Ex/PG/MMD/T/128A/2024

Master in Multimedia Development Examination, 2024
(First Year, Second Semester)

MULTIMODAL DATA ANALYSIS

Full Marks: 100

Answer any***Five*** questions

1 a) What do you understand by the term "Multimodality" in the context of data? Mention any two modes used to capture learning experience of students when they are using an E-learning platform (2+2)

b) The average test score of all students in a class was measured to be 75 with a known standard deviation 20. If margin of error was 3.9 at a 90% confidence level (z-score=1.645), what is the total number of students in the class? (4)

c) There are two dataframes df1 and df2 with the following data: (2+2)

S#	class	left
1	1	L1
2	2	L2
3	2	L3
4	4	L4

df1

S#	class	right
1	2	R1
2	4	R2
3	5	R3

df2

Assuming that you have imported pandas with an alias pd, what will be the merged dataset on execution of the following Python statements?

i) pd.merge(df1,df2, on="class")
ii) pd.merge(df1,df2, how="left")

d) Differentiate between the following using suitable examples to illustrate where possible (any two): (4+4)

i) Data & Information
ii) Quality Assurance & Quality Control
iii) Nominal & Ordinal Data
iv) Data Mining & Data Analysis

2 a) What is Exploratory Data Analysis? Mention any two reasons why you think EDA is important. (2+2)

b) The following table shows the header and first three rows of a housing related dataset: (1*2) + (2*4)

Area	Bedrooms	Bathrooms	Mainroad	AC	Garage	Status	Price
7420	4	2	yes	yes	2	furnished	6600000
8960	4	4	yes	yes	3	furnished	4500000
9960	3	2	yes	no	1	semi-furnished	4500000

Assuming that the above dataset is imported into a Pandas dataframe df1, what will be the code for doing the following data exploration tasks:
(Attempt any two from Set A and four from Set B)

Set A

i. Remove duplicate rows in the dataframe
ii. View the last 8 rows of the dataframe
iii. Display a correlation matrix of df1 using a heatmap (Seaborn library)

Set B

i. Check if there is any missing value in any of the columns
ii. Handle missing values in the column named 'Price' which contains the dependent variable.
iii. Fill missing values in the column 'Bedrooms' using the most frequently occurring value of the column.
iv. Fill missing values in the column 'Area' using the average value of the column.
v. Create 5 equally spaced bins for the column 'Area'

c) If the p-value associated with two variables X and Y is 0.7, what does it indicate? (2)

d) The following table depicts comparison of scores for different Machine Learning(ML)models related to one particular dataset: (4)

Model	MSE	R2 score
Simple Linear Regression	0.45	0.3
Multiple Linear Regression	0.014	0.7
Random Forest Regression	0.0055	0.95

Based on the above data, which do you think is the best of the three models and why? Explain the significance of MSE and R2 scores.

3 a) Mention any two good practices for making a Visualization effective. (2)

b) Name any plot/chart which satisfies the criteria given below: (5)
i) Used for univariate graphical analysis of numerical variables
ii) Seaborn plot used for multivariate analysis
iii) Used for bivariate analysis, where one of the variables is categorical
iv) Combination of box-plot and distribution plot
v) Used for analyzing trends in the dataset

c) Why is the box plot called a 5-number summary? With reference to IQR of the plot, how are outliers detected in a Box Plot? (2+3)

d) Using Chi Square Test of Association, find if gender has anything to do with political party preferences, using sample data given in the table below: (8)

	Party-A	**Party-B**	**Total**
Male	110	90	200
Female	140	60	200
Total	250	150	400

4 a) Suppose you have a confusion matrix as given below: (6)

	Prediction = 0	Prediction = 1
Actual = 0	350	60
Actual = 1	40	550

Calculate the Precision, Recall and F1 score from the confusion matrix. Assume a proper value of β used for calculating the F1 score

b) Why does a ML Model need to be regularized? How is the regularization used in Lasso Regression different from that in Ridge Regression? (2+3)

c) Explain any one ML algorithm of your choice, which is used for Regression or Classification of data sets. (6)

d) A transaction list of a grocery shop consists of the following eight transactions: (3)

1. Milk, Eggs, Bread, Butter
2. Bread, Butter, Eggs
3. Milk, Bread, Butter, Cookies
4. Milk, Cookies
5. Milk, Butter, Bread
6. Bread, Butter, Eggs, Ketchup
7. Bread, Butter, Eggs, Cornflakes
8. Bread, Butter, Cornflakes

Using Association Rule Mining, what is the Confidence of the rule that if Bread and Butter are brought, Eggs will also be bought?

5 a) For the data set in Q 2(b) above, assume that the data of the independent variables above is stored in a dataframe X, while the data for the dependent variable (price) is stored in a dataframe Y. (6)
Write the code in Python to do the following:

i) Split the data into separate sets for training and testing: X will be split into X_train, X_test and Y into Y_train, Y_test respectively using 10% data for testing.
ii) Train the classifier using X_train and Y_train
iii) Test the model using X_test and store the output in Y_pred

b) What are k-fold Classification and Grid Search? Why are they used? (3*2)

c) Name any two Dimensionality Reduction Techniques and briefly explain the basic difference between the two. (4)

d) How is Descriptive Statistics fundamentally different from Inferential Statistics? Name any one statistical method used with each of them. (4)

6 a) Explain why a TF-IDF Vectorizer is considered better than a Count Vectorizer. (4)

b) Name a commonly used algorithm in Topic Mining. What are the main assumptions used in this algorithm. (1+2)

c) Mention any one type of text content that is considered noise. Why is it necessary to remove this type of text noise as part of text pre-processing in NLTP? Write sample code (using Python/NLTK) to show how this type of noise is removed. (2+3)

d) Explain the basic difference between Stemming and Lemmatization with an example. Mention one advantage of each over the other. (4)

e) Suppose there is a string variable named "movie_review" that contains multiple sentences containing punctuations and multiple white spaces. As part of text-preprocessing, remove these noises using code with generalized regular expressions (any two): (2*2)

i) Remove all special characters like punctuations by single space.
ii) Remove all single characters at the end of any sentence.
iii) Replace multiple white spaces with single white space.

7 a) Explain any three image processing techniques briefly, along with one practical application related to that technique. (6)

b) Name the following (any two): (2)
   i) A global processing method for edge linking after detection
   ii) A Corner Detection Technique
   iii) A Face Detection Technique

c) Name one first order and another second order derivative operator used for edge detection in images. What are the pros and cons of using a first order derivative for edge detection with respect to second order derivative? (4)

d) What is thresholding? How does global thresholding differ from adaptive thresholding? (4)

e) Suppose you have a sharpening kernel as follows: (4)

```
[0  -1  0
-1   5  -1
0`  -1  0]
```

What is the new value of the central pixel, when the kernel shown above is applied to the following 3*3 block of a Black & White Image?

```
[33   116   129
3778   127
36    78    115]
```

8 Write short notes (any four): (5*4)
   i. Data Wrangling and its importance
   ii. Finding Document Similarities
   iii. Techniques of Image Retrieval
   iv. Qualitative vs Quantitative Data
   v. Primary vs Secondary Data Collection
   vi. Bias & Variance in ML models with respect to underfitting and overfitting