Name: Pola Gnana Shekar

Roll No: 21CS10052

Report of Assignment 1

Strengths and weaknesses of Logistic Regression, Support Vector Machines (SVM), Decision Tree, and K-Nearest Neighbors (KNN):

Logistic Regression:

Strengths:

- Logistic Regression is a simple and interpretable algorithm.
- It works well when the relationship between features and the target is approximately linear.
- It provides probabilities, making it useful for ranking predictions.

Weaknesses:

- May not capture complex non-linear relationships in the data effectively.
- Sensitive to outliers.
- May not perform well when there is multicollinearity among features.

Support Vector Machines (SVM):

Strengths:

- SVM can handle non-linear relationships through kernel functions.
- Effective in high-dimensional spaces.
- Works well when there is a clear margin of separation between classes.

Weaknesses:

- SVM can be computationally expensive, especially with large datasets.

- Selection of the appropriate kernel and hyperparameters can be challenging.
- SVM may not perform well when there is a high degree of class overlap.

Decision Tree:

Strengths:

- Decision Trees are interpretable and can be visualized.
- They can handle both numerical and categorical data.
- Require little data preprocessing (e.g., no need for feature scaling).

Weaknesses:

- Prone to overfitting, especially when the tree is deep.
- Decision Trees can be sensitive to small changes in the data.
- Not suitable for capturing complex relationships in data without deep trees.

K-Nearest Neighbors (KNN):

Strengths:

- KNN is simple to understand and implement.
- It is a non-parametric algorithm, making it suitable for non-linear relationships.
- Effective when data has localized patterns.

Weaknesses:

- Computationally expensive for large datasets, as it requires calculating distances for each prediction.
- Sensitive to the choice of the distance metric and the number of neighbors (k).
- Not suitable for high-dimensional data due to the "curse of dimensionality."

Identifying the Most Suitable Model:

The most suitable model for predicting heart disease depends on the specific criteria you prioritize, such as accuracy and efficiency:

- ❖ If **accuracy** is the primary concern, **SVM** (with the "SMOTE" resampling strategy) achieved the highest accuracy of 0.8467, followed by SVM (with the "RandomUnderSampler" resampling strategy) with an accuracy of 0.8514.
- ❖ If **efficiency** (considering both computational resources and model simplicity) is important**, Logistic Regression** might be a better choice. It achieved a reasonable accuracy of 0.6604 (with "SMOTE" resampling) and 0.6450 (with "ADASYN" resampling) while being computationally less demanding compared to SVM.

It's essential to consider the trade-offs between accuracy and efficiency when choosing the most suitable model for a real-world application.