UNIT 5			
Outliers: Introduction:Challenges of outlier detection ,Outlier detection methods: Introduction ,Supervized and Semi-supervized methods ,Unsupervized methods.Statistical and Proximity based methods ,Statistical approaches,Statistical data mining,Data mining and recommender systems,Data mining for financial data analysis,Data mining for Intrusion detection			
PART-A (Multiple Choice Questions)			
Q. No	**Questions**	**Course Outcome**	**Competence BT Level**
1	**The most representative average value of data set which consists of too many outliers is** a) Mean value b) Mode value c) Median value d) Value of Standard Deviation **Ans : c) Median value**	CO5	BT4
2	**When a data set contains only one outlier, what statistical measures will sustain when the outlier has been removed?[L2]** a) Standard deviation b) Range c) Mode d) Mean **Ans: c) Mode**	CO5	BT4
3	**When a bowler scores the following scores after 5 games: 205, 196, 280, 202 and 197.** **What is the bowler's mean score increase when the outlier is considered; compared to when there is no outlier to be considered?** a) 22 b) 16 c) 19 d) 13 **Ans: b) 16**	CO5	BT4
4	**John calculated participant's ages in a picnic with the below data set: 12, 12, 14, 15, 16, 20, 24, 28, 32, 35, 36. What will be difference between the mean and the median ages of people at the picnic when the arrival of John's 72-year old grandfather?** a) Difference will be increased by 4.3 b) Difference will be increased by 3.3	CO5	BT5

	c) Difference will be increased by 4.1 d) Difference will be increased by 2.1 **Ans: d)** Difference will be increased by 2.1		
5	**A real estate owner proposed to deliver the customers with an idea of housing prices in a neighborhood. All the houses are priced within $40,000 of one another, except one house that is much larger and more expensive than the rest. Which measure(s) of central tendency will be most affected by the one expensive house?** a) Mean only b) Mode only c) Both mode and mean d) Both mean and median **Ans: a) Mean only**	CO5	BT5
6	**With the given data set: 56, 64, 73, 59, 98, 65 and 59. which measure of central tendency will be least affected If the outlier is removed** a) mean b) median c) mode d) range **Ans: c) mode**	CO5	BT5
7	**What is the most appropriate strategy for cleaning the data before performing clustering analysis, by giving less than desirable number of data points?** **i) Capping and flouring of variables** **ii) Removal of outliers** a) i) only b) ii) only c) i) and ii) d) Not possible **Ans: a) 1 only**	CO5	BT4
8	**Point out the algorithm which is most sensitive to outliers.** a) K-medoids clustering b) K-medians clustering c) K-modes clustering d) K-means clustering **Ans d) K-means clustering**	CO5	BT4
9	**Identify the correct recommendation system's algorithm(s) from given options.**	CO5	BT3

	a) Collaborative page ranking b) Collaborative filtering c) Item based recommendation system d) Object based recommendation system **Ans: b) Collaborative filtering**		
10	**Statistical significance is referred as** A) The science of gathering, organizing and applying numerical facts or data B) A measure of probability that some hypothesis is incorrect with the given observations. C) An aspect of a data warehouse that has been specially built around the existing applications of operational data D) Art of specifying a Knowledge data base with respect to future data set **Answer: B**	CO5	BT4
11	**Among the following which one is NOT a statistical processing software package?** a) Minitab b) SAS c) Mahout d) Vertica **Ans: d) Vertica**	CO5	BT3
12	**Statistical significance and transparency is closely related to_____** a) Classification Accuracy b) Search Complexity c) Statistical significance d) Transparency **Ans: d) Transparency**	CO5	BT3
13	**Which is not a type of outlier from the following?**	CO5	BT2

	a. Collective Outliers b. Contextual Outliers c. Extrinsic Outliers d. Global Outliers **Ans: c. Extrinsic Outliers**		
14	**Self organizing maps are example for _____** a) Semi supervised learning b) unsupervised learning c) Supervised learning d) Missing data imputation **Ans: b) unsupervised learning**	CO5	BT3
15	**In supervised learning** a) classes are not predefined b) classes are predefined c) classes are not required d) classification is not done **Ans: b) classes are predefined**	CO5	BT3
16	**With a large dataset of medical records for patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments. Point out what kind of learning problem is this?** a. Supervised Learning b. Unsupervised Learning c. None of the given answers d. Reinforcement Learning **Ans: b. Unsupervised Learning**	CO5	BT3
17	**What are the characteristics of signature based IDS?** a) Most are based on simple pattern matching algorithms b) It is programmed to interpret a certain series of packets c) It models the normal usage of network as a noise characterization	CO5	BT4

