ADS Assignment 9.

1. Load in the data. The target column should be considered as whether a patient will develop heart disease or not.
2. Explore the data. Notice all columns are numerical. Therefore separate the continuous from the discrete features.
3. Identify any presence of outliers in the continuous features and resolve them using the IQR method.
4. Binned the continuous column values apart from the column ‘oldpeak’.
5. Separate the features from the labels and use the most appropriate feature selection technique(s).
6. Slice the data and scale the features.
7. Identify the data if the data is balanced. If not, sample the data using the most appropriate method keeping the size of the data in mind.
8. Using at least 4 classification methods, identify the best machine learning model using their training and testing accuracy scores.
9. Hyper parameter tune the best model using grid search to identify the best performing model.
10. Redefine the model instance based on the grid search results, train it and evaluate it using:
    a. A classification report.
    b. A visual representation and well labelled confusion matrix.
    c. AUC score. (Explain the score in a markdown cell.)
    d. ROC curve.
11. Based on the results on the ROC curve, which threshold would be ideal given the nature of the data? (Explain in a markdown cell.)
12. Save the model as ‘classification_model’.