TAYLOR'S UNIVERSITY

Wisdom · Integrity · Excellence

MASTER OF APPLIED COMPUTING

FINAL ASSESSMENT

AUGUST 2022 SEMESTER (BLOCK 2)

MODULE NAME	**:**	**PRINCIPLES OF DATA SCIENCE**
MODULE CODE	**:**	**ITS70404**
EXAM DURATION	**:**	**24 HOURS**
EXAM DATE / TIME	**:**	**13/12/22 08:00 – 14/12/22 08:00 (M'SIA TIME GMT+8)**

This paper consists of SEVEN (7) printed pages, inclusive of this page.

Instruction to Candidates:

1. Answer ALL questions.
2. This is an open book assessment. Student is not allowed to transcribe directly (copy and paste) any material from another source into their submission. Students are also not allowed to consult with any party about any part of the questions throughout the assessment period.
3. The Turnitin similarity for this module is 20% and 1% from a single source excluding program source code.
4. Severe disciplinary action will be taken against those caught violating assessment rules.
5. The final assessment answers handed in should be within 5 - 12 pages in total for non-programming modules, with a spacing of 1.5 and a font of 12pt Times New Roman.
6. Name your answer file as ITS70404_XXXXXX_FinalAssessment.pdf where XXXXXX is your Student ID. Then, submit to MYTiMES portal via the link "Final Assessment submission" in module page. (Do not submit the question paper)

Section	Marks
Answer all **FOUR (4)** questions. All questions carry equal marks.	100 Marks

Assessment Criteria

Assessment Task	Weightage	MLO Assessed	Formative / Summative	Assessment Instrument	Topics	Week	MCQ2.0
Question 1,2	50%	MLO 1	Summative	Individual Assignment	1-7	8	C1, C3A C3D
Question 3,4	50%	MLO 2					

C1 = Knowledge & Understanding, **C2** = Cognitive Skills, **C3A** = Practical Skills, C3B = Interpersonal Skills, **C3C** = Communication Skills, **C3D** = Digital Skills, **C3E =** Numeracy Skills**, C3F** = Leadership, Autonomy & Responsibility, **C4A** = Personal Skills, **C4B** = Entrepreneurial Skills, **C5** = Ethics & Professionalism

QUESTIONS (100 marks)

QUESTION 1 (25 marks)

Outliers Detection and Removal

An observation in a dataset is considered an outlier if it differs significantly from the rest of the data points. Detection and removal of outliers or anomalies in a dataset is a fundamental task in data cleaning without which the analysis of the data can be misleading. Furthermore, the existence of anomalies in the data can heavily degrade the performance of machine learning algorithms. Suppose that you are given a dataset for house price prediction in the city of Bangalore, India which contains many outliers that degrade the performance of your machine learning model. Your task is to investigate *price_per_sqft* column and do following.

a. Create a Python program to load the *'banglore_house_price.csv'* dataset into a Pandas dataframe and plot a histogram to visualize the *price_per_sqft* column.

(5 marks)

b. Create a Python program to detect and remove outliers that fall within 4 standard deviations from the mean.

(5 marks)

c. Create a Python program to plot the histogram for the newly generated dataframe from Q1b. Furthermore, plot bell curve on the same histogram.

(5 marks)

d. Create a Python program to detect and remove outliers from the original dataframe created in Q1a using Z score of 4.

(5 marks)

e. Discuss and explain two other methods that can be used to detect and remove outliers from the dataset.

(5 marks)

QUESTION 2 (25 Marks)

a. Assume that your data science manager asks you to create a model for house rents in Kuala Lumpur. What model would be helpful in predicting rents and explain your answer? Give examples of three features you think might be useful.

(5 marks)

b. Suppose that you are working on a multiclass classification problem using random forest classifier with 10000 trees. Your training error is 0.00, but the testing error is 27.12. What is the problem of your model and give some suggestions to solve it.

(5 marks)

c. Write a Python program to create a Pandas Data Frame from the following dictionary data.

