeline, the stakeholders can potentially reduce the burden of cardiovascular diseases through early detection, thus improving patient outcomes and reducing healthcare costs. The model's precision and recall provide a strong foundation for trust in its predictions, making it suitable for real-world deployment.