Quiz – Supervised Learning

1) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college. Which of the following statement is true in following case?

A) Feature F1 is an example of nominal variable.
B) Feature F1 is an example of ordinal variable.
C) It doesn't belong to any of the above category.
D) Both of these

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

1. **Number of Trees**
2. **Depth of Tree**
3. **Learning Rate**

A) Only 1
B) Only 2
C) Only 3
D) 1 and 2
E) 2 and 3
F) 1,2 and 3

3) Imagine, you are working with a website, and you want to develop a machine learning algorithm which predicts the number of views on the articles.
Your analysis is based on features like author name, number of articles written by the same author on this website in past and a few other features. Which of the following evaluation metric would you choose in that case?

1. **Mean Square Error**
2. **Accuracy**
3. **F1 Score**

A) Only 1
B) Only 2
C) Only 3
D) 1 and 3
E) 2 and 3
F) 1 and 2

4) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.
You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?

A) All categories of categorical variable are not present in the test dataset.
B) Frequency distribution of categories is different in train as compared to the test dataset.
C) Train and Test always have same distribution.
D) Both A and B
E) None of these

5) Adding a non-important feature to a linear regression model may result in.

1. **Increase in R-square**
2. **Decrease in R-square**

A) Only 1 is correct
B) Only 2 is correct
C) Either 1 or 2
D) None of these

6) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

A) 1 and 3
B) 1 and 4
C) 2 and 3
D) 2 and 4

7) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models. Which of the following statements is / are true for weak learners used in ensemble model?

1. **They don't usually overfit.**
2. **They have high bias, so they cannot solve complex learning problems**
3. **They usually overfit.**

A) 1 and 2
B) 1 and 3
C) 2 and 3
D) Only 1
E) Only 2
F) None of the above

8) Which of the following options is/are true for K-fold cross-validation?

1. **Increase in K will result in higher time required to cross validate the result.**
2. **Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.**
3. **If K=N, then it is called Leave one out cross validation, where N is the number of observations.**

A) 1 and 2
B) 2 and 3
C) 1 and 3
D) 1,2 and 3

9) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. **Number of samples used for split**
2. **Depth of tree**
3. **Samples for leaf**

A)1 and 2
B) 2 and 3
C) 1 and 3
D) 1, 2 and 3
E) Can't say

10) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

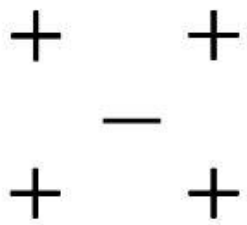
n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

1. **Accuracy is ~0.91**
2. **Misclassification rate is ~ 0.91**
3. **False positive rate is ~0.95**
4. **True positive rate is ~0.95**

A) 1 and 3
B) 2 and 4
C) 1 and 4
D) 2 and 3

Below are five samples given in the dataset. Answer question 11 & 12



Note: Visual distance between the points in the image represents the actual distance.

11) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?
A) 0
D) 0.4
C) 0.8
D) 1

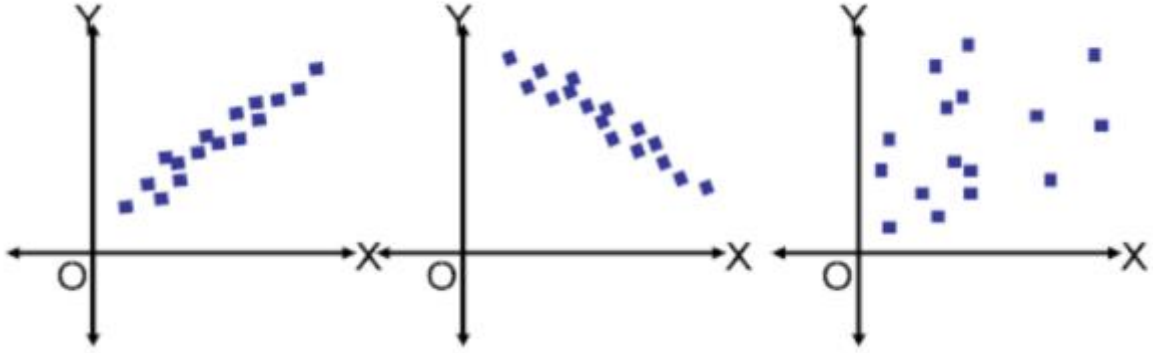
12) Which of the following value of K will have least leave-one-out cross validation accuracy?
A) 1NN
B) 3NN
C) 4NN
D) All have same leave one out error

13) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.
Note: All other hyperparameters are same and other factors are not affected.

1. **Depth 4 will have high bias and low variance**
2. **Depth 4 will have low bias and low variance**

A) Only 1
B) Only 2
C) Both 1 and 2
D) None of the above

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right). Answer questions 14-15



14) In the above images, which of the following is/are example of multi-collinear features?

A) Features in Image 1
B) Features in Image 2
C) Features in Image 3
D) Features in Image 1 & 2
E) Features in Image 2 & 3
F) Features in Image 3 & 1

15) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. **Remove both collinear variables.**
2. **Instead of removing both variables, we can remove only one variable.**
3. **Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.**

A) Only 1
B) Only 2
C) Only 3
D) Either 1 or 3
E) Either 2 or 3

Ref: AnalyticsVidhya.com