|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **P.V.P Siddhartha Institute of xTechnology** | | | | | | | | | | | | | | |
| **Department of Computer Science and Engineering** | | | | | | | | | | | | | | |
| **Course: B.Tech** | | | **Year: III** | | **Semester: I** | **Descriptive: I** | | | **A.Y: 2023-24** | | | | | |
| **Subject Code:**  **20CS4501A** | | | **Subject Name: Data Science** | | | | | | **Regulation: PVP20** | | | | | |
| **Duration: 1 hr 30 min** | | | | **Maximum Marks:15 Marks** | | | **Date: 31-08-2023** | | | | | | | |
| **Answer all the Questions. Each Question carries 5Marks 3×5M=15M** | | | | | | | | | | | | | | |
| **Q.No** |  | | | | | | | **Marks** | | | | **CO** | **Level** | |
| **1.** | **a)** | Explain Data Science life cycle phases and write names of the tools used for Data Science. | | | | | | **3M** | | | | **CO1** | **L4** | |
| **b)** | Explain various Hyperparameter optimization techniques with suitable examples. | | | | | | **2M** | | | | **CO1** | **L3** | |
|  | | | | | | | | | | | | | | |
| **2.** | **a)** | In real-world data, tuples with missing values for some attributes are a common occurrence and had many attributes or dimensions. Apply various pre-processing methods for handling this problem. | | | | | | **2.5M** | | | **CO2** | | | **L3** |
|  | **b)** | The following are the sorted data for price attribute (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.   1. Bin means 2. Bin medians 3. Bin boundaries.   and Apply Histograms and clustering with an example | | | | | | **2.5M** | | | **CO2** | | | **L3** |
|  | | | | | | | | | | | | | | |
| **3.** | **a)** | Explain the various probability functions for discrete and continuous random variables. | | | | | | **2.5M** | | **CO1** | | | | **L2** |
| **b)** | The probability density function is  y = k(3x2-1) in -1<=x<=2   1. else where   Find the value of k and find P(-1<=x<=0) | | | | | | **2.5M** | | **CO3** | | | | **L3** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **P.V.P Siddhartha Institute of Technology(Autonomous)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **Invigilator Signature with date:** | | **Marks Obtained:** | |
| **Department of Computer Science and Engineering** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Course: B.Tech** | | | **Year: III** | | | | | | | **Semester: I** | | | | | | | | **Objective: I** | | | | | | | | | | | | | |
| **Regulation: PVP20** | | | **Maximum Marks:10 Marks** | | | | | | | | | | | | | | | | | | | **Session: F.N.** | | | | | | | | | |
| **A.Y: 2023-24** | | | **Date: 31.08.2023** | | | | | | | | | | | | | | | | | | **Duration: 20 min** | | | | | | | | | | |
| **Subject Code: 20CS4501A** | | | | | | | | **Course Name: Data Science** | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Registered Number:** | | | | | | | | | | | | | | | | | | | | **Name:** | | | | | | | | | | | | | | | |
| **Answer all the Questions. Each Question carries ½ Mark 20×½ M =10M** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **S.No** | **Question** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | **CO** | **Level** | **Answer** |
| 1. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a confluence of multiple disciplines like Machine learning, Substantial research, Hacking skills | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Data Science | | | | | b) Data Analysis | | | | | | | | c) Descriptive Analytics | | | | | | | | | | | | | | | | | d) None | |
| 2. | Point out the correct statement. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Pre-processed data is original source of data | | | | b) Raw data is original source of data | | | | | | | | | | | c) Raw data is the data obtained after processing steps | | | | | | | | | | | | | d) None of the mentioned | | | |
| 3. | Raw data should be processed only one time. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) True | | | b) False | | | | | | |  | | | | | | | | | | | | | |  | | | | | | | |
| 4. | Among the following which is not correct application for Data Science in health care? | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) seismology | | | | b) Medical Imaging | | | | | | | | | | c) Drug Discovery | | | | | | | | | | | | | d) None | | | | |
