**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.

**Steps:**

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

**Performance Before Feature Selection:**

* **Accuracy:** 0.6383
* **Precision:** 0.6148
* **F1-score:** 0.6242

**Performance After Feature Selection:**

* **Accuracy:** 0.8369
* **Precision:** 0.8175
* **F1-score:** 0.8249

**Classification Report:**

| **Class** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.92 | 0.96 | 0.94 | 79 |
| 1 | 0.83 | 0.78 | 0.81 | 37 |
| 2 | 0.50 | 0.75 | 0.60 | 8 |
| 3 | 0.64 | 0.54 | 0.58 | 13 |
| 4 | 0.00 | 0.00 | 0.00 | 4 |
| **Accuracy** | **0.84** | **141** |  |  |
| **Macro avg** | 0.58 | 0.61 | 0.59 | 141 |
| **Weighted avg** | 0.82 | 0.84 | 0.82 | 141 |

**5. Conclusion**

The KNN model performed better before feature selection with an accuracy of 83.69%. Feature selection based on correlation threshold (0.25) reduced model performance to 63.83%. This indicates that some features excluded in the selection process were important for predicting heart disease.

**Recommendations:**

* Investigate other feature selection methods.
* Try different values of k in the KNN algorithm.
* Explore other machine learning models.

**References:**

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