Birla Institute of Technology and Science, Pilani

Mid-semester examination

Course name: Predictive Analytics | Course code: MPBAG513

Time: 4:00 PM – 5:30 PM (1.5 hours) | Total marks: 25

Attempt all the questions

One mark for each correct answer

Question no. 1-5

Briefly explain and mention the formula to calculate following performance metric using the confusion matrix notation given below.

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

1. Prevalence
2. Matthews correlation coefficient
3. F1-Score
4. Diagnostic odds ratio (DOR)
5. False Discovery Rate

Question no. 6-15

For the given data matrix (D), X1, and X2 are independent variables, and Y is a categorical dependent variable.

D

X1	X2	Y
10	2	A
20	4	A
30	8	B

6. Write the dimension/shape/size of covariance matrix
7. The covariance matrix obtained using the dot product between D^T x D would be the same as the correlation matrix. (True/False)
8. Perform Z-scaling and write the scaled data matrix as DS
9. The covariance matrix obtained using the dot product between DS^T x DS would be the same as correlation matrix. (True/False)
10. Write the correlation matrix of DS (Scaled data matrix) as C
11. Which variable (X1/X2) shows more variance
12. Calculate the Eigen values and Eigen vectors from C matrix
13. For the Eigen values calculated above, calculate the variance and cumulative variance explained from them
14. PCA loadings are calculated by taking dot product between scaled data matrix and Eigen vectors. (True/False)
15. Formula to calculate PCA scores is
    Scores = Eigenvectors $\cdot \sqrt{\text{Eigenvalues}}$ (True/False)

16. What is the Python library function to perform Z-scaling?
17. Write the answers to the following PCA-based questions
    a. Mention the methods to choose the minimum number of principal components in PCA.
    b. Write two applications of PCA.
18. Answer the following questions from KNN's distance metrics
    a. Minkowski distance with p^{th} norm where p = 1 is a special case which is also known as _____ distance
    b. Minkowski distance with p^{th} norm where p = 2 is special case which is also known as _____ distance
19. Briefly explain only one merit and one demerit of normalization of a histogram
20. Removal of a nearly unary variable from the analysis would be seen as a 'commission mistake. (True/False)
21. Keeping a variable for analysis which is the linear function of some other variable, is seen as an 'Omission mistake. (True/False)
22. Seaborn's jointplot function can be used to visualize a 2-d histogram. (True/False)
23. A probability plot can be used to check the normality of a distribution - (True/False)
24. Pandas' **iloc** property can be used to subset the Pandas' data frame into X (Predictors) and Y(Target), but the **loc** property can not be used for the same purpose. (True/False)
25. A 2x2 contingency table can not be used to draw a clustered/grouped/stacked bar chart. (True/False)