CSCI-4146 --- Process of Data Science
Exercises list

1. Using your own words, characterize the process of data science.
2. Explain all the four steps of the process of data science.
3. Characterize the meaning of each of the four Vs of Big Data?
4. Why can the data collection step be the most time-consuming in the process of data science?
5. What is the difference between raw and pristine data and why it is essential to keep a record of the operations done in the raw data?
6. Why is it important to handle missing data? Explain at least three different ways to handle missing data and compare them concerning advantages and disadvantages.
7. Why is Python considered an exciting programming language choice to apply a data science process?
8. What are data descriptives and why its analysis is important?
9. What is a probability mass function and how can they be used to compare two different variables?
10. Why creating visualizations is important? Where can they be used?
11. A good visualization must be expressive and effective. Explain this sentence.
12. Characterize the four ways to plot a 3D space in a 2D space. Compare them concerning advantages and disadvantages.
13. Explain in general lines the steps of the workflow for analyzing a statistical model.
14. What is the difference between a beta-hat and a y-hat oriented analysis?
15. What is the difference between the Bayesian and frequentist schools?
16. Explain at least two ways to test if a distribution is normal. Detail what is the output of these tests.
17. What is machine learning? What is the difference between supervised and unsupervised learning?
18. What is feature selection task and why this task should be considered in a supervised learning task?
19. Explain in general lines how decision trees, naive Bayes, KNNs and SVMs work. Compare these algorithms regarding advantages and disadvantages.
20. What are ensembles and why they tend to improve a classification metric when compared its result is compared with a single model?
21. What is the difference between bagging and boosting?
22. Explain in general lines how AdaBoost and Random Forest works. Compare the two strategies regarding advantages and disadvantages.
23. Explain in few lines what is the task solved by clustering techniques.
25. What is dimensionality reduction and why to use such techniques?