**Final Report: Heart Disease Analysis**

**1. Introduction**

**Problem Statement**

Heart disease is one of the leading causes of mortality worldwide. Early detection and proper analysis of risk factors can help improve treatment and prevention strategies. The goal of this project is to analyze heart disease data using data science techniques to identify key patterns and predictive insights.

**Objectives**

1. **Data Exploration & Cleaning**: Inspect and clean the dataset to ensure it is ready for analysis.
2. **Exploratory Data Analysis (EDA)**: Identify patterns and correlations in the dataset.
3. **Modeling**: Train different machine learning models to predict the likelihood of heart disease.
4. **Evaluation**: Compare model performance and determine the most effective predictive approach.
5. **Deployment**: Use **Streamlit** to build an interactive web app that visualizes insights and predictions.

**Dataset**

The dataset contains information about various health indicators such as cholesterol levels, blood pressure, age, gender, and presence of heart disease. It was preprocessed to remove missing values and outliers before modeling.

**2. Methodology**

**Step 1: Data Exploration and Cleaning**

* **Loaded the dataset** using Pandas.
* **Checked for missing values** and removed any incomplete rows.
* **Converted categorical values** into numerical format if necessary.
* **Handled outliers** using standard statistical techniques.

**Step 2: Exploratory Data Analysis (EDA)**

* **Visualized distributions** of numerical features using histograms.
* **Generated correlation matrices** to identify key relationships.
* **Boxplots and scatter plots** were used to find trends in the data.
* **Compared health attributes** between patients with and without heart disease.

**Step 3: Modeling**

* Applied **Logistic Regression**, **Random Forest**, and **Support Vector Machines (SVM)**.
* Used **train-test split (80-20)** for model evaluation.
* Optimized hyperparameters using **GridSearchCV**.
* Measured performance using **accuracy, precision, recall, and F1-score**.

**Step 4: Model Evaluation**

* **Random Forest** had the highest accuracy (~85%).
* Logistic Regression performed well but had slightly lower accuracy (~78%).
* SVM worked well for high-dimensional data but required careful tuning.

**Step 5: Deployment with Streamlit**

* Developed a web-based application where users can input patient data and receive a prediction.
* Added visualizations and explanations for model outputs.
* Deployed the Streamlit app to **GitHub** and shared via **Streamlit Cloud**.

**3. Results**

**Key Findings**

* **Cholesterol levels and blood pressure** were strongly correlated with heart disease.
* **Age and gender** also played a significant role in determining risk.
* **Random Forest performed best** due to its ability to handle complex relationships between features.

**Insights**

* Early identification of high-risk individuals can significantly improve treatment plans.
* Machine learning models can assist doctors in diagnosing heart disease with high accuracy.
* Future work can include incorporating more patient history data for better predictions.

**4. Code Submission**

**File Structure**

/heart\_disease\_analysis

│── app.py (Streamlit App)

│── data\_processing.py (Data Cleaning & EDA)

│── modeling.py (Machine Learning Models)

│── requirements.txt (Dependencies)

│── README.md (Project Documentation)

**Code Overview**

* **app.py**: Implements the Streamlit web application.
* **data\_processing.py**: Handles loading, cleaning, and exploring the dataset.
* **modeling.py**: Trains and evaluates machine learning models.
* **requirements.txt**: Lists dependencies (Pandas, Scikit-learn, Streamlit, Matplotlib, etc.).

**5. Conclusion**

This project successfully analyzed heart disease data, built predictive models, and deployed a user-friendly web application using **Streamlit**. The insights gained can help in the early detection of heart disease and provide valuable assistance to healthcare professionals.

Next steps include integrating more features, improving model interpretability, and deploying the app for public use.

**6. References**

* [Kaggle Heart Disease Dataset](https://www.kaggle.com/datasets)
* Scikit-learn Documentation
* Streamlit Documentation