UNIT 3

Classification: Basic concepts -General approach to Classification - Decision tree induction - Algorithm for Decision tree induction - Numerical example for Decision tree induction - Attribute selection measure - Tree pruning - Scalability and Decision tree induction - Bayes' Theorem - Naïve Bayesian Classification - IF-THEN rules for classification - Rule extraction from a decision tree - Metrics for evaluating classifier performance -Cross validation - Bootstrap - SLO-2 Ensemble methods-Introduction - Bagging and Boosting - Random Forests: Introduction.

PART-A (Multiple Choice Questions)

Q. No	Questions	Course Outcome	Competence BT Level

1	Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?
	     a. Decision Tree
	     b. Regression.
	     c. classification
	     d. **Random Forest**
		CO3	BT 4

2	To find the minimum or the maximum of a function, we set the gradient to zero because:
	     a. **The value of the gradient at extrema of a function is always zero.**
	     b. Depends on the type of problem
	     c.Confusion Matrix
	     d. Classification
		CO3	BT5

3	Which of the following is a disadvantage of decision trees?
	     a.Factor analysis
	     b.decision trees are robust to outliers
	     c.decision trees are robust to outliers
	     d.decision trees are prone to overfit
		CO3	BT4

4	Identify the kind of learning algorithm for "facial identities for facial expressions" a. Prediction **b. Recognition patterns** c. Recognition anomalies d. Generating patterns	CO3	BT4
5	What is the application of machine learning methods to a large database called? a. Big data computing b. Internet of things **c. Data Mining** d. Artificial Intelligence	CO3	BT4
6	What is Machine learning? **a) The autonomous acquisition of knowledge through the use of computer programs** b) The autonomous acquisition of knowledge through the use of manual programs c) The selective acquisition of knowledge through the use of computer programs d) The selective acquisition of knowledge through the use of manual	CO3	BT4

	programs		
7	Which of the following machine learning techniques helps in detecting the outliers in data? a. classfication. b. Clustering. **c. Anamoly detection** d. Regression	CO3	BT4
8	A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. **a) Decision tree** b) Graphs c) Classifiers d) Neural Networks	CO3	BT4
9	_____is used for cutting or trimming the tree in Decision trees. **a. Pruning** b. Stemming c. classifiers d. Regressors	CO3	BT2
10	_____ measure of the randomness in the information being processed in the Decision Tree. **a. Entropy** b. Information Gain c. Confusion matrix. d. Accuracy	CO3	BT4
11	_____is a statistical property that measures how well a given attribute separates the training examples according to their target classification. a Entropy **b. Information Gain** c. Confusion matrix. d. Accuracy	CO3	BT4
12	_____computes the difference between entropy before the split and average entropy after the split of the dataset based on given attribute values. **a. Information gain** b. Gini ratio c. Pruning d. Accuracy	CO3	BT5
13	Information gain is biased towards choosing attributes with a large number of values as _____. a) Branch nodes **b) Root nodes** c) Leaf nodes d) internal nodes	CO3	BT4
14	_____is an algorithm used for continuous target variables that are used for regression problems in Decision Tree. **a) Reduction in Variance** b) Collinearity c) Correlation	CO3	BT4

	d) Multicollinearity		
15	Decision tree is used for _____. a) Regression. b) Classification. **c) Regression and Classification** d) Clustering.	CO3	BT4
16	Choose the correct sequence of typical decision tree structure – (I) Take the entire data set as input (II) Divide the input data into two part (III) Reapply the split to every part recursively (IV) Stop when meeting desired criteria (V) Cut the tree when we went too far while doing splits(pruning) a) (I), (II),(V),(IV),(III). b) (V),(I),(III),(II),(IV). c) (I),(III),(II),(V),(IV). **d) (I),(II),(III),(IV),(V).**	CO3	BT5
17	_____ denotes the entire population or sample and it further divides into two or more homogeneous sets. a) Leaf node. b) Terminal node. **c) Root node.** d) Internal Nodes	CO3	BT4
18	The Process of removing sub-nodes from a decision node is called _____. a) Splitting. b) Breaking. **c) Pruning.** d) Pooling	CO3	BT4
19	_____ is a measurement of likelihood of an incorrect classification of a new instance for a random variable, if the new instance is randomly classified as per the distribution of class labels from the data set. **a) Gini impurity.** b) Entropy. c) Information gain. d) Regularization	CO3	BT4
20	Which nodes have the maximum Gini impurity in a decision tree? a. 0.6 b. 0.1 **c. 0.5** d. 0.2	CO3	BT4
21	Where does the bayes rule can be used? a) Solving queries b) Increasing complexity c) Decreasing complexity **d) Answering probabilistic query**	CO3	BT4
22	How the bayesian network can be used to answer any query? a) Full distribution **b) Joint distribution**	CO3	BT4

