1. What is the difference between Data-Driven Decision Making, Data Science, and Data Engineering?
2. How Data Science and Big Data are related?
3. One of the fundamental principles of data science is that data, and the capability to extract useful knowledge from data, should be regarded as key strategic assets. Why?
4. What is the difference between supervised and unsupervised learning?
5. In the following problem descriptions, identify what type of methods should be used (e.g., supervised or unsupervised) and why.
   a. Several photos are available with elements that are on them (e.g., a person, a car, a street sign, etc). You need to create a model to recognize these elements in unseen photos.
   b. Several photos of people are available and attributes from it are extracted using some pre-processing technique. You need to create a model to place people with similar characteristics in the same group.
   c. Based on some prior knowledge about the weather (e.g., when it’s sunny, the temperature is higher; when it’s cloudy, humidity is higher, etc.) and event attendance, you need to create a model to determine when more people will attend to a future event.
   d. Examples of animals' characteristics of several birds, cats and dogs are available in your dataset. There’s no information regarding what is the animal you’re analyzing. You need to build a model to place each animal example in the same group.
6. What is the difference between classification and regression?
7. Explain with your own words all the six steps of CRISP-DM.
8. Given the following attributes, which one would be chosen by a Decision tree classifier? Why?

9. Given the following entropy graphs, which attribute is more informative? Why?
10. How do decision trees algorithms use measures like information gain to build a predictor?
11. The following figure shows decision boundaries from a decision tree classifier. Build its tree with its probabilities.

![Decision Tree Diagram](image)

12. What are the differences between supervised segmentation and parametric modeling strategies?
13. You have a problem with the following two variables. Which model would be a better choice in terms of complexity (e.g., number of decision boundaries), a decision tree or a linear classifier?

![Variable Scatter Plot](image)

14. SVMs, if properly configured, tend to generalize a problem better than a Logistic regression. Why?
15. What are ensembles and why they tend to improve a classification metric when its result is compared with a single weak model?
16. What is the difference between bagging and boosting?
17. Explain in general lines how AdaBoost and Random Forest works. Compare the two strategies regarding advantages and disadvantages.

18. Given the following regression models, what models are underfitting, overfitting and robust? Why?

![Regression Models](image)

19. What is the motivation behind the use of robust strategies for model evaluation such as cross-validation?

20. What are learning curves and fitting graphs? How can they be used to control overfitting?

21. Explain in a few lines what is the task solved by clustering techniques.


23. You are supposed to use the DBScan clustering algorithm using an eps = 1 and minPts = 5. Consider the euclidean distance and show the final result only.

![DBScan Clustering](image)

24. What would be the result if the parameters are eps = 0.5 and minPts = 3?