BSc Degree in Information Technology

End Semester Examination 2023 (III)

IT 256 Data Science and Analytics

Time: Two Hours

Date: 19th December 2023 **Time: 9.00 a.m. – 11.00 a.m.**

Answer **Question 01(Compulsory) AND Any Two (2)** Questions.
(This paper contains 4 questions.)

1.

1.1. Briefly explain the two branches of statistical data analysis with examples where necessary.

1.2. What are the six steps of data analysis? Elaborate with examples.

1.3. Identify the typical challenges encountered by data analysts during the course of their analytical work.

1.4. Write short notes on the following

a) Data visualization techniques

b) Big Data

c) Deep Learning

1.5. A bottle of sherry nominally contains 1000 milliliters. After the introduction of new method of filling the bottles, there is a suspicion that mean volume of the bottle has changed.

In order to investigate this suspicion, a random sample of 12 bottles of sherry is taken and the volume of sherry in each bottles is measured.

The volumes, in milliliters, of sherry in these bottles are found to be,

996	1006	1009	999	1007	1003
998	1010	997	996	1008	1007

Assuming that the volume of sherry in a bottle is normally distributed,

a) Compute mean, median and mode for the above data set.

b) State the null and alternative hypothesis for the above claim.

c) Calculate the T and p values for the above data set.

d) What is your inference on the above data set at 0.05 significance level based on the results obtained in part (c)?

e) If the T and p value of the above test obtained via a program is 1.91 and 0.083 respectively, what is your inference on the hypothesis made on part b at a 0.05 significance level? Justify your answer.

f) Write python codes to compute the measures you obtain in part(a).

(30 Marks)

2.

2.1. What is Exploratory Data Analysis (EDA)? Briefly explain with an example emphasizing on the importance of EDA.

2.2. Briefly explain the steps to check the awareness level on the impact of social media (Facebook, Twitter, WhatsApp ..etc.) to the daily routine among users in Sri Lanka. You need to include the following on your process

a) Sampling method

b) Age group

c) Data collection technique

d) Suitable statistical measures

e) Conclusion based on sample test results (You may give some hypothetical values on the measures you take for part d)

2.3. ABC company manager want to test whether a new training program has a significant impact on the performance of employees. The manager collected performance scores from a sample of 12 employees before and after the training. Following is the dataset representing their performance scores (arbitrary units):

Employee	1	2	3	4	5	6	7	8	9	10	11	12
Before Training	45	50	55	48	52	47	50	53	49	46	51	48
After Training	55	58	60	57	59	56	58	61	57	54	59	56

a) Compute mean, and standard deviation for the above data set.

b) Write null and alternative hypothesis to test the claim of the researcher.

c) Compute T and P value for the sample data set and check the validity of your hypothesis at 95% confidence level.

(15 Marks)

3. Python is a powerful language used for data science. It is simple but has many powerful libraries for data science.

3.1. Name two most commonly used built-in libraries available in python for data analysis.

3.2. Examine the following quiz marks obtained by 20 students in a class (Marks are given out of 30).

10,15,15,12,29,30,30,15,18,17,30,20,20,16,21,23,24,21,21,20

a) Construct an ungrouped frequency and a grouped frequency distribution for the above data. Use a class interval of 5 for marks for the grouped distribution.

b) Find the mean, standard error, and a 95% confidence interval for the average quiz marks for students for the distribution you created on part a.

3.3. Write python codes to get the following results for the dataset in question 3.2(a)

a) Create a data frame with suitable variables

b) Mean, Median and standard deviation

c) Standard error

d) Quartile I, II and III (Q1,Q2, and Q3)

e) Scatterplot for Marks Vs Number of students

(15 Marks)

4.

4.1. Explain the nature of a probabilistic experiments with suitable examples and sample preparation in brief.

4.2. In a group of 125 students, 70 passed in mathematics, 55 passed in statistics and 30 passed in both. Calculate the probability that a student selected at random has passed,

a) At least in one subject

b) In only one subject

4.3. Amal plays 12 game of chess with computer and he wins 6 games while computer wins 4 games and 2 games end in a tie. Amal again decides to play 3 games more. Find the probability that

a) Amal wins all three games.

b) Two games end in a tie.

4.4. What are the different types of Machine Learning? Briefly explain with examples for each type.

4.5. Design an experiment to check the efficiency of a medicine for diabetes using a suitable machine learning technique. Explain the steps briefly with examples.

(15 Marks)