1.High entropy means that the partitions in classification are

Pure-ok

not Pure

useful

useless

2. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?

Model A

Model B

both are performing equal

Data Insufficient

3.What can be the disadvantage if the learning rate is very high in gradient descent?

Gradient Descent algorithm can diverge from the optimal solution.

Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.

Both of them

None of them

**Question 4**

Marks: 1

In which of the following you can say that the model is overfitting?

High R-squared value for train-set and High R-squared value for test-set.

Low R-squared value for train-set and High R-squared value for test-set.

High R-squared value for train-set and Low R-squared value for test-set.

None of the above

**Question 5**

True or False: Training data is used to optimize the parameter settings of a supervised learner model.

True

False

If k-NN is very likely to overfit due to very high number of features. Which of the following can be used to handle such problem?

Dimensionality Reduction ok

SMOTE

Feature Selection

None of the above

**Question 7**

Marks: 1

Which of the following statements about Naive Bayes is incorrect?

 Attributes are equally important.

Attributes are statistically dependent of one another given the class value.-ok

 Attributes are statistically independent of one another given the class value.

Attributes can be nominal or numeric

**Question 8**

Which statement about outliers is true? Select one:

Outliers should be part of the training dataset but should not be present in the test data.

Outliers should be identified and removed from a dataset.

The nature of the problem determines how outliers are used-ok

Outliers should be part of the test dataset but should not be present in the training data.

**Question 9**

Which of the following in sklearn library is used for hyper parameter tuning?

GridSearchCV()

train\_test\_split()

K-fold Cross Validation

None of the above

**Question 10**

Marks: 1

K-Mean algorithm has some limitations. One of the limitation it has is, it makes hard assignments (A point either completely belongs to a cluster or not belongs at all) of points to clusters.  
Note: Soft assignment can be consider as the probability of being assigned to each cluster: say K = 3 and for some point xn, p1 = 0.7, p2 = 0.2, p3 = 0.1)  
Which of the following algorithm(s) allows soft assignments?  
Gaussian mixture models  
Fuzzy K-mean

s  
Options:

A only

B only

**A and B**

None of these

**Question 11**

Marks: 1

Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Logistic Regression

Naïve Bayes Classifier

Decision Tree Classifier

Support Vector Classifier

**Question 12**

Marks: 1

Which of the following is an ensemble technique?

SVM

Logistic Regression

Random Forest

AdaBoost

**Question 13**

Marks: 1

In which of the below ensemble techniques trees are trained in parallel?

Random forest

Adaboost

Gradient Boosting

All of the above

**Question 14**

Marks: 1

Which of the following cannot be used for dimensionality reduction?

Lasso Regularisation

PCA

Recursive feature elimination

Ridge Regularisation

**Question 15**

Marks: 1

Clustering

is \_\_\_\_\_\_\_\_\_\_\_ and is an example of \_\_\_\_\_\_\_\_\_\_\_\_learning Select one:

Predictive and supervised

Predictive and unsupervised

Descriptive and supervised

Descriptive and unsupervised

**Question 16**

Marks: 1

A nearest neighbour approach is best used:

with large-sized datasets

when irrelevant attributes have been removed from the data

when a generalized model of the data is desirable

when an explanation of what has been found is of primary importance

**Question 17**

Marks: 1

Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?

Accuracy

Sensitivity

Precision

None of the above.

**Question 18**

Marks: 1

How do you handle missing or corrupted data in a dataset?

Drop missing rows or columns

Replace missing values with mean/median/mode

Assign a unique category to the missing values

All of the above

**Question 19**

Marks: 1

Arbitrary shaped clusters can be found by using Select one:

Density methods

Partitional methods

Hierarchical methods

Agglomerative Show

The value of correlation coefficient will always be:

between 0 and 1

greater than -1

between -1 and 1

between 0 and -1

6. Feature scaling is an important step before applying the K-Mean algorithm. What is the reason behind this?

In distance calculation it will give the same weights for all features

You always get the same clusters. If you use or don’t use feature scaling

In Manhattan distance it is an important step but in Euclidean it is not

None of these

**Question 22**

Marks: 1

To find the minimum or the maximum of a function, we set the gradient to zero because:

The value of the gradient at extrema of a function is always zero

Depends on the type of problem

Both A & B

None of the above

**Question 23**

Marks: 1

With Bayes classifier, missing data items are:

treated as equal compares

treated as unequal compares

replaced with a default value

ignored

**Question 24**

Marks: 1

In which of the following cases you can say that it is over-fitting case?

