**Project Report: Heart Disease Prediction using Logistic Regression with Feature Selection**

**1. Introduction** This project aimed to predict the presence of heart disease in patients using the Logistic Regression algorithm. The dataset used was the "Heart Disease UCI" dataset from Kaggle. We evaluated the impact of feature selection on the model's performance.

**2. Data Preparation** **Dataset:** Heart Disease UCI dataset from Kaggle. **Steps:**

1. **Data Cleaning:** Handled missing values and duplicates.
2. **Feature Selection:** Applied filter method using correlation with the target variable and Recursive Feature Elimination (RFE).

**3. Correlation Matrix and Feature Selection** A correlation matrix was calculated to determine the relationship between features. A threshold of 0.25 was set to select features that had a significant correlation with the target variable.

**Selected Features:** • 'age' • 'sex' • 'chest\_pain\_type' • 'thalch' • 'exang' • 'oldpeak' • 'ca' • 'target'

**4. Model Training and Evaluation** **Splitting Data:** The data was split into training and testing sets with a ratio of 80:20, using a random state of 42 for reproducibility. **Standardization:** Features were standardized using StandardScaler. **Logistic Regression Model:** The Logistic Regression classifier was trained with a maximum of 1000 iterations.

**Evaluation Metrics:** • Accuracy • Precision • Recall • F1-score

**Performance Results**

**Logistic Regression Performance:**

* **Accuracy:** 0.7163
* **Precision:** 0.6701
* **Recall:** 0.7163
* **F1-score:** 0.6901

**Logistic Regression Performance with RFE:**

* **Accuracy:** 0.6950
* **Precision:** 0.6396
* **Recall:** 0.6950
* **F1-score:** 0.6623

**Logistic Regression Performance with Correlation-based Feature Selection:**

* **Accuracy:** 0.6667
* **Precision:** 0.6134
* **Recall:** 0.6667
* **F1-score:** 0.6332

**Conclusions**

* • **Feature Selection Impact:** Both the filter method and RFE resulted in a decrease in model performance. This suggests that some of the features removed during feature selection might have been important for the prediction task. • **Best Approach:** Although both feature selection methods reduced the performance metrics, the model without feature selection achieved the highest accuracy, precision, recall, and F1-score. • **Model Performance:** The Logistic Regression model performed best without any feature selection, indicating that the removed features contained valuable information for prediction.

### Recommendations

• **Cautious Feature Selection:** Ensure that feature selection methods are carefully evaluated to avoid removing informative features. • **Further Exploration:** Experiment with other feature selection techniques and hyperparameter tuning to optimize the model further. • **Consider Interactions:** Explore models and feature selection techniques that account for interactions between features.

**References:** • Dataset: [Heart Disease UCI](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data) • Scikit-learn library for model implementation and evaluation