DS325 Applied Data Science Exam 2 name:_____
Spring 2025, Professor Roth

sec:_____

I - Multiple Choice (30 pts)

1. In performing a logistic regression, you notice you have many highly correlated features. Which of the following could resolve the issue?
   a. use lasso regression
   b. apply PCA to your features
   c. scale and encode your features
   d. a and b
   e. all of the above

2. Which of the following are differences between Random Forests and Gradient Boosted Trees?
   a. The trees in a random forest are trained in parallel; the trees in a gradient boosted tree are trained sequentially.
   b. Random forests have low variance; gradient boosted trees have high variance.
   c. Random forests use bagging; gradient boosted trees focus on previously misclassified samples
   d. b and c
   e. all of the above

3. In the context of a decision tree, the homogeneity of the samples at a given node is described by:
   a. R^2
   b. leaf variance
   c. gini impurity
   d. class
   e. none of the above

4. Which of the following is **not** a classification problem?
   a. predicting whether a Spotify customer will like or dislike a particular song based on other songs saved to their playlists
   b. estimating the likelihood of rain (as a percentage) based on humidity and temperature
   c. predicting which of streaming service's customers are likely to cancel their subscription in the coming month
   d. using factors like income and past credit history to calculate whether a loan will default
   e. more than one of the above

5. Which of the following is an example of an ***unsupervised*** learning algorithm?
   a. PCA
   b. K-Nearest Neighbors
   c. K-Means
   d. Random Forest
   e. more than one of the above

6. In K-Nearest Neighbors, what does K refer to?
   a. the number of nearby data points to poll in voting.
   b. the radius around the unlabeled point that defines the neighborhood.
   c. the number of votes a given category has to receive to win.
   d. the number of different candidate categories a point can be assigned to.
   e. none of the above

DS325 Applied Data Science Exam 1
Spring 2025, Professor Roth

7. In K-Means, what does K refer to?
   a. the number of nearby data points to poll in voting.
   b. the radius around the unlabeled point that defines the neighborhood.
   c. the number of votes a given category has to receive to win.
   d. the number of different candidate categories a point can be assigned to.
   e. none of the above

8. In K-fold Cross Validation, what does K refer to?
   a. the number of samples in the validation set.
   b. the number of times a candidate model is fit and validated.
   c. the number of partitions the training set is divided into.
   d. a and b
   e. b and c

9. Which of the following methods require numerical data to be scaled:
   a. random forests
   b. gradient boosted trees
   c. decision trees
   d. all of the above
   e. none of the above

10. Which of the following require categorical data to be encoded:
    a. random forest
    b. logistic regression
    c. K-nearest neighbors
    d. all of the above
    e. none of the above

11. Which of the following are examples of ordinal data?
    a. 'morning', 'afternoon', 'evening', 'night'
    b. *'rarely', 'sometimes', 'often', 'always'*
    c. *'socks', 'pants', 'shirt', 'hat'*
    d. two of the above
    e. all of the above

12. Which of the following would likely lead to over-fitting?
    a. small value for K (n_neighbors) in K-nearest neighbors
    b. small max_depth in a decision tree
    c. large n_estimators (number trees) in a random forest
    d. large k in K-fold cross validation
    e. none of the above

13. Which of the following is FALSE about PCA?
    a. the first principle component represents the direction in which the data varies most
    b. even if features are co-linear, principle components will not be
    c. you can create as many principle components as there are features
    d. you can often approximate your data using fewer principle components than original features
    e. none of the above

14. You are optimizing a decision tree using GridSearchCV with ***5-fold cross validation***. You choose the following candidate values for your hyper-parameters:
    i. max_depth = [2, 4, 6]
    ii. min_samples_split = [5, 10]

    How many times will this search algorithm fit a model?
    a. 5
    b. 6
    c. 25
    d. 30
    e. 32

15. If a decision tree is over-fitting, which of the following might fix the issue?
    a. increasing the max_depth
    b. using a larger training data set
    c. using more features
    d. applying Lasso regularization
    e. none of the above

II - Very short answer (10 pts)

16. In K-means clustering, which metric is used to choose the correct number of categories?
17. Which encoder creates new columns for each different value of a feature?
18. What is the name of the method for random selection with replacement (used in Random Forests)?
19. For any classifier, which metric represents the percentage of correct classifications across all tested data?
20. What is the name for the top-most node in a decision tree? What is the name for a terminal node where a decision is made?

III - Short Answer (20 pts)

21. Below is a dataframe for the pets available at an animal rescue. How would you encode the data?
    a. For each kind of encoder you'll use, list the corresponding features.
    b. Fill out the blank dataframe with the encoded values. Give any new columns easily interpretable names.

	names	type	weight	age	good_with_children	activity	special_needs
0	Bink	Dog	20	Puppy/Kitten	Yes	Very Active	Fenced Yard
1	Rascal	Dog	15	Adult	Yes	Lazy	Medicine
2	Purrburger	Cat	15	Senior	Yes	Active	None
3	Sage	Dog	50	Adult	No	Very Active	Fenced Yard
4	Cleo	Dog	25	Puppy/Kitten	No	Normal	None
5	Maru	Cat	5	Puppy/Kitten	Yes	Normal	Medicine

DS325 Applied Data Science Exam 1
Spring 2025, Professor Roth

22. At an apple processing facility, apples are scored on several categories and a classifier determines whether each apple is 'good quality' (good for eating) or 'not good' (used for juice, baby food, baked goods, etc).

    Below are the first five rows of a dataframe and a decision tree classifier trained on a sample of 100 apples.

    a. Which is the most informative feature in determining the quality of apples?
    b. Fill out the values of the confusion matrix according to this decision tree.
    c. Which metric would be most important for this classification? Explain very briefly.
    d. Calculate the value of the metric you chose?

	type	weight	color	firmness	blemishes	good_quality
0	G	87	10	4	1	True
1	S	67	10	3	0	True
2	F	77	4	4	0	True
3	G	102	10	3	0	False
4	S	107	8	5	2	False