| **INSTRUCTIONS:** | |
| --- | --- |
| 1. | Figures to the right indicate maximum marks for that question. |
| 2. | The symbols used carry their usual meanings. |
| 3. | Assume suitable data, if required & mention them clearly. |
| 4. | Draw neat sketches wherever necessary. |

| Examination | : Sessional III | Seat No. | : |
| --- | --- | --- | --- |
| Date | : 07/05/2022 | Day | : Saturday |
| Time | : 9:00 a.m. to 10:15 a.m. | Max. Marks | : 36 |

| Q.1 | Do as directed: |  |
| --- | --- | --- |
| (a) | Fill in the blanks:  (I) According to Gartner, \_\_\_\_\_\_\_\_\_ is “the information assets organizations  collect, process and store during regular business activities, but generally fail to  use for other purposes.”  (II) Decision trees are \_\_\_\_\_ to address the problem of overfitting/anomalies in the  training due to noise or outliers. | [02] |
| (b) | State true or false.  (I) The agglomerative approach of hierarchical method based clustering is also  called the bottom-up approach.  (II) To reduce the number of parameters, we make the Naive Bayes conditional  independence assumption (attribute values are independent of each other given  the class). | [02] |
| (c) | For nominal data, provide ways to generate concept hierarchy. | [02] |
| (d) | Match the following about advanced classification methods:   | Column A | Column B | | --- | --- | | Back propagation | Overcomes the disadvantage of rule-based systems that they involve sharp cut-offs for continuous attributes. | | Fuzzy approach | Attempt to incorporate ideas of natural evolution. | | Genetic algorithm | Use/s a database of problem solutions to solve new problems. | | Case based reasoning classifiers | Learns by iteratively processing a data set of training tuples, comparing the network’s prediction for each tuple with the actual known target value | | [02] |
| (e) | How is the elbow method used to determine heuristically the number of clusters in cluster analysis? | [02] |
| (f) | What is the drawback of K-Means clustering algorithm? Explain in brief. | [02] |
|  |  |  |
| Q.2 | Attempt the following: |  |
| (a) | (I) Describe any three requirements (area of research/development) of cluster  analysis in data mining.  (II) Describe clustering methods in brief. | [03]  [03] |
| (b) | Provide introductory information about Hadoop EcoSystem components (HBse, Hive, Pig) and steps on how to set up a Hadoop cluster. | [06] |
|  |  |  |
| Q.3 | Attempt the following: |  |
| (a) | (I) How are decision trees used for classification?  (II) What is the purpose of the confusion matrix? Also, Define accuracy as a function of sensitivity and specificity. | [03]  [03] |
| (b) | Perform Attribute Oriented Induction for Data Characterization on the following initial working relation and explain your reasoning.   | name | gender | major | birth\_place | birth\_date | residence | phone | gpa | | --- | --- | --- | --- | --- | --- | --- | --- | | [06] |