Heart Disease Prediction Project Writeup

# Introduction

Heart disease is one of the leading causes of mortality worldwide, making its early detection crucial. The goal of this project is to predict the presence of heart disease in patients using various diagnostic measurements. Accurate prediction models can aid healthcare professionals in making timely decisions and improving patient outcomes.

## Importance and Usefulness

Predicting heart disease is essential for early intervention and treatment, potentially saving lives. This model can support healthcare providers by identifying high-risk individuals who may benefit from preventive measures or early treatment. For stakeholders, the potential benefits include improved patient care, reduced healthcare costs, and enhanced efficiency in the healthcare system.

## Stakeholder Pitch

To gain buy-in from stakeholders, the problem can be pitched as follows: 'Heart disease remains a top cause of death, but early detection can significantly improve patient outcomes. By leveraging predictive modeling, we can identify at-risk individuals earlier, enabling timely intervention and potentially saving lives. Investing in this project means contributing to a healthier society and reducing long-term healthcare costs.'

## Data Source

The dataset used in this project combines five popular heart disease datasets from various hospitals, harmonized to include 11 common features: Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol Levels, Fasting Blood Sugar, Resting Electrocardiogram Results, Maximum Heart Rate Achieved, Exercise Induced Angina, ST Depression Induced by Exercise, and Slope of the Peak Exercise ST Segment. The target variable indicates the presence (1) or absence (0) of heart disease.

# Summary of Milestones

## Milestone 1: Data Exploration and Visualization

**Introduction:**

The purpose of this analysis is to predict heart disease based on various diagnostic measurements. Heart disease is a leading cause of death globally, making it a critical area for accurate diagnosis using predictive modeling.

**Graphical Analysis:**

Age Distribution by Heart Disease Status:

* The histogram shows older age groups are more likely to have heart disease. This aligns with known medical insights that age is a significant risk factor for heart disease.

Max Heart Rate by Heart Disease Status:

* Individuals without heart disease tend to have higher maximum heart rates. This suggests that lower heart rates could be associated with increased cardiovascular risk, potentially due to poorer heart function or reduced fitness levels.

Cholesterol Levels by Heart Disease Status:

* Higher cholesterol levels are prevalent among those with heart disease. High cholesterol is a well-known risk factor for cardiovascular diseases, contributing to plaque build-up in arteries.

Resting Blood Pressure by Heart Disease Status:

* Higher median blood pressure is observed in patients with heart disease. Elevated blood pressure is a major risk factor for heart disease, indicating higher strain on the cardiovascular system.

## Milestone 2: Data Preparation and Feature Selection

**Feature Selection:**

* Correlation matrix was computed to ensure minimal multi-collinearity among features. Most features did not have very high correlations with each other, indicating minimal redundancy. No features were dropped as all showed relevance, making them potentially useful for building the logistic regression model.

**Data Standardization:**

* Continuous features (Age, Resting Blood Pressure, Cholesterol, Maximum Heart Rate, Oldpeak) were standardized to ensure each contributes equally to the model. Logistic regression can be sensitive to the scale of input features, and standardizing ensures that larger-scaled features do not dominate the training process.

**Data Splitting:**

* Data was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data. This step is crucial for preventing overfitting and ensuring that the model generalizes well to new data.

**Feature Transformation**:

* Log transformation was applied to the 'Oldpeak' feature to correct skewness. The feature ‘Oldpeak’ was highly positively skewed, meaning the data in this variable was right-tailed. Transforming this feature helped in normalizing its distribution, leading to better model performance.

**Handling Missing Data and Dummy Variables**:

* No missing data was found, and categorical variables were already appropriately encoded. If there were missing values, imputation techniques such as using the average or median values would be considered. For categorical variables, one-hot encoding was used to convert them into numerical format.

## Milestone 3: Model Building and Evaluation

**Logistic Regression Model:**

* A logistic regression model was built using the standardized dataset. Logistic regression is a simple yet effective classification algorithm, particularly suitable for binary classification problems like predicting the presence or absence of heart disease.
* Model training and testing were performed, with predictions made on the test set. The model used a regularization parameter to prevent overfitting and improve generalization.

**Model Performance:**

* Accuracy: 83.19%
* Precision: 82% for class 0, 84% for class 1
* Recall: 83% for class 0, 83% for class 1
* F1-Score: 83% for class 0, 84% for class 1

**Confusion Matrix**:

* The confusion matrix provided insights into true positive, true negative, false positive, and false negative rates. High values along the diagonal indicated correct predictions, while low off-diagonal values indicated fewer misclassifications.

**Interpretation of Confusion Matrix:**

* True Positives (TP): Patients correctly predicted to have heart disease.
* True Negatives (TN): Patients correctly predicted not to have heart disease.
* False Positives (FP): Patients incorrectly predicted to have heart disease (Type I error).
* False Negatives (FN): Patients incorrectly predicted not to have heart disease (Type II error).

**ROC Curve:**

* The ROC curve was plotted to visualize the model's performance in distinguishing between classes. The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between positive and negative classes. A high AUC indicates good model performance.

# Conclusion

## Analysis Insights

The analysis and model building reveal that the logistic regression model performs well in predicting heart disease with an accuracy of 83.19%. Key factors such as age, cholesterol levels, and maximum heart rate significantly influence the likelihood of heart disease. The graphical analysis highlighted the importance of these features, and the model's performance metrics indicate its reliability.

## Model Deployment Readiness

While the model shows good performance, further validation and testing on diverse datasets are recommended before deployment. This ensures robustness and generalizability across different populations. Additionally, the model's interpretability makes it suitable for deployment in clinical settings, where understanding the rationale behind predictions is crucial.

## Recommendations

1. Further Model Improvement:
   * Experiment with more complex models (e.g., Random Forest, Neural Networks) to potentially improve accuracy. Ensemble methods or deep learning techniques might capture more complex patterns in the data.
2. Regular Updates:
   * Continuously update the model with new data to maintain accuracy and relevance. Incorporating recent patient data ensures that the model adapts to any changes in patient demographics or disease patterns.
3. Integration:
   * Integrate the model into healthcare systems to assist clinicians in real-time decision-making. This can be achieved through user-friendly interfaces or integration with electronic health record (EHR) systems.

## Challenges and Opportunities

1. Data Quality:
   * Ensuring high-quality, representative data is crucial for model accuracy. Regular audits and updates to the dataset are necessary to maintain the model's performance.
2. Ethical Considerations:
   * Addressing privacy concerns and ensuring unbiased predictions are critical.
3. Future Opportunities:
   * Integrate the model into healthcare systems to assist clinicians in real- Expand the model to predict other cardiovascular diseases or health conditions using similar methodologies. Leveraging similar data-driven approaches can help in early detection and intervention for a range of diseases.