Mock Interview Questions

1. Tell me a little about yourself. (ask everyone)
2. Can you tell me about a data science problem you’ve worked on? (ask everyone)
3. Technical Questions (choose 2-3 per person)
   1. What is the difference between AI, machine learning and data science?
   2. Can you explain the difference between Bag of Words and TF-IDF?
   3. What is the difference between supervised & unsupervised learning? Give examples of each.
   4. Name two supervised machine learning models you are familiar with. Give examples of when you’ve used each.
   5. What is deep learning?
   6. When might you want to use a random forest model versus a decision tree?
   7. What is a false positive? What is a false negative? Give an example when a false negative is worse than a false positive.
   8. Why would we want to use cross validation?
   9. Say I am building a model to predict the salary of someone based on factors such as education level and industry. What error metric would be appropriate for this model and why?
   10. What are some ways you can handle missing data?
   11. Say I want to predict whether or not someone will graduate college based on factors such as their high school gpa and their major. Name two machine learning models that may be appropriate for this problem.
   12. Draw a confusion matrix and explain each of the pieces.
4. Python white board (choose 1 per person - answers will vary)
   1. Imagine I have a dataset that contains two columns – height and weight. Use pseudocode to do the following: 1) check for nulls 2) remove rows with nulls, 3) filter the data to only return the rows where the weight is over 100 lbs.
   2. Imagine I have a dataset that contains ten columns. Use pseudocode to show how you could plot each column in X.
   3. Use pseudocode to show how you could try to determine the best number of splits (i.e., depth) in a decision tree.
   4. Say we have data saved in 5 csv files. Each csv file contains one column of data. Use pseudocode to show how you could read in the data in each of those files and combine it into one data structure.
   5. Imagine X is a dataset that contains two columns – the heights and weights of 100 people. Using pseudocode, calculate the BMI for each person and add that to a new column in the data frame. Note that the formula for BMI is weight/height^2.
5. SQL (choose 1 per person)
   1. What is an alias in SQL? Why might you use an alias?
   2. What SQL commands would allow you to only select dates from 2020-01-01 to 2020-12-31?
   3. What type of join only returns rows that have matching values in both tables involved in the join?
   4. What wildcard symbol matches zero or more characters?
   5. Name one aggregate function in SQL.