1. What does one mean by the term "machine learning"?

Ans. Machine learning is a kind of artificial intelligence in which computers use algorithms to learn from data, recognise patterns, and make judgements with little human intervention. It entails training models with data in order to enhance their performance over time. The three main types are supervised learning (with labelled data), unsupervised learning (without labelled data), and reinforcement learning (learning through rewards and penalties). Applications include picture and speech recognition, natural language processing, recommendation systems, and predictive analytics.

2.Can you think of 4 distinct types of issues where it shines?

Ans. NLP, Speech Recognition, sentiment analysis, predictive analysis.

3.What is a labeled training set, and how does it work?

Ans. A labeled training set is a dataset used in supervised learning that consists of input data paired with corresponding correct output labels. These labels provide the "ground truth" that the machine learning model uses to learn the relationship between inputs and outputs. Here’s how it works:

1. **Data Collection**
2. **Labeling**
3. **Training**
4. **Model Optimization**
5. **Validation**
6. **Testing**

4.What are the two most important tasks that are supervised?

Ans. Classification and Regression

5.Can you think of four examples of unsupervised tasks?

Ans. The examples of tasks commonly associated with unsupervised learning:

1. Clustering: Grouping similar data points together based on their features without predefined labels.
2. Dimensionality Reduction: Reducing the number of features in a dataset while preserving important information.
3. Anomaly Detection: Identifying data points that deviate significantly from the normal pattern.
4. Association Rule Learning: Discovering interesting relationships or associations between variables in large datasets.

6.State the machine learning model that would be best to make a robot walk through various unfamiliar terrains?

Ans. Reinforcement Learning (RL) is the ideal machine learning paradigm to utilise for teaching a robot to navigate unknown terrains. Advanced RL techniques, such as Deep Reinforcement Learning (Deep RL), which combines reinforcement learning and deep learning, have demonstrated notable success in robotic control tasks.

7.Which algorithm will you use to divide your customers into different groups?  
Ans. To divide customers into different groups, clustering algorithms are typically used in unsupervised learning. One of the most commonly used and effective algorithms for this task is K-Means Clustering.

8.Will you consider the problem of spam detection to be a supervised or unsupervised learning problem?

Ans. Spam detection is supervised learning problem.

9.What is the concept of an online learning system?

Ans. An online learning system, in the context of machine learning, refers to a method of learning in which the model continuously updates itself as new data becomes available, often sequentially. This is in contrast to batch learning, which involves training the model on a predefined dataset and then deploying it without any further changes until a fresh batch of data is acquired.

10.What is out-of-core learning, and how does it differ from core learning?

Ans. Out-of-core learning is a machine learning technique that handles datasets that are too huge to fit in a single computer's main memory (RAM). It permits models to be trained on data stored on disc or in external storage rather than loading the complete dataset into memory all at once. This method is especially beneficial in big data cases where standard in-memory processing is not possible owing to resource constraints.

11.What kind of learning algorithm makes predictions using a similarity measure?

Ans. Instance-based learning algorithms, often known as lazy learners, use a similarity metric to create predictions. These algorithms make predictions based on the similarity between fresh data points and points from the training dataset. Here are some significant algorithms.

1. **K-Nearest Neighbors (K-NN)**
2. **K-Nearest Neighbors Regression (K-NNR)**

12.What's the difference between a model parameter and a hyperparameter in a learning algorithm?

Ans. Model Parameters:

* What They Are: These are internal settings or weights that a machine learning model learns from training data.
* Example: In a linear regression y=w0+w1xy = w\_0 + w\_1 xy=w0​+w1​x, w0w\_0w0​ and w1w\_1w1​ are model parameters that the model adjusts to fit the data.

Hyperparameters:

* What They Are: These are external settings that you choose before training begins.
* Example: Number of layers in a neural network, learning rate, or batch size. They influence how the model learns and performs but are not directly learned from the data.

13.What are the criteria that model-based learning algorithms look for? What is the most popular method they use to achieve success? What method do they use to make predictions?

Ans. Model-based learning algorithms aim to create a mathematical model from the training data that can generalize well to make predictions on new, unseen data. These algorithms typically look for several criteria to achieve success:

* The model should accurately capture the patterns and relationships present in the training data. This involves minimizing the difference between the model's predictions and the actual observed values.
* The model should generalize well to new data that it has not seen before. This means the model should not just memorize the training data but should capture underlying patterns that apply to unseen instances.
* In some cases, especially in fields like medicine or finance, interpretability of the model is crucial. It should provide insights into how different features influence the predictions, allowing humans to understand and trust the model's decisions.
* The ability to scale to large datasets efficiently is important, as real-world datasets can be vast and complex.