	d) Anything distinct from the noise is assumed to be intrusion activity **Answer: a**		
18	**A test is administered annually. The test has a mean score of 100 and a standard deviation of 10. If Raju's z-score is 2.50, what was his score on the test? [L2]** a) 125 b) 140 c) 20 d) 50 **Ans a) 125**	CO5	BT5
19	**A test of statistical significance indicates how confident the researcher is about:** **a) understanding the data set** **b) passing a test about their significant other.** **c) the inter-coder reliability of their structured interview schedule.** **d) generalising their findings from the sample to the population.** **Ans: d)** generalising their findings from the sample to the population.	CO5	BT4
20	**What does the term 'outlier' mean?** **a) A score that is left out of the analysis because of missing data** **b) The arithmetic mean** **c) A type of variable that cannot be quantified** **d) An extreme value at either end of a distribution** **Ans: d)** An extreme value at either end of a distribution	CO5	BT4
21	**What kind of algorithm is used for facial identifiers or facial expressions?** a) Prediction b) Recognition patterns c) Recognizing anomalies d) Generating patterns **Ans: b) Recognition patterns**	CO5	BT4

22	**Machine learning techniques differ from statistical techniques in that machine learning methods** a) typically assume an underlying distribution for the data. b) are better able to deal with missing and noisy data. c) are not able to explain their behavior. d) have trouble with large-sized datasets **Ans: b) are better able to deal with missing and noisy data.**	CO5	BT4
23	**Which statement about outliers is true?** a) Outliers should be identified and removed from a dataset. b) Outliers should be part of the training dataset but should not be present in the test data. c) Outliers should be part of the test dataset but should not be present in the training data. d) The nature of the problem determines how outliers are used. **Ans: d) The nature of the problem determines how outliers are used.**	CO5	BT3
24	**This supervised learning technique can process both numeric and categorical input attributes.** a) linear regression b) Bayes classifier c) logistic regression d) backpropagation learning **Ans: a) linear regression**	CO5	BT3
25	**Which of the following is a common use of unsupervised clustering?** a) detect outliers b) determine a best set of input attributes for supervised learning c) evaluate the likely performance of a supervised learner model d) determine if meaningful relationships can be found in a dataset **Ans: a) detect outliers**	CO5	BT4
26	**Figure out outlier detection method based on the model given below scenarios:Outlier detection is modeled based on classification problem labeled via domain experts and It models data normality and abnormality** **a) Semi-Supervised methods**	CO5	BT4

	b) Statistical methods **c) Supervised methods** **d) Unsupervised methods** **Ans. c) Supervised methods**		
27	**Figure out outlier detection method based on the model given below:** Outlier detection methods are used for applications where sample objects are labeled as 'normal' or 'outlier' and are not available **a) Semi-Supervised methods** **b) Statistical methods** **c) Supervised methods** **d) Unsupervised methods** **Ans. d) Unsupervised methods**	CO5	BT4
28	**Figure out outlier detection method based on the model given below:** Outlier detection methods are used where number of labeled samples available are relatively very small. **a) Semi-Supervised methods** **b) Statistical methods** **c) Supervised methods** **d) Unsupervised methods** **Ans. a) Semi-Supervised methods**	CO5	BT4
29	**Figure out outlier detection method based on the model given below:** Outlier detection methods usually make assumptions of data normality. They assume that the normal data objects are generated by stochastic models and the data not following the model are outliers. **a) Semi-Supervised methods** **b) Statistical methods** **c) Supervised methods** **d) Unsupervised methods** **Ans. b) Statistical methods**	CO5	BT4
30	**Guess the outlier detection method based on the description given.** **a) Statistical methods** **b) Clustering based methods** **c) Semi-Supervised methods** **d) Proximity-Based methods** The parameter for an object to be an outlier or not depends on the distance between the object and its nearest neighbor. **Ans. d) Proximity-Based methods**	CO5	BT4