Reg. No. : | E | N | G | G | T | R | E | E | . | C | O | M |

Question Paper Code : 50010

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2024.

Third/Fourth Semester

Artificial Intelligence and Data Science

AD 3491 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

(Common to: Computer Science and Business systems)

(Regulations 2021)

Time : Three hours | Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. How does confusion matrix define the performance of classification algorithm?

2. Differentiate between structured and unstructured data.

3. How regression toward the mean differs other parameters? Give an example.

4. What are outliers in the data?

5. Compare between one–tailed and two–tailed tests.

6. State the Central Limit Theorem.

7. Differentiate t–test and ANOVA.

8. Write a note on F–test.

9. What is survival analysis?

10. Why do we need weighted resampling?

PART B — (5 × 13 = 65 marks)

11. (a) (i) Explain in detail about the benefits and uses of data science with counter examples. (6)

(ii) Describe in depth about exploratory data analysis techniques. (7)

Or

(b) (i) Illustrate in detail about different facets of data with examples. (7)

(ii) Draw and outline the step-by-step activities in the data science process. (6)

12. (a) The frequency distribution for the length, in seconds, of 100 telephone calls was:

Time (seconds)	Frequency
0–20	0
21–40	5
41–60	7
61–80	14
81–100	28
101–120	21
121–140	13
141–160	9
161–180	3

Compute mean, median and variance.

Or

(b) (i) The wind speed X in miles per hour and wave height Y in feet were measured under various conditions on an enclosed deep water sea, with the results shown in the table.

X	0	2	7	9	13	22
Y	0	5	10	14	22	31

Create a scatter plot and predict the type of correlation. (6)

(ii) Assume that an r of $-.80$ describes the strong negative relationship between years of heavy smoking *(X)* and life expectancy *(Y)*. Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares:

$\overline{X} = 5 \quad \overline{Y} = 60$

$SS_x = 35 \quad SS_y = 70$

Determine the least squares regression equation for predicting life expectancy from years of heavy smoking. (7)

13. (a) Imagine that one of the following 95 percent confidence intervals estimates the effect of vitamin C on IQ scores:

95% Confidence Interval	Lower Limit	Upper Limit
1	100	102
2	95	99
3	102	106
4	90	111
5	91	98

(i) Which one most strongly supports the conclusion that vitamin C increases IQ scores? (4)

(ii) Which one implies the largest sample size? (3)

(iii) Which one most strongly supports the conclusion that vitamin C decreases IQ scores? (3)

(iv) Which one would most likely stimulate the investigator to conduct an additional experiment using larger sample sizes? (3)

Or

(b) (i) Exemplify in detail about the significance of z–test, its procedure and decision rule with example. (6)

(ii) A study finds that racism in cricket event more often takes place when the game is played in England or Australia or New Zealand (Say EAN countries). Given that

- Racism takes place or Game is played in EAN is 9/13
- Racism takes place and Game is played in EAN is 5/7
- Game is played in EAN given that Racism takes place is 4/5

Find the probability of

- No Racism takes place
- Game is played in EAN
- Racism takes place given that Game is played in EAN (7)

14. (a) A manufacturer of a gas additive claims that it improves gas mileage. A random sample of 30 drivers tests this claim by determining their gas mileage for a full tank of gas that contains the additive (X_1) and for a full tank of gas that does not contain the additive (X_2). The sample mean difference, $\overline{D}$, equals 2.12 miles (in favor of the additive), and the estimated standard error equals 1.50 miles.

(i) Using t, test the null hypothesis at the .05 level of significance. (5)

(ii) Specify the p – value for this result. (4)

(iii) Are there any special precautions that should be taken with the present experimental design? (4)

Or

(b) (i) A library system lends books for periods of 21 days. This policy is being reevaluated in view of a possible new loan period that could be either longer or shorter than 21 days. To aid in making this decision, book-lending records were consulted to determine the loan periods actually used by the patrons. A random sample of eight records revealed the following loan periods in days: 21, 15, 12, 24, 20, 21, 13, and 16. Test the null hypothesis with t–test, using the .05 level of significance. (7)

(ii) A random sample of 90 college students indicates whether they most desire love, wealth, power, health, fame, or family happiness. Using the .05 level of significance and the following results, test the null hypothesis that, in the underlying population, the various desires are equally popular using chi-square test.

Desires of college students							
Frequency	Love	Wealth	Power	Health	Fame	Family Hap.	Total
Observed (f_0)	25	10	5	25	10	15	90

(6)

15. (a) Illustrate in depth about time series forecasting, its components, moving averages and its various methods with examples.

Or

(b) (i) Compare and contrast between multiple regression and logistic regression techniques with examples. (6)

(ii) A company manufactures an electronic device to be used in a very wide temperature range. The company knows that increased temperature shortens the life time of the device, and a study is therefore performed in which the life time is determined as a function of temperature. The following data is found:

Temperature in Celcius (t)	10	20	30	40	50	60	70	80	90
Life time in hours(y)	420	365	285	220	176	117	69	34	5

Find the linear regression equation. Also find the estimated life time when temperature is 55. (7)

PART C — (1 × 15 = 15 marks)

16. (a) An investigator polls common cold sufferers, asking them to estimate the number of hours of physical discomfort caused by their most recent colds. Assume that their estimates approximate a normal curve with a mean of 83 hours and a standard deviation of 20 hours.

(i) What is the estimated number of hours for the shortest– suffering 5 percent? (3)

(ii) What proportion of sufferers estimate that their colds lasted longer than 48 hours? (2)

(iii) What proportion suffered for fewer than 61 hours? (2)

(iv) What is the estimated number of hours suffered by the extreme 1 percent either above or below the mean? (2)

(v) What proportion suffered for between 1 and 3 days, that is, between 24 and 75 hours? (3)

(vi) What proportion suffered for between 2 and 4 days? (3)

Or

(b) Admission to a state university depends partially on the applicant's high school GPA. Assume that the applicants' GPAs approximate a normal curve with a mean of 3.20 and a standard deviation of 0.30.

(i) If applicants with GPAs of 3.50 or above are automatically admitted, what proportion of applicants will be in this category? (4)

(ii) If applicants with GPAs of 2.50 or below are automatically denied admission, what proportion of applicants will be in this category? (3)

(iii) A special honors program is open to all applicants with GPAs of 3.75 or better. What proportion of applicants are eligible? (4)

(iv) If the special honors program is limited to students whose GPAs rank in the upper 10 percent, what will Brittany's GPA have to be for admission to this program? (4)