**Capstone Project on Heart Disease Prediction**

**Introduction**

This document outlines the end-to-end process, solution architecture, and methodology employed for a heart disease prediction project.

**Data Overview**

The dataset used for this project consists of the following columns:

**age**

**sex**

**chest pain type (4 values)**

**resting blood pressure**

**serum cholestoral** in mg/dl

fasting blood sugar > 120 mg/dl

**resting electrocardiographic results** (values 0,1,2)

maximum heart rate achieved

**exercise:** induced angina

**oldpeak:** ST depression induced by exercise relative to rest

**slope:** the slope of the peak exercise ST segment

number of major vessels (0-3) colored by fluoroscopy

**thal:** 0 = normal; 1 = fixed defect; 2 = reversible defect

**target:** Presence of heart disease (1 for Yes, 0 for No)

**Data Preprocessing**

* Creation of Age Group Column: Loaded the data into MySQL, created a view, and added a new column named age\_group based on age ranges.
* Import Data into Jupyter Notebook: Imported the view into Jupyter Notebook using pd.read\_sql\_query.

**Data Exploration:**

* Checked information and statistical description of the data.
* Checked for missing values (none found).
* Explored the distribution of the target variable using value\_counts.
* Plotted countplots for 'sex' and 'age\_group' variables.

**Exploratory Data Analysis (EDA):**

* Created a pairplot for selected variables ('age', 'trestbps', 'chol', 'thalach', 'oldpeak').
* Checked correlation heatmap to identify relationships between variables.

**Data Encoding:**

* Converted categorical data in the 'age\_group' column using label encoding.

**Model Building**

**Feature Scaling:**

* Chose 'target' as the dependent variable and other variables as independent variables.
* Scaled down the features using Standard Scaler.

**Model Selection:**

* Created five different models: Logistic Regression, Decision Tree, Random Forest, K-nearest Neighbors (KNN), and Support Vector Machines (SVM).

**Model Evaluation:**

* Evaluated models based on training accuracy, testing accuracy, precision score, and F1-score.
* Logistic Regression performed well across all metrics.
* Hyperparameter Tuning and Model Refinement:
* Used hyperparameter tuning for Random Forest to find optimal parameters.
* Used error plot method for K-nearest Neighbors to determine the optimal number of neighbors.

**Model Comparison and Selection**

The following table summarizes the performance of each model:

| Algorithm Name | Training Accuracy | Testing Accuracy | Precision Score | F1-Score |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.847 | 0.967 | 0.974 | 0.974 |
| Decision Tree | 0.913 | 0.934 | 0.900 | 0.947 |
| Random Forest | 0.930 | 0.918 | 0.966 | 0.918 |
| K-nearest Neighbors | 0.822 | 0.951 | 0.917 | 0.957 |
| Support Vector Machine | 0.901 | 0.951 | 0.917 | 0.957 |

**Model Selection**

Considering testing accuracy, precision score, and F1-score, Logistic Regression is chosen as the ideal model for this project.

**Further Analysis**

A loop method was employed to find the optimal random\_state for maximum testing accuracy.

**Conclusion**

In conclusion, the heart disease prediction project successfully achieved its objectives, delivering a reliable model for identifying individuals at risk of heart disease. The choice of Logistic Regression as the ideal model is supported by its balanced performance across various evaluation metrics.

The comprehensive nature of the data preprocessing, exploratory data analysis, and model development processes ensures the reliability and interpretability of the final model. Furthermore, the iterative approach, including hyperparameter tuning and additional analyses, contributes to the robustness of the project.

The insights gained from this project can potentially aid healthcare professionals in early detection and intervention for individuals at risk of heart disease, ultimately improving patient outcomes.

Overall, the success of this project underscores the effectiveness of data-driven approaches in healthcare analytics and highlights the potential for further advancements in predictive modeling for cardiovascular health.