Code No: **20CS4501A**

**PVP20**

PVP Siddhartha Institute OF TECHNOLOGY

(Autonomous)

III B.Tech - I Semester Regular Examinations, JULY-2022

**DATA SCIENCE**

**Duration: 3 Hours Max. Marks: 70**

Note:

1. Contains 5 essay questions with an internal choice. Each question carries 14 Marks.
2. All parts of the Question paper must be answered in one place.

 5 x 14 = 70 Marks

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Blooms Level | CO | Max. Marks |
| **UNIT-I** |
| 1 | (a) | Outline the various phases of Data Science. | 2 | CO1 | 7 |
| (b) | What is Exploratory Data Analysis? Explain any two types of visualization. | 2 | CO1 | 7 |
| **OR** |
| 2 | (a) | Explain various Hyperparameter optimization techniques with suitable examples. | 2 | CO1 | 7 |
| (b) | Briefly explain the role of Data Science in various fields. | 2 | CO1 | 7 |
| **UNIT-II** |
| 3 | (a) | Given data = {2, 3, 4, 5, 6, 7; 1, 5, 3, 6, 7, 8}. Use PCA Algorithm to compute the principal component. | 3 | CO2 | 7 |
| (b) | Normalize the following group of data by using the following techniques. 200, 300, 400, 600, 1000 1. min-max normalization technique
2. z-score normalization
3. Decimal scaling.
 | 3 | CO2 | 7 |
| **OR** |
| 4 | (a) | In real-world data, tuples with missing values for some attributes are a common occurrence. Apply various pre-processing methods for handling this problem. | 3 | CO2 | 7 |
| (b) | The following are the sorted data price (in rupees) of certain items in the supermarket. 4, 8, 15, 21, 21, 24, 25, 28, 34, 36, 39, 42, 51, 57, 60 Smooth the data by using the following smoothing techniques. Consider the bin size as 3. i) Bin means ii) Bin medians iii) Bin boundaries | 3 | CO2 | 7 |
| **UNIT-III** |
| 5 | (a) | There is a class of 25 students, and the mean score of their test is 60 out of 100, with a standard deviation 4 marks from the mean. While other students of the school have a mean score of 50 on the same test. What will be the t-score for calculating the probability that school students scored not less than 60 on their tests? | 3 | CO3 | 7 |
| (b) | Find the mean and standard deviation of a normal distribution in which 7% are under 35, and 89% are under 63. | 3 | CO3 | 7 |
| **OR** |
| 6 | (a) | In a sample of 1000 cases, the mean of a particular test is 14, and the standard deviation is 2.5. assuming the distribution to be normal, Determine 1. how many students score between 12 and 15?
2. how many scores above 18
3. how many scores below 18
 | 3 | CO3 | 7 |
| (b) | Differentiate Stratified sampling and Cluster sampling techniques with examples. | 2 | CO1 | 7 |
| **UNIT-IV** |
| 7 | (a) | Using the logistic regression model with probability cutoff = 0.75, classify the following 6 customers as “Owner” or "Nonowner" : if p≥ 0.75 then the case as a “Owner". Present the results in a classification matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| Customer# | Income | Lot\_size | Ownership |
| 1 | 60.0 | 18.4 | Owner |
| 2 | 64.8 | 21.6 | Owner |
| 3 | 84.0 | 17.6 | Nonowner |
| 4 | 59.4 | 16.0 | Nonowner |
| 5 | 108.0 | 17.6 | Owner |
| 6 | 75 | 19.6 | Nonowner |

 | 3 | CO4 | 7 |
| (b) | Which specific regressors seem essential in multiple regression? How will you address this question? Discuss. | 2 | CO1 | 7 |
| **OR** |
| 8 | (a) | Apply Linear Discriminative analysis for 2D data set into a 1-dimensional plane.X1 = {(4,1) (2,4) (2,3) (3,6) (4,4)}X2 = { (9,10) (6,8) (9,5) (8,7) (10,8) | 3 | CO4 | 7 |
| (b) | Discuss the need for fitting the model in multiple regression. | 2 | CO1 | 7 |
| **UNIT-V** |
| 9s | (a) | Explain the bias/ variance dilemma about the model complexity. | 2 | CO1 | 7 |
| (b) | Given data set STr = {(xi, yi), i=1,….6}, xi**∈**ℝ a feature scalaryi∈{-1,+1} a class label.Data points in the data set are(x1,y1)=(2,-1) (x2,y2)=(7,-1) (x3,y3)=(4,+1) (x4,y4)=(1,-1) (x5,y5)=(3,+1) (x6,y6)=(6,+1)Suppose you are training a Linear Classifier f(x;a,b) = sign(ax+b) with 2-fold Cross Validation where sign(z) = $\left\{\begin{array}{c}+1,z\geq 0\\-1, z<0\end{array}\right.$Split STr into S1={(x1,y1) (x2,y2) (x3,y3)} and  S2={(x4,y4) (x5,y5) (x6,y6)} After training the classifier f on S1, we have a1=-1, b1=5 and then try to validate the classifier on S2.After training the classifier f on S2, we have a2=2, b2=-3 and then try to validate the classifier on S1Calculate the average training error in the 2-fold cross-validation.  | 3 | CO4 | 7 |
| **OR** |
| 10 | (a) | What is the holdout approach? What is the limitation of this approach? Name four alternative approaches for it. | 2 | CO1 | 7 |
| (b) | Given data set TTr = {(xi, yi), i=1,….4}, xi**∈**ℝ a data pointyi∈{-1,+1} a corresponding label.The Data points are(x1,y1)=(5,-1) (x2,y2)=(6,-1) (x3,y3)=(1,+1) (x4,y4)=(4,-1) Suppose you are training a Linear Classifier f(x;a,b) = sign(bx+a) with 2-fold Cross Validation where sign(z) = $\left\{\begin{array}{c}+1,z\geq 0\\-1, z<0\end{array}\right.$Split STr into T1={(x1,y1) (x2,y2)} and  T2={(x3,y3) (x4,y4)} Subsequently training the classifier f(x;a,b) on T1, we have a1=-3, b1=4 and then try to validate the classifier on T2.Training the classifier f(x;a,b) on T2, we have a2=1, b2=-5 and then try to validate the classifier on T1Calculate the average validation error (i.e. the cross-validation error) in the 2-fold cross-validation. | 3 | CO4 | 7 |