|  |  |
| --- | --- |
|  | **Instructions** |
| • | This is an **individual** assignment. |
| • | Each student will be required to submit a fully commented jupyter notebook for the Python code required for this assignment. |
| • | Students will have to attend a short Question and Answers session regarding their work. |
| • | The jupyter notebook needs to be submitted in **ipynb** format |
| • | All notebooks will be checked for plagiarism using Turnitin and late submissions will **not** be accepted. |
|  |  |

# Problem Statement

In this assignment, you been provided with a dataset containing information about the diagnoses of heart disease patients. A brief description of every feature present in the dataset is provided in Annex 1. Machine learning is needed in order to determine whether a person has a heart disease or not and in turn assist with their medical treatment. You have tasked by a senior data scientist to **explore and prepare** the dataset and to provide inferences on the causality of heart conditions based on the available information. In addition to data exploration, you will be required to make a presentation to your company’s management with regards to the data provided. Therefore, you will also need to perform an in-depth data representation and visualisation activity.

Using the knowledge you have acquired in the field of Machine Learning, prepare a comprehensive and well documented jupyter notebook for the above tasks.

***(25 marks)***

# Marking

This assignment contributes **25%** towards the final grade for this module. The work submitted will be assessed on a total of 25 marks as per the below breakdown:

|  |  |
| --- | --- |
| Notebook structure and comments | 4 Marks |
| Exploratory Data  Analysis and Data Preparation | 7 Marks |
| Data Visualisation | 7 Marks |
| Dataset Inferences | 4 Marks |
| Conclusions | 3 Marks |

## Annex 1: Feature Descriptions

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Sample Data** |
| **Age** | Patient Age (in years) | 63; 37; ... |
| **Sex** | Gender of patient (0 = male; 1 = female) | 1; 0; ... |
| **cp** | Chest pain type (4 values: 0, 1, 2, 3) | 3; 1; 2; ... |
| **trestbps** | resting blood pressure (in mm Hg) | 145; 130; ... |
| **chol** | Serum cholestoral (in mg/dl) | 233; 250; ... |
| **fbs** | Fasting blood sugar > 120 mg/dl  (1 = true; 0 = false) | 1; 0; ... |
| **restecg** | Resting  electrocardiographic results (values 0, 1, 2) | 0; 1; ... |
| **thalach** | Maximum heart rate achieved | 150; 187; ... |
| **exang** | Exercise induced angina (1 = yes; 0 = no) | 1; 0; ... |
| **oldpeak** | ST depression induced by exercise relative to rest | 2.3; 3.5; ... |
| **slope** | The slope of the peak exercise ST segment  (values 0, 1, 2) | 0; 2; ... |
| **ca** | number of major vessels  (0-4) colored by flourosopy | 0; 3; ... |
| **thal** | (3 = normal; 6 = fixed defect; 7 = reversable  defect) | 1; 3; ... |
| **Target** | Target column (1 = Yes; 0 = No) | 1; 0; ... |