Strathmore UNIVERSITY

Faculty of Information Technology

Master of Science in Information Technology

End of Semester Examination

MCS 8116/ MSSET 88502: Data Science Concepts

Date: 14th August 2023 **Time** 18:00-20:30 Hours

Instructions: Answer Question **ONE** and any other **TWO** Questions

Question ONE (20 Marks)-Compulsory

a) A data scientist is expected to have an array of skill sets. One of the skills set is business skills. Explain how relevant this category of skill sets would be to a data scientist working in a company with international outlook. [2 marks]

b) The algorithms for Machine Learning can be split into categories, Supervised Learning– Unsupervised Learning. In supervised learning, the following characteristics; generalization and over-fitting are critical. Explain their role in machine learning. [4 marks]

c) An investigative team has previously used people's appearance to identify them as good or bad. They hope to build a machine-learning model to help them accurately identify and classify them. To do this, they have divided their data set into training and test data set as shown in the table below.

Name	**Attributes**						**Class**
	Sex	**Mask**	**Cape**	**Tie**	**Ears**	**Smokes**	
Training Data set							
X_1	Male	Yes	Yes	No	Yes	No	Good
X_2	Male	Yes	Yes	No	No	No	Good
X_3	Male	No	No	Yes	No	No	Good
X_4	Male	No	No	Yes	No	Yes	Bad
X_5	Female	Yes	No	No	Yes	No	Bad
X_6	Male	No	No	No	No	No	Bad
Test data							
X_7	Female	Yes	Yes	No	Yes	No	??
X_8	Male	Yes	No	No	No	No	??

Illustrate how to use decision tree induction classification technique can be applied into the data set to achieve the desired goal. [4 marks]

d) The company "General Electric Inc." claims that a certain brand of its flashlight battery lasts on average 300 hours of flashlight use. You suspect that the population of batteries average fewer

than 300 hours. You select a random sample of 49 batteries and obtain a sample mean of 290 and a sample standard deviation S=70.

i. Perform a one-tailed hypothesis test with the company's claim in the null hypothesis. Use a level of significance of .10. [3 marks]

ii. Calculate the probability value of the test statistic. Interpret the results. [3 marks]

e) Explain the importance of Data mining analysis and give examples of data mining techniques. [4 marks]

Question TWO (15 Marks)

a) Wasike and Makori did a study on feelings of stress and life satisfaction due to low quality of electrical energy in their locations. Participants completed a measure how stressed they were feeling on a scale of 1 to 30 and a measure of how satisfied they felt with their lives measured on a scale of 1 to 10. The table below indicates the participants' scores. Use this data to answer the questions that follow:

Participant #	Stress score (x)	Life satisfaction(y)
1	11	7
2	25	1
3	19	4
4	7	9
5	23	2
6	6	8
7	11	8
8	22	3
9	25	3
10	10	6
Sum Mean Sd	159 15.9 7.23	51 5.1 2.70

i. Calculate the correlation (r) between stress and life satisfaction. [4 marks]

ii. Write a brief interpretation of the correlation, including the strength, direction and an explanation of the effect of l low quality of electrical energy in their locations. [2 marks]

b) Outliers, which are observations that appear far away and diverge from an overall pattern in a sample, are likely to be found in a sample dataset. Assuming you have designed an experiment in the field to collect data using two categories of sensors, that is, humidity and temperature sensors and data is collected at specified intervals.

i. With reasons, identify 4 potential causes of outliers in your data set. [2 marks]

ii. Discuss the potential effect of the outliers on a dataset obtained from your experiment. [3 marks]

c) Suppose a sample of 16 light trucks is randomly selected off the assembly line. The trucks are driven 1000 miles and the fuel mileage (MPG) of each truck is recorded. It is found that the mean MPG is 22 with a SD equal to 3. The previous model of the light truck got 20 MPG.

i. State the null hypothesis for the problem above. [1 marks]

ii. Conduct a test of the null hypothesis at p=0.05 and make appropriate conclusion. [3

marks]

Question THREE (15 Marks)

a) When conducting data analysis, different statistics are obtained that help in the interpretation of your results. One of these statistics is the p-value. Explain how you will interpret a p-value from your data analysis. [2 marks]

b) Most data scientists describe data as structured or unstructured. With examples, show the distinction of the two grouping. [3 marks]

f) The following table shows the hours of sunshine, x, during nine days in October and the number of ice creams y sold by a beach shop in Mombasa.

x	4.3	6.9	0.0	10.4	5.2	1.8	8.0	9.2	2.1
y	224	208	123	419	230	184	362	351	196

i. Establish an equation of the regression line of y on x. [3 marks]

ii. Calculate the residuals for the days when the number of hours of sunshine was 8.0 and 6.9. [2 marks]

c) Predictive modelling continues to play a central role in determining future outcomes using historical data. They have been used by corporate organizations to predict potential future profits based on current investments. Consider the data set below, which relates to mid-year and end-year investments results of ten business entities.

Business Entity	Mid – Year	End-year
1.	98	90
2.	66	74
3.	100	98
4.	96	88
5.	88	80
6.	45	62
7.	76	78
8.	60	74
9.	74	86
10.	82	80

Develop a regression model which may be used to predict end-year investment results given the mid – year business performance in the future assuming the business environment will remain the same business investment. [5 marks]

Question FOUR (15 Marks)

a) Suppose that the thickness of a part used in a semiconductor is its critical dimension and that measurements of the thickness of a random sample of 18 such parts have the variance s2 = 0.68, where the measurements are in thousandths of an inch. The process is considered to be under control if the variation of the thickness is given by a variance not greater than 0.36.

Assuming that the measurements constitute a random sample from a normal population, state the null hypothesis and test it against the alternative hypothesis at the $\alpha = .05$ significance level. [6 marks]

b) Suppose the cooking system in your school is powered by firewood. An experimental study is commissioned to find the best firewood to use. The burning rate of the firewood is an important product characteristic. It is assumed the expected mean burning rate must be 50 centimeters per second. Further, if it is assumed that the standard deviation of burning rate is 2 centimeters per second.

The experimenter decides to specify a type I error probability or significance level of 0.05 and selects a random sample of 25 and obtains a sample average burning rate of centimeters per second. What conclusions should be drawn? [5 marks]

c) A large company dealing with production of household products has 15,000 workers. If the arithmetic mean of salaries is sh.500 per day and the standard deviation is 100, how many workers have a salary between sh. 400 and sh. 650 per day? [4 marks]