Data Analysis and Results

At first, we imported the following library Pandas to read our dataset, following our dataset being uploaded to the environment we created two variables ‘X’ and ‘Y ’to assign all the features of the dataset to ‘X’ except for the column “Target’ which was assigned to Y.

After successfully reading and writing variables we move on to select a suitable algorithm, in this case, we selected RandomForest Classifier.
We imported RandomForestClassifier from the Scikit-learn package and instantiated it to ‘Clf’.

We then split the dataset into training and testing, for our model we selected 75 percent of the dataset to be used for our training purpose and the remaining 25 percent for testing.
Using Scikit-learn ‘Model Selection’ method we were able to train, test, and fit our dataset.

Upon training, we evaluated our model’s score which was 98.83% Accuracy. Our model was able to predict if the patient has heart disease with high accuracy.

Furthermore, to determine if tuning hyperparameters yielded a better model, we created a loop to change one of its parameters called ‘N_estimators’ and increased it by 10 until 40, and recorded the results.
However, we found in our case that the default hyperparameters for RandomForest Classifier yielded the same accuracy.

Table 1. Hyperparameter: N_Estimator Results

Accuracy	N_Estimator
98.83%	10
98.83%	20
98.83%	30
98.83%	40

Table 2. Confusion Matrix Score

Precision	Recall	F1-Score	Support
0.98	1.00	0.99	140

1) **Precision:**
It is defined by as the ratio of 'True Positive' to the sum of 'True Positive' and 'False positive'. In other words, it is the accuracy of the positive prediction. The mathematical formula is TP / (TP + FP).

2) **Recall:**
This is defined as the ratio of 'True Positives' to the sum of 'True Positive' and 'False Negative', it is the fraction of positives that were correctly defined. The mathematical formula is TP / (TP + FN)

3) **F1-Score:**
It is the value of weighted mean of 'Precision' and 'Recall'. This score would address the question of 'What percent of positive predictions were right?
The mathematical formula is 2*(Recall*Precision) / (Recall + Precision)

In medical diagnosis a high recall is extremely important. A greater number of false negatives would signal a patient is classified as a normal, but in reality, it is a patient with heart disease. Therefore, a heart disease diagnosis machine learning model should aim exceedingly high recall percentage. We can notice that our model achieved a recall score of 1.

Conclusion:
With the help of the RandomForest Classifier algorithm, we were able to build a machine-learning model.
Our model was trained and tested by a dataset from the UCI repository.

The dataset consisted of labelled 1025 patients, it included both diagnosed heart disease patients and normal patients.

After the model was trained and then tested, we achieved an accuracy of 98.83% with the default hyperparameter. While we tried to tune RandomForest

Classifier's hyperparameter; N_estimator in the hope of higher accuracy, we noticed that it resulted the same accuracy as default.

We can conclude that machine learning and data mining can play an important role in our healthcare system. Traditionally, diagnosis of the disease was performed by standard procedures and doctor's intuitions which had limitations and led to costly expenses, but with machine learning models, diagnosis can be done on large datasets cost-effectively.