**SIT720 Machine Learning**
**Assessment Task 1: Problem solving task.**

This document supplies detailed information on Assessment Task 1 for this unit.

**Key information**

- Due: **Monday 26 July 2021** by 8.00 pm (AEST)
- Weighting: 5%

**Learning Outcomes**

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

| Unit Learning Outcome (ULO) | Graduate Learning Outcome (GLO) |
|---|---|
| **ULO1 -** Perform Python programming to solve a given problem. | **GLO1 -** through the assessment of student ability to use data acquisition.<br>techniques to obtain, manipulate and represent data.<br>**GLO2 -** through the assessment of communicating the results in specific format.<br>**GLO3 -** through student ability to use specific programming language and modules to obtain, pre-process, transform and analyse data. |

**Purpose**

This assessment task is for student to apply Python programming skills for loading, visualising, manipulating and exporting data using various modules and packages.

**Assessment 1** **Total marks = 30** **15 * 2 = 30**

**Submission Instructions**

a) Submit your solution into **a notebook file with ".ipynb**" extension.
b) Insert your Python code or text responses into the cell of your submitted file followed by the question i.e., copy the question by adding a cell before the solution cell. If you need multiple cells for better presentation of the code or answer, add question only before the first solution cell.
c) Your submitted code should be executable. If your code does not generate the submitted solution, then you will **get zero** for that part of the marks.
d) For answers regarding discussion or explanation, **maximum five sentences are suggested**.
e) Answers must be **relevant and precise**.
f) No **hard coding** is allowed. Avoid using specific value that can be calculated from the dataprovided.
g) Submit your assignment after running each cell individually with the output.
h) The submitted notebook file name should be of this form "**SIT720_A1_studentID.ipynb**". For example, if your student ID is 1234, then the submitted file name should be SIT720_A1_1234.ipynb.

**Background**

According to World Health Organization (WHO), cardiovascular diseases (CVDs) are the number1 cause of death globally, taking an estimated 17.9 million lives each year and affecting the quality of life of a large number of people worldwide. Prerequisites of the treatment of these types of disease involve proper diagnosis method to identify its occurrence and its type.

Diagnosis of such diseases always involve a large number of parameters to help the Cardiologists to identify them. Electrocardiography (ECG) is most commonly used to observepatient's cardiac states. The following image shows different ECG waves (P, Q, R, S, T, etc).

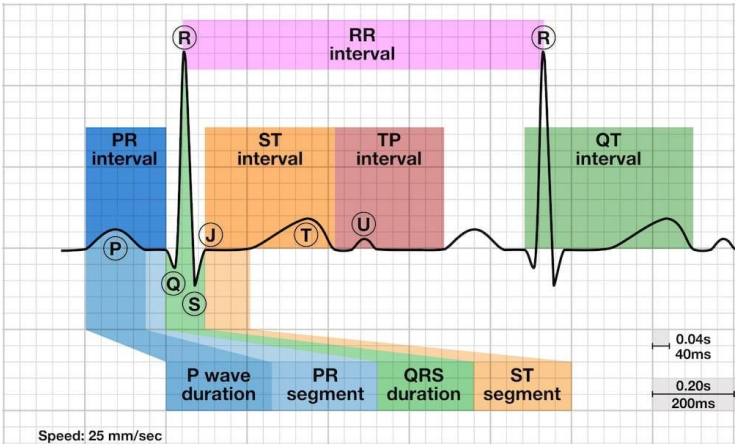


Fig: Electrocardiogram trace with respective biomarkers. [Image source: https://litfl.com/wp-content/uploads/2018/10/ECG-waves-segments-and-intervals-LITFL-ECG-library-3.jpg]

In this assignment, you will have a look at such a dataset containing different parameters along with the decision of the Cardiologists about the level of a sample heart disease. There will be a list of tasks to check your ability to use of programming skill, basic logics, and reasoning.

## Dataset

*Dataset file name*: A1_heart_disease_dataset.csv

*Dataset description*: Dataset contains different features along with the disease state. It contains total 13 features and an additional disease state, in total 14 columns. It contains different types of data including int, float and string. Feature names, data type and values are described in the following section with their proper unit details. Data may contain 'null' or 'nan' values. Each observation is a datapoint along the row of the dataset. Patient and observation areused interchangeably in this case.