When the curve you fit is passing through each and every point of the training data.

When the training MSE is much less than Test MSE.

When the MSE on both the training and test set is high.

None of the above

**Question 25**

Marks: 1

The model will be trained with data in one single batch is known as ?

Batch learning

Offline learning

Both A & B

None of the above

**Question 26**

Marks: 1

Which of the following sequences is correct for a K-Means algorithm using Forgy method of initialization?  
Specify the number of clusters  
Assign cluster centroids randomly  
Assign each data point to the nearest cluster centroid  
Re-assign each point to nearest cluster centroids  
Re-compute cluster centroids  
Options:

A, B, C, E, D

A, C, B, D, E

B, A, C, D, E

None of these

**Question 27**

Marks: 1

2. What among the following is the intrinsic cluster evaluation parameter?

Cluster homogeneity

Silhouette effect

Cluster completeness

BCubed precision and recall

**Question 28**

Marks: 1

2. Which of the following are interestingness measures for association rules? Select one:

recall

lift

accuracy

compactness

**Question 29**

Which of the following is a reasonable way to select the number of principal components "k"?

Choose k to be the smallest value so that at least 99% of the

is retained.

Choose k to be 99% of m (k = 0.99\*m, rounded to the nearest integer).

Choose k to be the largest value so that 99% of the variance is retained.

Use the elbow method.

**Question 30**

4. OPTICS is a type of?

Density based clustering

Partition clustering

Decision classification

frequent item-set mining algorithm

**Question 31**

Marks: 1

Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

Decision Tree

Random Forest

    Classification

Regression

**Question 32**

Marks: 1

1. The number of iterations in apriori \_\_\_\_\_\_\_\_\_\_\_ Select one:

increases with the size of the data

decreases with the increase in size of the data

increases with the size of the maximum frequent set

decreases with increase in size of the maximum frequent set

**Question 33**

Marks: 1

Which Association Rule would you prefer Select one:

High support and medium confidence

High support and low confidence

Low support and high confidence

Low support and low confidence

**Question 34**

Marks: 1

8. Which of the following are the spatial clustering algorithms?

Partitioning based clustering

K-means clustering

Grid based clustering

All of the above

**Question 35**

Marks: 1

6. \_\_\_\_\_\_\_\_\_ is an example for case based-learning Select one:

Decision trees

Neural networks

Genetic algorithm

K-nearest neighbour

**Question 36**

Marks: 1

Which of the following statements are true regarding soft margin SVM classifier?

They are less sensitive to outliers and can be used even in their presence

They make sure that there is no data point present in the margin area

They allow some degree of errors or misclassification

They can be used in case data is not completely linearly separable

**Question 37**

Marks: 1

7. Which of the following methods is used for finding the optimal cluster in the K-Mean algorithm?

Elbow method

Manhattan method

Euclidean method

All of the above

**Question 38**

Marks: 1

Which of the following is/are not true about DBSCAN clustering algorithm:  
For data points to be in a cluster, they must be in a distance threshold to a core point  
It has strong assumptions for the distribution of data points in dataspace  
It has substantially high time complexity of order O(n3)  
It does not require prior knowledge of the no. of desired clusters  
It is robust to outliers  
Options:

A only

B only

D only

B and C

**Question 39**

Marks: 1

Another name for an output attribute is:

predictive variable

independent variable

estimated variable

dependent variable

**Question 40**

Marks: 1

How can SVM be classified?

It is a model trained using unsupervised learning. It can be used for classification and regression.

It is a model trained using unsupervised learning. It can be used for classification but not for regression

It is a model trained using supervised learning. It can be used for classification and regression.

It is a model trained using supervised learning. It can be used for classification not for regression.

**Question 41**

Marks: 1

Which statement is true about neural network and linear regression models?

Both models require input attributes to be numeric

Both models require numeric attributes to range between 0 and 1

The output of both models is a categorical attribute value

Both techniques build models whose output is determined by a linear sum of weighted input attribute values.

**Question 42**

Marks: 1

10. Choose the correct option(s) from the following.

When working with a small dataset, one should prefer low

/high variance classifiers over high bias/low variance classifiers.