(5 marks)

Sample Dictionary Data:

```
exam_info = {

'Student_name': ['Adam', 'David', 'Kathy', 'Jack', 'Emily', 'Michael', 'Max', 'Laura', 'Ali', 'Sarah'],
'score': [70.5, 80, 60.4, np.nan, 90, 20, 44.5, np.nan, 80, 89],
'attempts': [2, 1, 2, 3, 2, 3, 2, 3, 1, 1],
'qualify': ['yes', 'yes', no, 'no', 'yes', 'no', 'no', 'no', 'yes', 'yes']
}
```

d. Write a Python program to convert the resulted Pandas Data Frame from Q2c into a NumPy array. Create an array that extracts only the feature data we want to work with (Student_name, score, and attempts) and another array that contains the classes (qualify).

(5 marks)

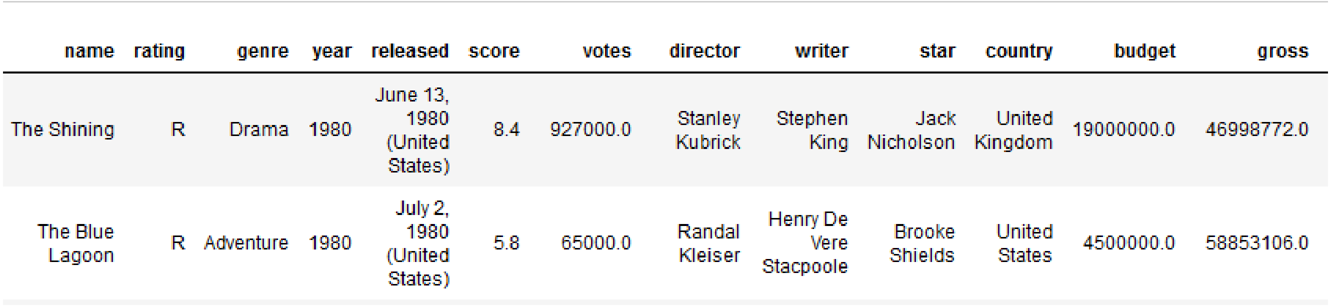
e. Write a Python program to handle the missing values in the score column using at least two techniques.

(5 marks)

QUESTION 3 (25 Marks)

Basic Statistics and Data visualization using Python

A newly launched movie production company has hired you as a data scientist to help the company find the best way to make profits from the choices they make in Movie Industry. You are required to perform a basic exploratory data analysis to explore the most important factors which are directly connected to the gross revenue. You think that there is a high correlation between the budget of the films and the gross revenue.



name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0
The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0

Check mytimes platform for movies.csv*

Based on the above scenario use **Python Pandas, seaborn and matplotlib** libraries to answer the following questions.

a. Create a scatter plot diagram to visualize the relationship between the budget and the gross revenue. Label the axes and add title.

(3 marks)

b. To better visualize the relationship between the budget and gross revenue, plot a Scatter plot with regression line**.**

(5 marks)

c. Write a python program to create a jointplot to compare the budget and gross columns. Does the correlation make sense?

(2 marks)

d. Write a python program to compute the Pearson's correlation between the budget and gross revenue. Explain the resulted correlation value.

(5 marks)

e. Write a python program to compute the Kendall correlation between all columns in the dataset. Based on the resulting Correlation Matrix is there any other factor that has a high correlation with the gross revenue? Justify your answer.

(5 marks)

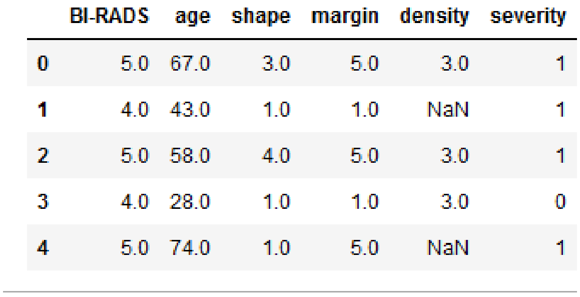
f. Write a python program to generate a correlation heatmap graph. What conclusions can you draw

(5 marks)

QUESTION 4 (25 Marks)

Case Study: Breast cancer prediction using ML and deep learning

Breast cancer is one of the leading causes of death in female cancer patients. The disease can be detected early using Mammography, an effective X-ray imaging technology. Suppose that you are a data scientist who has been recruited to help doctors to predict if a mass detected on mammogram is benign or malignant.



