**Quiz 1: Machine Learning Paradigms**

**Questions:**

1. Supervised Learning:
   1. Define supervised Learning and provide an example.
      * Definition: A subcategory of machine learning that uses datasets with both inputs and outputs to train the algorithm.
      * Example: detecting spam messages or emails: to train the algorithm by giving it examples of messages that were confirmed by an expert as spam messages, and messages that were confirmed as safe messages.
2. Unsupervised Learning:
   1. Describe unsupervised learning and mention common algorithms used in this paradigm.
      * Description: A subcategory of machine learning that uses datasets only with inputs. The goal is to find patterns and relationships for these inputs.
      * Common Algorithms:   
        K-Means Clustering: is used to divide the data into clusters where each point belongs to the cluster with the nearest centroid.  
        Another method is called Hierarchical Clustering.
3. Reinforcement Learning:
   1. What is the objective of reinforcement learning? Name and briefly describe one of the challenges in training reinforcement learning models.
      * Objective: After getting assigned to a goal to reach, it starts to learn how to achieve that goal by trial-and-error approach.
      * Challenges: One famous challenge is that you need an immense amount of interaction with the environment. This can be costly and time-consuming.
   2. Define the terms "agent", "environment", "reward", and "policy" in the context of reinforcement learning.
      * Agent: the entity or decision-maker that interacts with the environment.
      * Environment: the external system that the agent has to deal with to gain some reward from it.
      * Reward: a signal that the agent receives from the environment after taking an action.
      * Policy: a strategy that determines the agent's actions to take in each state.
4. Semi-Supervised Learning:
   1. Explain semi-supervised learning and why it can be beneficial over supervised and unsupervised learning in certain scenarios.
      * Explanation: it is a machine learning paradigm that uses datasets of both inputs and outputs, but the number of outputs is way less than the number of inputs. The goal is to achieve the best performance and results by using these limited outputs with the help of the large number of inputs.
      * Benefits:  
        - Main advantage is that it requires less labeled data compared to supervised learning, which is very important with models that are very difficult in terms of colleting labeled data.  
        - Having unlabeled data helps models reducing overfitting and improve performance.  
        - The fact that collecting labeled data is not always easy and available, so in this method it is not a problem to work with less amount of labeled data.
   2. Given a large dataset with only a small portion of labeled data, describe a strategy that could be used to train a semi-supervised model.
      * The first step is the semi-supervised classification step, which is designing a basic classification model on the few labeled data instances.
      * The next step is that the model operates unsupervised on the remaining unlabeled data instances and assigns them to the classes from the first step.
5. Evaluation Metrics:
   1. Discuss why accuracy is not always the best for evaluating the performance of a classification model, especially in cases of class imbalance.
      * Accuracy is an important metric for evaluating the performance of classification model but not in all cases.
      * For example, if the distribution of classes in the dataset is highly imbalanced, the model will be very accurate, but at the same time fails when dealing with other classes, which is very common to occur.
   2. For a binary classification problem, define precision and recall. Provide a scenario in which you might prefer one over the other.
      * Definitions:
        + Precision:
          - It measures the accuracy of positive predictions.
          - The number of correctly labeled instances to either the positive or negative class, divided by the total number of positive predictions.
          - It is a measure of the model's exactness.
        + Recall:
          - The number of correctly labeled instances to either the positive or negative class, divided by the total number of instances in this class.
          - It is a measure of the model's completeness.
      * Scenario:
        + Precision is definitely preferred in an environment where the cost of false positives is high.  
          For example, if we are trying to diagnose cancer patients, and we want to avoid telling cancer patients that they are healthy more than telling healthy people that they have cancer.
        + Recall is the opposite. It is preferred in an environment where the cost of false negative is high.  
          In the previous example, our main concern was to avoid telling cancer patients that they are healthy. But Recall is preferred if our main concern is to avoid telling healthy people that they have cancer.