*Features:*

i. age (int): age of the patient in year
ii. sex (str): gender of the patient (M: male, F: female)
iii. cp (str): chest pain type (tap: typical angina, aap: atypical angina, nap: non-anginal pain,asp: asymptomatic pain)
iv. trestbps (float): resting blood pressure (in mm Hg on admission to the hospital)
v. chol (float): serum cholesterol in mg/dl
vi. fbs (bool): is fasting blood sugar higher than standard 120 mg/dl? (yes: if true; no: if false)
vii. restecg (int): resting electrocardiographic results (0: normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), 2:showing probable or definite left ventricular hypertrophy by Estes' criteria)
viii. thalach (float): maximum heart rate achieved
ix. exang (bool): exercise induced angina (true: if yes; false: if no)
x. oldpeak (float): ST interval depression induced by exercise relative to rest

xi. slope (float): the slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3:down-sloping)

xii. ca (int): number of major vessels (0-3) affected

xiii. thal (int): thalassemia state (3: normal; 6: fixed defect; 7: reversable defect)

xiv. state (int): heart disease risk state (0: no disease, 1-4: level of risk) [Decided bycardiologists]

---

Questions

---

1. Load the data from supplied data file. Remove the observations/samples where the heart diseases are not diagnosed by the Cardiologists. Print the data dimension before and after removing the observations/samples.

2. Continue from question 1. Display the number of rows and their indices that have missing data in one or more cells. Now, replace the missing data by the lowest value of the corresponding feature if it is a continuous variable. In case of categorical variable, remove the sample. Print the median values of all features before and after replacing missing data.

3. Continue from question 2. Is there any change in data type? If yes, convert them back to appropriate data types. Print all variables with corresponding data type.

4. Continue from question 3. Print the total numbers and ration of male and female patients who are at highest risk of heart disease.

5. Continue from question 3. Is there any association between heart rate and severity of heart disease? Explain your results from given dataset.

6. Continue from question 3. Print the average cholesterol level **for different number of blocked blood vessels** across gender. Please report the pattern found in the result, if any.

7. Print the percentage of patients at risk of heart disease having abnormality in both ECG and blood sugar with asymptomatic chest pain.

8. Calculate and print the average blood pressure of all observations with non-flat ST slopes of ECG.

9. Create and print a dataframe of the heart rate, blood pressure and cholesterol levels for different age groups (based on 10 years interval).

10. Continue from question 3. Find the average cholesterol level of across gender for each age group. Please explain the results.

11. Continue from question 3. Draw two scatter plots of cholesterol level, one against blood pressure and another against heart rate. Draw them in two subplots of the same plot.

12. Visualize the cholesterol level against number of blood vessel blocked for male and female using line plot. Explain the graph base on your observation.

13. Draw a group bar diagram of heart rate, blood pressure and total number of patients, based on age groups defined in question 9. Explain your observation from the graph.

14. Continue from question 9. Add two more columns named ['num_male_patients', 'num_female_patients'] and having values of the number of male and female patients affected by heart disease in each age group respectively. Save the combined dataset to a csv file named 'age_group_stat.csv' in the same directory of your code file.

15. Continue from question 1. Replace all the rows where the 'state' is null with its immediate previous row. Finally, display and save the resultant dataset to a csv file named 'clean_data.csv' in the same directory of your code file.

**Submission details**

Deakin University has a strict standard on plagiarism as a part of Academic Integrity. To avoid any issues with plagiarism, students are strongly encouraged to run the similarity check with the Turnitin system, which is available through Unistart. A Similarity score MUST NOT exceed 39% in any case. Late submission penalty is 5% per each 24 hours from**- Monday 26 July 2021 by 8.00 pm (AEST)**, No marking on any submission after 5 days (24 hours X 5 days from- Monday 26 July 2021 by 8.00 pm (AEST),).

**Extension requests**

Requests for extensions should be made to Unit/Campus Chairs well in advance of the assessment due date. If you wish to seek an extension for an assignment, you will need to submit a request using the "Extension Request" link of the "Assessment" menu in the unit site, as soon as you become aware that you will have difficulty in meeting the scheduled deadline, but at least 3 days before the due date. When you make your request, you must include appropriate documentation (medical certificate, death notice) and a copy of your draft assignment. Conditions under which an extension will normally be approved include:

*Medical* To cover medical conditions of a serious nature, e.g. hospitalisation, serious injury or chronic illness. Note: Temporary minor ailments such as headaches, colds and minor gastric upsets are not serious medical conditions and are unlikely to be accepted. However, serious cases of these may be considered.

*Compassionate* e.g. death of close family member, significant family and relationship problems.

*Hardship/Trauma* e.g. sudden loss or gain of employment, severe disruption to domestic arrangements, victim of crime. Note: Misreading the timetable, exam anxiety or returning home will not be accepted as grounds for consideration.

*Special consideration*

You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time. See the following link for advice on the application process: http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration.

**Assessment feedback**

The results with comments will be released within 15 business days from the due date.

**Referencing**

You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

**Academic integrity, plagiarism and collusion**

Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.
Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.
Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.
Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity.