Problem Statement

The primary objective of this project is to develop an effective predictive model for diagnosing heart disease within a 2-week timeline. Utilizing a comprehensive heart disease dataset curated from five popular sources (Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog), recorded over various periods, the project aims to create a robust predictive model. This model will be deemed successful if the predictions closely align with the real data, achieving an accuracy rate of 85% or more. Key variables, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak (ST depression), the slope of the peak exercise ST segment, and the target class (presence or absence of heart disease), will be crucial in achieving this level of precision. The success criteria emphasize the importance of providing accurate and reliable predictions to contribute to a comprehensive understanding of factors influencing heart disease and enhancing early diagnosis and treatment.

Context

The focus on predicting heart disease is driven by the overarching goal of improving early diagnosis and treatment, thereby reducing mortality rates and improving patient outcomes. By developing an accurate predictive model, we aim to contribute to a deeper understanding of the factors influencing heart disease. This initiative aligns with the broader mission of advancing healthcare through data-driven approaches and optimizing patient care.

The significance of this project lies in its potential to enhance our ability to diagnose heart disease effectively. By accurately forecasting the presence of heart disease, we can inform strategic decision-making in clinical settings, ultimately contributing to the advancement of early diagnosis and treatment protocols. The project's outcomes have the potential to impact the healthcare sector positively, promoting more efficient and reliable heart disease diagnosis.

Criteria for Success

The success of this project will be determined by the accuracy and reliability of the predictive model in diagnosing heart disease. The key criteria for success include achieving a prediction accuracy rate of 85% or more, as measured against the real data. The model's ability to closely align its predictions with actual diagnoses within a 2-week timeline will be the primary indicator of success. This high level of accuracy is crucial for providing valuable insights into the factors influencing heart disease and, by extension, improving early diagnosis and treatment efficiency.

Scope of Solution Space

The focus of this business initiative is exclusively on developing a predictive model for diagnosing heart disease. The project's key components include the analysis and utilization of data collected from the combined heart disease dataset. Specific variables, such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak (ST depression), the slope of the peak exercise ST segment, and the target class, will be the primary

focus. The objective is to create a robust model within a 2-week timeline that accurately forecasts the presence of heart disease, with a particular emphasis on achieving a prediction accuracy rate of 85% or more. The exclusive concentration on these elements aims to contribute to a comprehensive understanding of factors influencing heart disease, driving advancements in early diagnosis and treatment.

Constraints

The accuracy and reliability of the predictive model heavily depend on the quality of the available data. Incomplete or inaccurate data could lead to biased or less reliable predictions.

Stakeholders

- Cardiologists
- Medical Researchers
- Healthcare Providers
- Data Scientists
- Patients and Patient Advocacy Groups

Data Sources

1. **Age**: Age in years.
2. **Sex**: 1 = male, 0 = female.
3. **Chest Pain Type**: Value 1 = typical angina, Value 2 = atypical angina, Value 3 = non-anginal pain, Value 4 = asymptomatic.
4. **Resting Blood Pressure**: Resting blood pressure in mm Hg.
5. **Serum Cholesterol**: Serum cholesterol in mg/dl.
6. **Fasting Blood Sugar**: 1 = fasting blood sugar > 120 mg/dl, 0 = fasting blood sugar ≤ 120 mg/dl.
7. **Resting Electrocardiogram Results**: Value 0 = normal, Value 1 = having ST-T wave abnormality, Value 2 = showing probable or definite left ventricular hypertrophy.
8. **Maximum Heart Rate Achieved**: Maximum heart rate achieved.
9. **Exercise Induced Angina**: 1 = yes, 0 = no.
10. **Oldpeak (ST Depression)**: ST depression induced by exercise relative to rest.
11. **The Slope of the Peak Exercise ST Segment**: Value 1 = upsloping, Value 2 = flat, Value 3 = downsloping.
12. **Target Class**: 1 = heart disease, 0 = normal.

Methodology

1. **Data Preprocessing**: Clean and preprocess the data to handle missing values and normalize the features.
2. **Feature Engineering**: Analyze and select the most relevant features for the predictive model.

3. **Model Development**: Develop various machine learning models including Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Gradient Boosting, XGBoost, AdaBoost, Naive Bayes, and MLP Neural Network.
4. **Model Evaluation**: Evaluate the models based on accuracy, precision, recall, and F1-score. Select the best-performing model.
5. **Model Deployment**: Deploy the final model for practical use in clinical settings.

Tools and Technologies

- **Programming Language**: Python
- **Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost, lightgbm
- **Development Environment**: Jupyter Notebook

Timeline

- **Week 1**: Data collection, preprocessing, and exploratory data analysis.
- **Week 2**: Model development, evaluation, and selection. Final report preparation and model deployment.