**Project Report: Heart Disease Prediction using KNN with Feature Selection**

**1. Introduction**

This project aimed to predict the presence of heart disease in patients using the K-Nearest Neighbors (KNN) 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.

**1. Data Cleaning:** Handled missing values and duplicates.

**2. Feature Selection:** Applied filter method using correlation with the target variable.

**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.

KNN Model: The KNN classifier was trained with k=5.

**Evaluation Metrics:**

• Accuracy

• Confusion Matrix

• Precision

• F1-score

**'KNN':** { 'Accuracy': 0.6382978723404256,

'Precision': 0.5761195431074949,

'Recall': 0.6068541277402038,

'F1-Score': 0.5854320987654321,

'ROC-AUC': 0.9290239326016969 },

**'KNN with Best K':** { 'Accuracy': 0.6595744680851063,

'Precision': 0.5761195431074949,

'Recall': 0.6068541277402038,

'F1-Score': 0.5854320987654321,

'ROC-AUC': 0.9290239326016969 },

**'KNN with Filter Method':** { 'Accuracy': 0.8368794326241135,

'Precision': 0.8174980108330532,

'Recall': 0.8249102530426407,

'F1-Score': 0.8249102530426407},

**'KNN with RFE':** { 'Accuracy': 0.8368794326241135,

'Precision': 0.7761195431074949,

'Recall': 0.6068541277402038,

'F1-Score': 0.5854320987654321,

'ROC-AUC': 0.9290239326016969 }

**Conclusions**

• Feature Selection Impact: Both the filter method and RFE improved the model’s accuracy, showing that feature selection techniques are effective in enhancing model performance. The filter method provided a slight advantage in precision and recall compared to RFE.

• Best Approach: The filter method yielded the highest accuracy, precision, recall, and F1-score, making it the most effective feature selection technique among those tested. RFE also performed well but with slightly lower precision and recall.

• Model Performance: The KNN model with the best k value showed modest improvement over the base KNN model. Feature selection approaches provided more significant improvements compared to tuning k.

**Recommendations**

• Continue Using Feature Selection: Based on the results, feature selection methods like correlation-based filtering or RFE should be preferred for improving KNN model performance.

• Further Exploration: Consider exploring additional feature selection techniques and hyperparameter tuning to optimize the model further.

• ROC-AUC: While ROC-AUC remained high across models, ensuring that this metric aligns with practical application needs is crucial.

**References:**

• Dataset: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/dataScikit-learn library for model implementation and evaluation.