In this assignment, I explored the heart disease dataset to predict whether individuals have heart disease or not based on various medical features using Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) models. The dataset, which I found from the UCI Machine Learning Repository, includes details on patient demographics and health metrics, the dataset consists of 303 samples with 13 features, such as age, sex, blood pressure, and cholesterol levels.

I began the analysis with data preprocessing, loading the dataset, first replacing the missing values with NaN and removed those rows. I also converted categorical features into numerical values and standardized the features using StandardScaler from sklearn. I first trained the SVM model using its default parameters then explored the C hyperparameter by testing a range of values, to see how this would impact accuracy. Surprisingly the best was the default value of 1. For KNN, I implemented the algorithm from scratch and trained it on the same training set, making predictions on the development which produced better results than SVM. For a baseline for my results, I tested two Dummy Classifiers, one that used the stratified strategy and another that used the most frequent strategy.

My results were interesting, the initial accuracy for the SVM model with default settings was about 55.56%. Even after testing various C values, the best accuracy I could achieve was still 55.56% with C = 1. Most other values led to slightly lower accuracies, which made me think that the SVM model wasn’t super responsive to changes in this hyperparameter for this dataset. On the other hand, the KNN model performed better, achieving an accuracy of 60% on the development set. This result suggested that KNN could be a better fit for this dataset and capture the relationships between the features and the target variable more effectively than SVM. The Dummy Classifier with the stratified strategy and got an accuracy of only 28.89% and the most frequent strategy yielded 53.33%, confirming that both of my machine learning models SVM and KNN were significantly outperforming these dummy approaches.

In summary, my analysis showed that the KNN classifier had the highest accuracy at 60.00%, while the SVM model peaked at 55.56%. Both models performed much better than the baselines set by the Dummy Classifiers. This experiment highlighted KNN as a solid option for predicting heart disease in this dataset, while the SVM model struggled somewhat regardless of the chosen C values for the hyperparameter.