Name: ____

In-class final, EC524/424

120 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) In the context of the *bias-variance tradeoff*, define what we mean by *bias*.

2. (4 points) In the context of the *bias-variance tradeoff*, define what we mean by *variance*.

3. (4 points) What does it mean to *tune* a model?

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

5. (4 points) Why do we prune decision trees?

6. (4 points) When we talk about more or less *parametric* models, what do we mean by *parametric*? Provide an example of a more parametric model and a less parametric model.

7. (4 points) Explain the concept of *cross-validation*.

8. (4 points) How does cross-validation help us trade between bias and variance?

9. (4 points) Explain the difference between supervised and unsupervised learning.

10. (4 points) Why do we normalize our predictors before fitting a penalized regression model?

11. (4 points) What is the difference between precision and sensitivity?

12. (4 points) Explain the concept of ensemble learning.

13. (4 points) Why do classification decision trees split with Gini or entropy rather than accuracy?

14. (4 points) How do you handle missing data in a dataset without dropping observations?

15. (4 points) What is the difference between L1 and L2 regularization (penalization)?

16. (4 points) How does a support vector machine (SVM) work?

17. (4 points) Describe (generally) how a decision tree is built.

18. (4 points) How do random forests decorrelate trees?

19. (4 points) Explain the ROC curve. Specifically: What is it and what information does it provide?

20. (4 points) Define the null classifier and explain why it is useful.

21. (4 points) Explain how the k-nearest neighbors (KNN) algorithm makes predictions.

22. (4 points) Explain why *accuracy* is not always the best metric to evaluate a classifier.

23. (4 points) Explain why we generally prefer *k*-fold cross-validation over the validation-set approach.

24. (4 points) Draw a confusion matrix and explain what it tells us.

25. (4 points) Explain how lasso regression differs from and relates to “standard” OLS regression.

26. (4 points) Are we typically “okay” with biased predictions in machine learning? Explain why/why not.

27. (4 points) Describe how increasing a model’s flexibility affects its bias **and** variance.

28. (4 points) Explain in which settings one might prefer linear-regression model over a random-forest model.

29. (4 points) How does bootstrapping help bagged trees?

30. (4 points) You have a dataset with one numeric outcome $\mathtt{y}$ and two numeric predictors $\mathtt{x}_1$ and $\mathtt{x}_2$. Each variable has values between 0 and 100.

You train a decision tree that makes the following splits:

1. $\mathtt{x}_1 < 50$
   (a) $\mathtt{x}_2 < 25$
2. $\mathtt{x}_1 \geq 50$
   (a) $\mathtt{x}_2 > 60$
      i. $\mathtt{x}_1 < 75$

Draw the implied *decision boundaries* in the predictor space.

Important: Make sure you label the axes.