**Module 4 Quiz**

Question 1

Which of the following is an example of clustering?

Separate the data into distinct groups by similarity

Question 2

Which of the following are advantages to using decision trees over other models?

Trees often require less pre-processing of data

Trees are easy to interpret and visualize

Question 3

What is the main reason that each tree of a random forest only looks at a random subset of the features when building each node?

To improve generalization by reducing correlation among the trees and making the model more robust to bias.

Question 4

Which of the following supervised machine learning methods are greatly affected by feature scaling?

KNN

Decision Trees

Support Vector Machines

Question 5

Select which of the following statements are true.

For a model that won’t overfit a training set, Naive Bayes would be a better choice than a decision tree.

For predicting future sales of a clothing line, Linear regression would be a better choice than a decision tree regressor.

Question 6

Match each of the prediction probabilities decision boundaries visualized below with the model that created them.

A picture containing text

Description automatically generated

1. Neural Network
2. KNN (k=1)
3. Decision Tree

Question 7

A decision tree of depth 2 is visualized below. Using the `value` attribute of each leaf, find the accuracy score for the tree of depth 2 and the accuracy score for a tree of depth 1.

Diagram

Description automatically generated

What is the improvement in accuracy between the model of depth 1 and the model of depth 2? (i.e. accuracy2 - accuracy1)

0.067

Question 8

For the autograded assignment in this module, you will create a classifier to predict whether a given blight ticket will be paid on time (See the module 4 assignment notebook for a more detailed description). Which of the following features should be removed from the training of the model to prevent data leakage?

compliance\_detail - More information on why each ticket was marked compliant or non-compliant

collection\_status - Flag for payments in collections

Question 9

Which of the following might be good ways to help prevent a data leakage situation?

If time is a factor, remove any data related to the event of interest that doesn’t take place prior to the event.

Remove variables that a model in production wouldn’t have access to

Sanity checks the model with an unseen validation set

Question 10

Given the neural network below, find the correct outputs for the given values of x1 and x2.

The neurons that are shaded have an activation threshold, e.g. the neuron with *>1?* will be activated and output 1 if the input is greater than 1 and will output 0 otherwise.

Diagram

Description automatically generated

|  |  |  |
| --- | --- | --- |
| x1 | x2 | output |
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |