Lesson 1

Create a document on https://writersintegrity.cs.miu.edu/, name it, "Lesson 1", and answer the following questions **briefly**.

1. How would you define Machine Learning?
2. Can you name four types of problems where it shines?
3. What is a labeled training set?
4. What are the two most common supervised tasks?
5. Can you name four common unsupervised tasks?
6. What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?
7. What type of algorithm would you use to segment your customers into multiple groups?
8. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?
9. What is an online learning system?
10. What is out-of-core learning?
11. What type of learning algorithm relies on a similarity measure to make predictions?
12. What is the difference between a model parameter and a learning algorithm's hyperparameter?
13. What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?
14. Can you name four of the main challenges in Machine Learning?
15. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?
16. What is a test set and why would you want to use it?
17. What is the purpose of a validation set?
18. What can go wrong if you tune hyperparameters using the test set?
19. What is repeated cross-validation and why would you prefer it to using a single validation set?

Interview Questions

- **Scenario**: You are working on a dataset that includes patient records in a healthcare database. After initial analysis, you identify that 15% of the patient age data is missing. The dataset is fairly large, with over 100,000 records. The age data is important for your analysis to track disease prevalence across age groups.
- **Question 1**: Describe how you would handle these missing values. Discuss the techniques you would consider and explain your choice. How would your approach change if missing values were identified in a more critical feature, such as diagnosis information?

- **Scenario**: You are tasked with cleaning a financial dataset used for forecasting stock prices. The dataset contains some apparent outliers in the volume of trades, which could potentially skew the predictive models. Preliminary analysis shows that these outliers represent days with significant market news.

- **Question 2**: What methods would you use to detect these outliers, and how would you decide whether to remove or adjust these data points in the dataset? Outline the potential impacts of your decision on the forecast model's performance.
- **Scenario**: Imagine you are preparing a dataset for a machine learning model that predicts real estate prices. The dataset features have varying scales and distributions, including property size in square feet and local crime rate per 1,000 residents.
- **Question 4**: Would you choose to normalize or standardize these features, and why? Provide a detailed explanation of how each process would affect the data and the model's learning process. What might be the implications of choosing one method over the other in terms of model performance and accuracy?

*** Read Lesson 2 and summarize it.