Questions

1. Write yes

**yes**

2. How many particles labeled "synecho" are in the file provided?

**18146**

3. What is the 3rd Quantile of the field fsc\_small? (the summary function computes this on your behalf)

**39184**

4. What is the mean of the variable "time" for your training set?

Given that the mean value depends on your random sample, the assignment accepts any value within a 99% confidence interval. If your answer is not accepted, try to resample. If your submission is chronically not accepted, check your program.

**341.1**

5. In the plot of pe vs. chl\_small, the particles labeled ultra should appear to be somewhat "mixed"

with two other populations of particles. Which two populations?

+ synecho

+ **pico**

+ crypto

+ **nano**

6. Use print(model) to inspect your tree. Which populations, if any, is your tree incapable of

recognizing? (Which populations do not appear on any branch?)

(It's possible, but very unlikely, that an incorrect answer to this question is the result of improbable

sampling.)

Hint: Look

+ pico

+ nano

+ ultra

**+ crypto**

+ synecho

7. Most trees will include a node near the root that applies a rule to the pe field, where particles with

a value less than some threshold will descend down one branch, and particles with a value greater than some threshold will descend down a different branch.

If you look at the plot you created previously, you can verify that the threshold used in the tree is evident visually.

What is the value of the threshold on the pe field learned in your model?

**5004**

8. Based on your decision tree, which variables appear to be most important in predicting the class population?

+ fsc\_big

+ **chl\_small**

+ **pe**

+ fsc\_perp

+ fsc\_small

+ chl\_big

9. How accurate was your decision tree on the test data? Enter a number between 0 and 1.

**0.85**

10. What was the accuracy of your random forest model on the test data? Enter a number between 0 and 1.

**0.921**

11. After calling importance(model), you should be able to determine which variables appear to be most important in terms of the gini impurity measure. Which ones are they?

+ **chl\_small**

+ fsc\_perp

+ **pe**

+ fsc\_big

+ fsc\_small

+ chl\_big

12. What was the accuracy of your support vector machine model on the test data? Enter a number between 0 and 1.

**0.921**

13. Construct a confusion matrix for each of the three methods using the table function. What appears to be the most common error the models make?

+ **pico is mistaken for ultra**

+ nano is mistaken for ultra

+ synecho is mistaken for pico

+ ultra is mistaken for pico

+ crypto is mistaken for ultra

+ ultra is mistaken for nano

14. After removing data associated with file\_id 208, what was the effect on the accuracy of your svm model?

Enter a positive or negative number representing the net change in accuracy, where a positive number represents an improvement in accuracy and a negative number represents a decrease in accuracy.

**-0.0083118**

15. The variables in the dataset were assumed to be continuous, but one of them takes on only a few discrete values, suggesting a problem. Which variable exhibits this problem?

+ pe

+ fsc\_small

+ chl\_big

+ chl\_small

+ fsc\_perp

+ **fsc\_big**