|  |
| --- |
| **P.V.P Siddhartha Institute of xTechnology** |
| **Department of Computer Science and Engineering** |
| **Course: B.Tech** | **Year: III** | **Semester: I** | **Descriptive: I** | **A.Y: 2024-25** |
| **Subject Code:****20CS4501A** | **Subject Name: Data Science** | **Regulation: PVP20** |
| **Duration: 1 hr 30 min** | **Maximum Marks:15 Marks** | **Date: 16-08-24** |
| **Answer all the Questions. Each Question carries 5Marks 3×5M=15M** |
| **Q.No** |  | **Marks** | **CO** | **Level** |
| **1.** | **a)** | Could you describe the phases and lifecycle of Data Science and identify the tools used for its implementation? | **2.5M** | **CO1** | **L4** |
| **b)** | Define Hyperparameter Optimization and identify various strategies for optimizing hyperparameter methods.  | **2.5M** | **CO1** | **L3** |
|  |
| **2.** | **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. | **2.5M** | **CO2** | **L3** |
|  | **b)** | The following are the sorted data price (in rupees) of certain items in the supermarket. 5, 9, 15, 21, 20, 24, 25, 28, 35, 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.
 | **2.5M** | **CO2** | **L3** |
|  |
| **3.** | **a)** | Explain the properties of Binomial Distribution and how mean and variance are computed. | **2.5M** | **CO1** | **L2** |
| **b)** | The severity of a disease and blood group were studied in a research project. The findings are given in the following table, known as the m x n contingency table. Can this severity of the condition and blood group are associated by using chi-square test. Calculate Chi square test statistic value.Severity of a disease classified by blood group in 1500 patients.

|  |  |  |
| --- | --- | --- |
| Condition | Blood Groups | Total |
| O | A | B | AB |  |
| Severe  | 51 | 40 | 10 | 9 | 110 |
| Moderate | 105 | 103 | 25 | 17 | 250 |
| Mild  | 384 | 527 | 125 | 104 | 1140 |
| Total | 540 | 670 | 160 | 130 | 1500 |

 | **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: 2024-25** | **Date:** | **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. | Which of the following is the first step in the data science lifecycle? | CO1 | L2 |  |
| a) Data Collection | b) Data Cleaning | c) Problem Definition | 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. | Model deployment involves: | CO1 | L2 |  |
| a)Integrating the model into a production environment | b) Monitoring model performance |  c) Retraining the model | d) All of the mentioned |
| 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. | What are some examples of data quality problems: | CO1 | L2 |  |
| a) Noise and outliers | b) Duplicate data | c) Missing values | d) All |
| 11. | Data integration involves combining data from multiple sources. Which challenge is commonly faced? | CO1 | L2 |  |
| a) Schema integration | b) Entity identification | c) Both | 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. | Which technique can be used to detect and correct inconsistencies in data? | CO1 | L2 |  |
| a) Data profiling | b) Data Integration | c)Data Reduction  | 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) only A | 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 one | d) All of the above |
| 17. | Discretization converts continuous data into categorical data. Which method divides data into equal-width intervals? | CO1 | L2 |  |
| a) Equal-width binning | b) Equal-frequency binning | c) Discretization by clustering | d) Discretization by supervised learning |
| 18. | Find the value of P(X=3) if X is the discrete random variable taking values x1, x2, x3 where P(X=0)=0, P(X=1) = 1/4 and P(X=2) = 1/4 | CO1 | L2 |  |
| a) 1/2 | b) 1/4 | 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 |