Name: _____

In-class final, EC524/424

100 points possible

Instructions A short response or derivation suffices for most of the following questions.

Do not write more than necessary. Excessively long responses will be penalized.

1. (4 points) Describe the bias-variance tradeoff.

2. (4 points) What is a confusion matrix? **Draw** an example for a binary classification task (numbers are not necessary) and **explain** how we use it to evaluate a classifier.

3. (4 points) Explain the concept of cross-validation.

4. (4 points) What are **two** key differences between causal inference and prediction?

5. (4 points) Explain the difference between supervised and unsupervised learning.

6. (4 points) **Define** *overfitting* and **explain** how can you prevent it.

7. (4 points) What is the difference between precision and sensitivity?

8. (4 points) How does a decision tree work?

9. (4 points) How does a random forest “extend” the concept of a decision tree?

10. (4 points) Explain the concept of ensemble learning.

11. (4 points) How do you handle missing data in a dataset?

12. (4 points) What is the difference between L1 and L2 regularization (penalization)?

13. (4 points) How does a support vector machine (SVM) work?

14. (4 points) What is the difference between bagging and boosting?

15. (4 points) Explain the ROC curve: **what** is and what does it **tell us**?

16. (4 points) Explain the concept of k-nearest neighbors (KNN).

17. (4 points) Explain why *accuracy* is not always the best metric to evaluate a classifier.

18. (4 points) Explain why some models require data standardization while others do not.

19. (4 points) Explain why we generally prefer *k*-fold cross-validation over the validation-set approach.

20. (4 points) Why might we prefer logistic regression over linear regression for classification tasks?

21. (4 points) What does it mean to *tune* a model?

22. (4 points) Give **two** techniques for variable selection and briefly **describe** how they work/differ.

23. (4 points) Explain why we use Gini or entropy as splitting criteria in decision trees.

24. (4 points) Describe how increasing a model's flexibility affects its bias **and** variance.

25. (4 points) You have a dataset with one numeric outcome y and two numeric predictors x_1 and x_2. Each variable takes values between 0 and 100.

    You train a decision tree that makes the following splits:

    1. $x_1 < 50$

       (a) $x_2 < 25$

    2. $x_1 \geq 50$

       (a) $x_2 > 60$

           i. $x_1 < 75$

    Draw either

    1. the implied **decision tree** or
    2. the implied decision boundaries in the predictor space.

    Bonus points (4 points) for correctly drawing both.