Practice Questions #2

Question #1 – Consider the dataset below, with T and F representing true and false. Choose the Gini impurity for the 'Rain' feature from the options below:

Temperature	Cloud	Rain
Low	T	T
Low	T	T
Medium	T	F
Medium	T	T
High	T	F
High	F	F

Options:

A. 0.99

B. 0.49

C. 0.29

D. 0.69

E. 1

Answer Key: B

Question #2 – Consider the concept of pruning in the context of decision trees. Which of the following statements is true?

A. Pruning is only useful for decision trees with depths greater than a specific level.

B. Pruning can lead to losing important decision rules if not carefully implemented.

C. A pruned decision tree is simpler than the original tree but might not be as accurate in its predictions.

D. The sole purpose of pruning is to handle missing data within the training set.

Answer Key: B

Question #3 – Assume we use a random forest model for a regression problem involving four observations. The projected outputs for observations 1, 2, 3, and 4 are 11, 14, 9, and 10, while the actual observed values are 8, 10, 12, and 14. What is the mean squared error?

A. 11.5

B. 14

C. 12.5

D. 10

Answer Key: C

Question #4 – Considering the gradual decline in model performance improvement with adding more trees in a bagging ensemble, what is the most efficient strategy for optimizing model complexity and computational resources?

A. Continuously add more trees indefinitely, as more trees result in better performance.

B. Evaluating the performance gain from additional trees and stopping at the point where improvements become negligible.

C. Using the smallest number of trees ensures the model runs on minimal computational resources.

D. Focusing solely on reducing the model's bias without considering the effects on variance or computational resources.

Answer Key: B

Question #5 – Given two machine learning models (Model 1 and Model 2) evaluated on their ability to predict disease presence in patients, their performance is summarized in the confusion matrix below:

Model 1 Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	80	30
Actual Negative	20	50

Model 2 Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	70	40
Actual Negative	10	90

Considering the importance of **accurately** identifying diseased patients while **minimizing** false alarms, which model demonstrates a more suitable trade-off for medical diagnosis?

A. Model 1, with its higher recall for detecting the disease, even at the cost of more false positives.

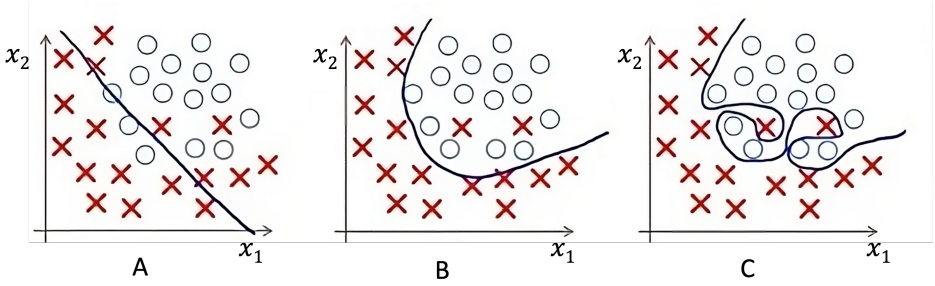
B. Model 2, prioritizing the precision of diagnosis and reducing the burden of false positives on patients and the healthcare system.

C. Model 1, with its higher number of false positives, is negligible in medical settings.

D. Model 2, as it compromises the detection of the disease for fewer false alarms, which may not be ideal in critical health scenarios.

Answer Key: B

Question #6 – What can be concluded from the visualization of the three regression models?



1. Compared to the more complex second (B) and third (C) models, the first model (A) likely has a higher training error due to its simplicity and potential underfitting.
2. Despite a minimum training error, the third model (C) may not be the best due to its complexity and potential overfitting, leading to poor generalization on unseen data.
3. The second model (B) appears more robust than the first (A) and third (C) models because it balances complexity and fit, likely performing better on unseen data.
4. The third model (C) shows signs of overfitting more than the first (A) and second (B) models, as it closely fits all training data points.
5. It's impossible to determine if all models will perform the same on unseen data without actual testing data.

Which combination of the above statements is true?

A. 1 and 3
B. 2 and 3
C. 1, 3, and 4
D. Only 5

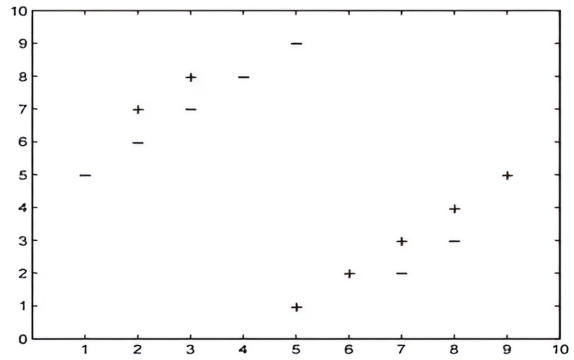
Answer Key: C

Question #7 – Determine the new centroids for each cluster as you prepare for the second iteration of the algorithm. Select the correct centroids from the options below:

A. C1 centroid is (4,4), C2 centroid is (2,2), and C3 centroid is (7,7)
B. C1 centroid is (6,6), C2 centroid is (4,4), and C3 centroid is (9,9)
C. C1 centroid is (2,2), C2 centroid is (0,0), and C3 centroid is (5,5)
D. None of the above is correct.

Answer Key: A

Question #8 – You are given a scatter plot with a binary classification: points are labeled with a plus (+) or minus (-) sign.



You want to use K-nearest neighbors (KNN) to create a model to classify new points. You use cross-validation to find the best value of K and ensure the model's generalizability. Which of the following options is the best approach for this situation?

A. Use the entire dataset to train the KNN model with K = 1 and predict new data points.

B. Divide the dataset into a training set and a test set, use K = 3 for the KNN model, and evaluate its performance on the test set.
C. Use 10-fold cross-validation to find the best K value for the KNN model and evaluate its average performance over the folds.
D. Pick the K value that classifies the training data with 100% accuracy and use it to predict new data points.

Answer Key: C

Question #9 – What does the Out-Of-Bag (OOB) score represent in a Random Forest model
A. The accuracy of the model on the training dataset.
B. The accuracy of the model on a separate test dataset.
C. The prediction error rate of the model on the bootstrapped sample.
D. The prediction error rate of the model on the data points that were not included in the bootstrap sample.

Answer Key: D

Question #10 – What is the principle behind clustering-based methods for outlier detection?
A. Assigning data points to predefined categories based on their characteristics
B. Identifying data points that conform to the general distribution of a dataset
C. Finding outliers by locating data points that do not belong to any cluster or are far from cluster centroids
D. Aggregating data points based on their temporal sequence

Answer Key: C