Module 5 Quiz

1. What impact does the depth of a decision tree model have on its complexity and likelihood to underfit or overfit?

1. Shallow decision trees are simpler models and may underfit data, while very deep decision trees are complex models and may overfit data.
2. The depth of a decision tree does not influence the model's complexity or fit
3. Deep decision trees are simpler models that commonly underfit data, while shallow decision trees are more complex and can result in overfitting
4. Regardless of depth, all decision tree models are simple and very likely to underfit

2. When used for a regression task such as predicting demand for a product, how does a decision tree make its

1. For each new observation, we look at the training data and find the most similar point, and then use the corresponding target label as the prediction for the new observation
2. We trace the route through the tree for each new observation until we reach a leaf. We then use the majority vote of the class of each point at the leaf as the prediction
3. For each new observation, it traces the route through the tree until it reaches a leaf. It then uses the mean target value of the training points at that leaf as the prediction.
4. For each new observation, we re-train the tree with the data including the observation, and then use the mean target value of the points at the leaf as the prediction.

3. Which of the following are correct statements about decision trees (select all that apply)? Single decision trees are highly interpretable

- Single decision trees are highly interpretable
- Decision trees do not handle non-linear relationships well
- Decision trees are prone to overfitting on the training data
- The fit of a decision tree is highly influenced by the hyperparameters, specifically those which control the tree's depth

4. What is the goal of using ensemble models rather than individual models?

1. Combining multiple models in an ensemble makes the aggregate model less likely to overfit and better at generalizing to new data
2. Ensemble models usually have better prediction performance on the training dataset
3. Ensemble models reduce the computational requirements to train and run the model
4. Ensemble models are more interpretable

5. When we create an ensemble model, what decisions do we need to make (select all that apply)?

- How we would like to aggregate the predictions of each individual member model to form the prediction for the ensemble (for example, using majority voting, simple average, or weighted average)

- Whether each individual member model is trained on the full dataset or different slices of the data
- How many individual member models we would like to create in the ensemble
- Whether we use the same algorithm or different algorithms for each member model

6. When we use tree models, why do we often choose to use a Random Forest ensemble model rather than a single decision tree?

1. Forests always generate better predictions than single decision trees on new data
2. Using a Random Forest reduces the risk of overfitting relative to a single decision tree
3. We can use Random Forest models for both regression and classification, while we can only use decision models for classification
4. Random Forest models require less computational power to train

7. What is a key benefit of tree-based models (decision trees or Random Forests) relative to linear models (linear or logistic regression)?

1. Decision tree and Random Forest models are easier to interpret than linear models
2. They can easily model complex, non-linear relationships between the input features and targets with no need for additional feature creation / feature engineering work
3. They always yield better performance than linear models do in generalizing to make predictions on new data
4. Tree-based models are less likely to overfit than linear models

8. For which of the below applications would we likely use a clustering approach (select all that apply)?

- Modelling and predicting flight delays using historical flight data
- Segmenting the potential customer base for a new product into groups for the purpose of developing targeted advertising campaigns for each
- Organizing food items into groups based on nutritional content
- Training a model for use in a car to autonomously parallel park

9. What is the most important (and typically first) decision to make when applying clustering to a problem?

1. Determining what basis we will use for measuring similarity / dissimilarity between datapoints (how we will measure which points are similar and which are not)
2. Determining which clustering algorithm we should apply
3. Which metric we will use to evaluate the quality of the clusters formed
4. How we will split our data between training and test sets

10. Which of the below are true about K-Means clustering (select all that apply)?

- It is easy to implement and generally converges quickly
- It is the most popular clustering technique
- It does not require the user to provide a specific number of clusters to use in the algorithm
- It forms linear decision boundaries between the data when separating into clusters