1 point

1. Which of the following is an example of clustering?

- (●) Separate the data into distinct groups by similarity
- ( ) Creating a new representation of the data with fewer features
- ( ) Accumulate data into groups based on labels
- ( ) Compress elongated clouds of data into more spherical representations

---

1 point

2. Which of the following are advantages to using decision trees over other models? (Select all that apply)

- [ ] Trees are naturally resistant to overfitting
- [ ] Decision trees can learn complex statistical models using a variety of kernel functions
- [■] Trees are easy to interpret and visualize
- [■] Trees often require less preprocessing of data

---

1 point

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 increase interpretability of the model
- ( ) To learn which features are not strong predictors
- ( ) To reduce the computational complexity associated with training each of the trees needed for the random forest.
- (●) To improve generalization by reducing correlation among the trees and making the model more robust to bias.

---

1 point

4. Which of the following supervised machine learning methods are greatly affected by feature scaling? (Select all that apply)

- [ ] Decision Trees
- [■] Support Vector Machines
- [■] Neural Networks
- [ ] Naive Bayes
- [■] KNN

1 point

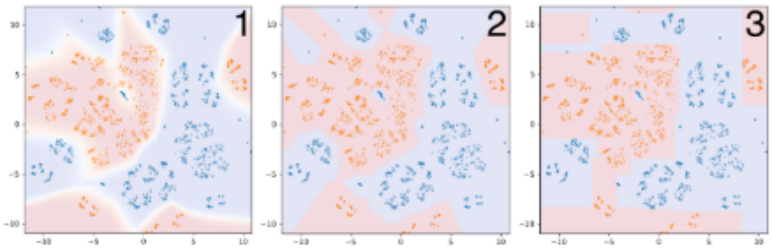
5. Select which of the following statements are true.

- [ ] For having an audience interpret the fitted model, a **support vector machine** would be a better choice than a **decision tree**.
- [x] For a model that won't overfit a training set, **Naive Bayes** would be a better choice than a **decision tree**.
- [ ] For a fitted model that doesn't take up a lot of memory, **KNN** would be a better choice than **logistic regression.**
- [x] For predicting future sales of a clothing line, **Linear regression** would be a better choice than a **decision tree regressor**.

---

1 point

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



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

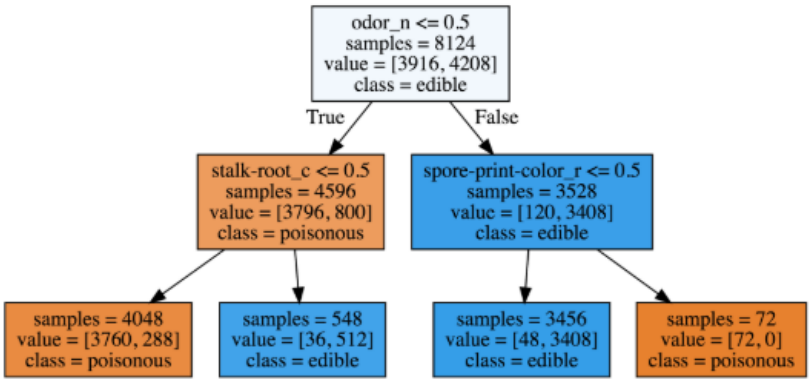
- ( ) 1. KNN (k=1)
  2. Neural Network
  3. Decision Tree

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

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

1 point

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.



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

0.067

---

1 point

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? (Select all that apply)

- [x] collection_status - Flag for payments in collections
- [ ] agency_name - Agency that issued the ticket
- [ ] grafitti_status - Flag for graffiti violations
- [ ] ticket_issued_date - Date and time the ticket was issued
- [x] compliance_detail - More information on why each ticket was marked compliant or non-compliant

1 point

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

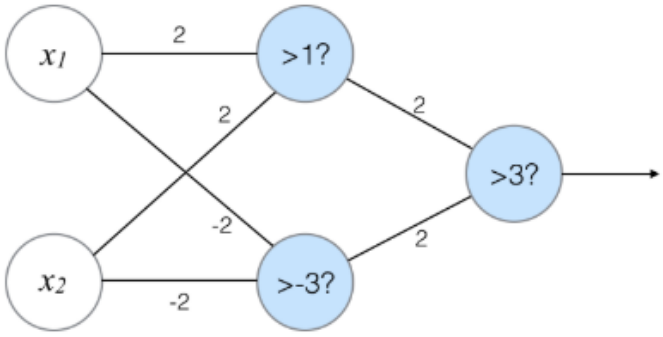
- [x] If time is a factor, remove any data related to the event of interest that doesn't take place prior to the event.
- [ ] Ensure that data is preprocessed outside of any cross validation folds.
- [x] Remove variables that a model in production wouldn't have access to
- [x] Sanity check the model with an unseen validation set

---

1 point

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.



x1	x2	output
0	0	0
0	1	0
1	0	0
1	1	1

x1	x2	output
0	0	0
0	1	1
1	0	1
1	1	0