When working with a small dataset, one should prefer high bias/low variance classifiers over low bias/high variance classifiers.

When working with a large dataset, one should prefer low bias/high variance classifiers over high bias/low variance classifiers.

All are correct

**Question 43**

In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called ?

mini-batches

optimizedparameters

hyperparameters

superparameters

**Question 44**

Marks: 1

Which of the statement/s is/are true for Gradient Descent (GD) and Stochastic GD?

In GD and SGD, we update a set of parameters in an iterative manner to minimize the error function.

In SGD, we have to run through all the samples in our training set for a single update of a parameter in each iteration.

In GD, we either use the entire data or subset of training data to update a parameter in each iteration

In SGD, we have to run through single random sample in our training set.

**Question 45**

Marks: 1

Suppose I have dataset having three predictors “TV”, “Radio”, “Newspaper” and response variable “Sales”. Where predictors are the budget spent on each of the advertising medium respectively and sales is the response capturing the quantity of product sold. Suppose I want to know the impact of each of the predictors on the Sales. I want to know the relative quantities of impact of each advertising medium on the “Sales”. Which of the following machine learning methods is best suited for this situation?

Adaboost

Xgboost

Linear Regression

Decision trees

**Question 46**

Marks: 1

Which among the following is a disadvantage of decision trees?

Decision trees are prone to outliers.

Decision trees are highly prone to overfitting.

Decision trees are not easy to interpret

None of the above.

**Question 47**

Which of the following hyperparameter(s), when increased may cause random forest to over fit the data?

Number of Trees

Depth of Tree

Learning rate

None of the above

**Question 48**

Marks: 1

Suppose I have a linear regression model which is performing as follows:  
Train accuracy=0.95  
Test accuracy=0.75  
Which of the following is true regarding the model?

model is underfitting

model is overfitting

model is performing good

None of the above

**Question 49**

Marks: 1

Which points are eliminated by the DBSCAN algorithm?

Core points

Border points

Noise points

None of these

**Question 50**

Marks: 1

How many clusters are generated by the k-means algorithm?

a.3

b. K

c. Number of points divided by K

d. K-1

**Question 51**

Marks: 1

5. Which statement is true in comparison to OPTICS and BIRCH?

In detection of clusters of arbitrary shape BIRCH performs better than OPTICS

In detection of clusters of arbitrary shape OPTICS performs better than BIRCH

BIRCH is more scalable than OPTICS

OPTICS is more scalable than BIRCH

**Question 52**

Marks: 1

How is the density of point p at the density-based clustering defined?

Min-Pts minus number of data points in an epsilon-neighbourhood

Number of data points in an epsilon-neighbourhood of p

Reciprocal value of the distance from p to the nearest neighbour

All of the above

**Question 53**

Marks: 1

Which of the following are the high and low bounds for the existence of F-Score?

[0,1]

(0,1)

[-1,1]

None of the above

**Question 54**

Marks: 1

Identified outliers can be treated differently. To change the data set as little as possible you..

...remove the entire outlier.

...remove the relevant component.

...correct the relevant component

None of these

**Question 55**

Marks: 1

8. Frequent item sets is Select one:

Superset of only closed frequent item sets

Superset of only maximal frequent item sets

Subset of maximal frequent item sets

Superset of both closed frequent item sets and maximal frequent item sets

**Question 56**

Marks: 1

Which of the following measure do we use for categorical variables in k-NN?

Hamming Distance

Euclidean Distance

Manhattan Distance

All of the above

**Question 57**

Marks: 1

A measure of goodness of fit for the estimated regression equation is the:

multiple coefficients of determination

mean squared due to error

mean squared due to regression

None

3. What type of clusters can density based clustering get?

Radial

High density

Arbitrary shaped

All of the above

**Question 59**

Marks: 1

What does the Term “Variance” mean in the

-Variance refer to?

It refers to the variance of the training dataset.

It refers to the variance of the test dataset.

It refers to the variance of the model that is how much the model changes with changes in the training dataset.

All of the above

10. Which of the following sequences is correct for a K-Means algorithm using Forgy method of initialization?  
Specify the number of clusters  
Assign cluster centroids randomly  
Assign each data point to the nearest cluster centroid  
Re-assign each point to nearest cluster centroids  
Re-compute cluster centroids  
Options:

A, B, C, E, D

A, C, B, D, E

B, A, C, D, E

None of these