	BI-RADS	age	shape	margin	density	severity
0	5.0	67.0	3.0	5.0	3.0	1
1	4.0	43.0	1.0	1.0	NaN	1
2	5.0	58.0	4.0	5.0	3.0	1
3	4.0	28.0	1.0	1.0	3.0	0
4	5.0	74.0	1.0	5.0	NaN	1

Check mytimes platform for masses.txt*

Type this code in your Jupyter Notebook to load the dataset.

```
import pandas as pd
data = pd.read_csv('masses.txt')
data = pd.read_csv('masses.txt', na_values=['?'], names = ['B', 'age','shape',
'margin', 'density', 'severity'])
```

a. Write a python program to prepare your data for modeling. Check for missing values and handle it. Create two NumPy arrays for the features and the classes. Perform data normalization if required.

(5 marks)

b. Write a python program to classify tumor into benign or malignant (Severity: benign=0 or malignant=1). Use Decision Trees classifier and K-Fold cross validation to measure the accuracy.

(5 marks)

c. Write a python program to classify tumor into benign or malignant using KNN classifier and K-Fold cross validation to measure the accuracy. (Hint: to get better classification results use a for loop to find the optimal value of K in KNN).

(5 marks)

d. Use Neural networks with one binary output neuron to predict whether the mass detected on mammogram is benign or malignant. (Hint: Use Keras library and experiment with different topologies, optimizers, and hyperparameters). Which algorithm achieved the highest accuracy? Justify your answer.

(10 marks)

END OF QUESTION SECTION

Marking Schemes & Rubrics

Question 1a, 1b, 1c, 1d:				
(5 Marks)	Completely misinterprets the problem Substantially inappropriate solution. No justifications **(0-1 Mark)**	Misinterprets major part of the problem Partially correct solution but with major errors. Tried to justify the results **(2 Marks)**	Misinterprets minor part of the problem Substantially correct solution with minor errors Good justifications of the results **(3-4 Marks)**	Complete understanding of the problem. Correct solution with no algorithmic errors. Excellent justifications of the results **(5 Marks)**

Question 1e:				
(5 Marks)	No evidence and minimum explanation for the possible methods to detect outliers. **(0-1 Marks)**	Made attempt to provide the justification for the problem and provided one method to detect outliers. **(2 Marks)**	Moderately explained and justified the possible problem and provided two methods to detect outliers. **(3-4 Marks)**	Good explanation and justification for the possible problem and provided more than 2 methods to detect outliers. **(5 Marks)**

Question 2a:				
(5 Marks)	No model and made minimum attempt to explain the features. **(1 Mark)**	Made attempt to provide the justification for the possible model and features. **(2-3 Marks)**	Moderately explained and justified the possible model and features **(4 Marks)**	Good, explanation and justification for the possible model and features **(5 Marks)**

Question 2b:				
(5 Marks)	No evidence and minimum explanation for the possible problem solutions. **(0-1 Marks)**	Made attempt to provide the justification for the problem and provided solution. **(2 Marks)**	Moderately explained and justified the possible problem and provided solutions **(3-4 Marks)**	Good, explanation and justification for the possible problem and provided excellent solution. **(5 Marks)**

Question 2c, 2d, 2e:
Python Code: 3 marks Output: 2 Marks

Question 3a:
Python Code: 2 Marks Output and explanation: 1 Marks

Question 3b:
Python Code: 3 marks Output and explanation: 2 Marks

Question 3c:
Python Code: 1 mark Output and explanation: 1 Mark

Question 3d:
Python Code: 3 marks Output and explanation: 2 Marks

Question 3e:
Python Code: 3 marks Output and explanation: 2 Marks

Question 3f:
Python Code: 3 marks Output and explanation: 2 Marks

Question 4a:
Python Code: 4 marks Output: 1 Mark

Question 4b:
Python Code: 4 marks Output: 1 Mark

Question 4c:
Python Code: 4 marks Output: 1 Mark

Question 4d:
Python Code: 6marks Output: 2 Marks Explanation: 2 Marks

END OF MARKING SCHEME SECTION