| 5. | Choose the correct components of data science. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) domain expertise | | | | b) data engineering | | | | | | | | | c) advance computing | | | | | | | | | | | | | d) all of the above | | | | | |
| 6. | Which of the following is not a part of the data science process? | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) communication building | | | | | | b) operationalize | | | | | | | | | c) model planning | | | | | | | | | | | | d) Discovery | | | | |
| 7. | \_\_\_\_\_ studies the collection, analysis, interpretation or explanation, and presentation of data. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Visualization | | | | | b) Statistics | | | | | | | | c) Data Mining | | | | | | | | | | | | | d) Clustering | | | | | |
| 8. | \_\_\_ are used when we want to visually examine the relationship between two quantitative variables. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Bar graph | | | | | b) Scatter plot | | | | | | | | c) line graph | | | | | | | | | | | | | d) pie chart | | | | | |
| 9. | The following set of data is categorical or numerical data? red, red, blue, green, red, green, red, red, blue, blue | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) numerical | | | | | b) categorical | | | | | | | | c) categorical and numerical | | | | | | | | | | | | | d) none | | | | | |
| 10. | Issues of data quality problems are: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Noise and outliers | | | | | | | | b) Duplicate data | | | | | | | | | | c) Missing values | | | | | | | | | | | d) All | | |
| 11. | Which is the correct statement for data pre-processing? | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) To improve quality of data | | | | | | | | b) To improve the size of data | | | | | | | | | | c) To ensure the data presence | | | | | | | | | | | d) None | | |
| 12. | Among the following options identify the one which is false regarding regression. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) prediction | | | b) discovers casual relationships | | | | | | | c) interpretation | | | | | | | | | | | | | d) relates inputs to outputs | | | | | | | | |
| 13. | The correct way of pre-processing the data should be- | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Imputation ->feature scaling-> training | | | | | | b) Feature scaling->imputation->training | | | | | | | | | | c) Feature scaling->label encoding->training | | | | | | | | | | | | | d) None | | |
| 14. | To remove noise and inconsistent data \_\_\_\_ is needed. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Data Transformation | | | | | b) Data Cleaning | | | | | | | | c) Data Reduction | | | | | | | | | | | | d) Data discretization | | | | | | |
| 15. | Which of the following is true about outliers - | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a)  Data points that deviate a lot from normal observations | | | | | | | b) Can reduce the accuracy of the model | | | | | | | | | | | c) None | | | | | | d) Both A and B | | | | | | | |
| 16. | Some of the Imputation methods are - | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Imputation with mean/median | | | | | b) Imputing with random numbers | | | | | | | | c) Imputing with default value | | | | | | | | | | | | | d) All of the above | | | | | |
| 17. | In transformation, the operation performed on data is \_\_\_\_\_\_\_\_\_\_\_\_ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) Sampling | | | | | b) Aggregation | | | | | | | | | c) PCA | | | | | | | | | | | | d) Cleaning | | | | | |
| 18. | Find the value of P(X=3) if X is the discrete random variable where P(X=0)=0, P(X=1) = 1/4 and P(X=2) = ¼ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) 1/2 | b) ¼ | | | | | | | | | | | c) 1 | | | | | | | | | | | | d) 1/3 | | | | | | | |
| 19. | Determine the value c so that the following function can serve as a probability distribution of the discrete random variable x:  f(x)=c(x+4), for x=0,1,2,3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) 1/20 | b) 1/22 | | | | | | | | | | c) 1/18 | | | | | | | | | | | d) 1/16 | | | | | | | | | |
| 20. | Let X be the random variable, P(X=x) is the Probability mass function is given by   |  |  |  |  |  | | --- | --- | --- | --- | --- | | X | 0 | 1 | 2 | 3 | | P(X=x) | 0 | k | 2k | 3k | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | CO1 | L2 |  |
| a) 1/2 | b) 2/5 | | | | | | | | | | c) 1/5 | | | | | | | | | | | d) 1/6 | | | | | | | | | |