	c) Partial distribution d) NULL		
23	Point out the wrong combination. a) True negative=correctly rejected b) False negative=correctly rejected **c) False positive=correctly identified** d) True positive=wrongly rejected	CO3	BT4
24	Which of the following is a categorical outcome? a) RMSE b) RSquared **c) Accuracy** d) Confusion Matrix	CO3	BT5
25	Which of the following algorithm is not an example of an ensemble method? A. Extra Tree Regressor B. Random Forest C. Gradient Boosting **D. Decision Tree**	CO3	BT5
26	What is true about an ensembled classifier? 1. Classifiers that are more "sure" can vote with more conviction 2. Classifiers can be more "sure" about a particular part of the space 3. Most of the times, it performs better than a single classifier A. 1 and 2 B. 1 and 3 C. 2 and 3 D. 1, 2 and 3	CO3	BT4
27	Which of the following option is / are correct regarding benefits of ensemble model? 1. Better performance 2. Generalized models 3. Better interpretability A. 1 and 3 B. 2 and 3 **C. 1 and 2** D. 1, 2 and 3	CO3	BT4
28	True positives is **a. These refer to the positive tuples that were correctly labeled by the classifier.** b. These are the negative tuples that were correctly labeled by the classifier. c. These are the negative tuples that were incorrectly labeled as positive. d. These are the positive tuples that were mislabeled as negative	CO3	BT4
29	Specificity is called as _____ a. accuracy **b. true positive rate**	CO3	BT5

	c. true negative rate d. precision		
30	The _____ method samples the given training tuples uniformly with replacement. **a. bootstrap** b. Accuracy c. precision d. specificity	CO3	BT4

<div align="center">

PART B (5 Marks)

</div>

1	Discuss the steps in General approach to Classification.	CO3	BT4
2	Write the basic algorithm for inducing a decision tree from training tuples.	CO3	BT3
3	Compare on various Attribute Selection Measures for building the decision tree.	CO3	BT4
4	Explain, how tree pruning avoid overfitting.	CO3	BT4
5	Write the steps followed in building the decision tree using Information gain.	CO3	BT3
6	Discuss on Naïve Bayes Classifier.	CO3	BT3
7	With a neat sketch explain a Cross-validation.	CO3	BT3
8	Discuss on confusion Matrix.	CO3	BT3
9	With suitable example, elaborate on Ensemble Method.	CO3	BT4
10	Explain about random forest.	CO3	BT3

<div align="center">

PART C (12 Marks)

</div>

1	Build a decision tree for the give dataset using ID3 algorithm. | age	income	student	credit_rating	Class: buys_computer | |---|---|---|---|---| | youth | high | no | fair | no | | youth | high | no | excellent | no | | middle_aged | high | no | fair | yes | | senior | medium | no | fair | yes | | senior | low | yes | fair | yes | | senior | low | yes | excellent | no | | middle_aged | low | yes | excellent | yes | | youth | medium | no | fair | no | | youth | low | yes | fair | yes | | senior | medium | yes | fair | yes | | youth | medium | yes | excellent | yes | | middle_aged | medium | no | excellent | yes | | middle_aged | high | yes | fair | yes | | senior | medium | no | excellent | no |	CO3	BT6
2	Identify whether student(X) will buy a computer or not using Naive Bayesian classifier when X= (YOUTH, MEDIUM, YES, FAIR, ?).	CO3	BT6

Training data is given below.

age	income	student	credit_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

3	Show how rules are extracted from the Decision tree, also discuss on Tree pruning.	CO3	BT5
4	Discuss the techniques to improve the classification accuracy.	CO3	BT4
5	Explain the Metrics for evaluating classifier performance.	CO3	BT4