Popular Methods for Success:

One of the most popular methods that model-based learning algorithms use to achieve success is supervised learning

Method for Making Predictions:

After training, model-based learning algorithms use the learned model to make predictions on new, unseen data. The specific method used for predictions depends on the type of model:

14.Can you name four of the most important Machine Learning challenges?

Ans. Four of the most important challenges in Machine Learning:

1. **Overfitting and Underfitting**
2. Data Quality and Quantity
3. Interpretability and Explainability
4. Computational Resources and Efficiency

15.What happens if the model performs well on the training data but fails to generalize the results to new situations? Can you think of three different options?

Ans. When a machine learning model performs well on training data but fails to generalise to new scenarios (i.e., performs poorly on unseen data), it is usually due to overfitting. Overfitting happens when a model detects noise or specific patterns in training data that do not translate well to fresh, unseen data. Here are three possible approaches of addressing this issue:

1. Regularization
2. Cross-Validation
3. Feature Selection and Engineering

16.What exactly is a test set, and why would you need one?

Ans. In machine learning, a test set is a subset of data that is excluded from the training process and is only used to evaluate the performance of a trained machine learning model. It provides an independent assessment of how well the model generalises to new, previously unknown data. Here's why a test set is important in machine learning.

Purpose of a Test Set:

1. Performance Evaluation
2. Generalization Assessment
3. Model Selection

17.What is a validation set's purpose?

Ans. The validation set serves a crucial purpose in the machine learning workflow, primarily during the training and development phases. Its main purpose is to aid in the selection and tuning of hyperparameters and to provide an unbiased estimate of the model's performance before final evaluation on the test set.

18.What precisely is the train-dev kit, when will you need it, how do you put it to use?

Ans. The term "train-dev kit" does not have a widely recognized definition in the context of machine learning or data science. It seems there might be confusion or a misunderstanding regarding terminology. However, based on common practices and terminology in machine learning, it's possible to infer a related concept.

Potential Interpretation: Train-Validation Set Split

In some cases, people might refer to a train-dev split or train-validation split, which is a specific way of dividing the dataset for training and validation purposes. Here’s how it typically works and when you might use it:

1. Train-Validation Set Split:
   * Definition: This refers to splitting the dataset into two parts: a training set and a validation set (sometimes called a development set or dev set).
   * Purpose: The training set is used to train the machine learning model, while the validation set is used to evaluate the model's performance and to tune hyperparameters.
   * Ratio: Common splits include 80-20% or 70-30% for training-validation, depending on the size of the dataset and the specific requirements of the problem.

When You Need It:

* Model Development: A train-validation split is crucial during the model development phase to ensure that the model generalizes well to new, unseen data.
* Hyperparameter Tuning: It helps in selecting and tuning hyperparameters, such as learning rates, regularization parameters, or model architectures, based on the validation set's performance metrics.
* Avoiding Overfitting: Regularly evaluating the model on the validation set helps monitor for signs of overfitting and allows adjustments to the model's complexity or training process as needed.

How to Put It to Use:

1. Dataset Preparation:
   * Split your dataset into training and validation sets using an appropriate ratio (e.g., 80-20%, 70-30%).
   * Ensure that the split maintains the distribution of classes or labels to avoid biased evaluation.
2. Model Training:
   * Train your machine learning model using the training set. Adjust model parameters and update weights based on the training data to minimize the training error.
3. Validation:
   * Evaluate the trained model on the validation set to assess its performance on unseen data.
   * Compute performance metrics (e.g., accuracy, precision, recall, F1-score) to understand how well the model generalizes and to compare different model configurations or hyperparameter choices.
4. Iterative Process:
   * Iterate between training and validation stages, making adjustments to the model architecture, hyperparameters, or data preprocessing based on the validation set's results.
   * Avoid using the validation set multiple times for tuning to prevent overfitting to the validation data.

19.What could go wrong if you use the test set to tune hyperparameters?

Ans. Using the test set to tune hyperparameters can lead to several issues that compromise the validity and reliability of your machine learning model evaluation:

1. Overfitting to the Test Set:
   * If you use the test set multiple times for evaluating different configurations of hyperparameters, there's a risk of overfitting to the test set. Essentially, you are indirectly fitting your model to the test set by optimizing hyperparameters based on its performance. This undermines the purpose of having a separate test set, which is meant to provide an unbiased estimate of the model's performance on truly unseen data.
2. Biased Performance Estimates:
   * Hyperparameter tuning using the test set can bias your performance estimates. The test set is supposed to simulate real-world, unseen data. If you optimize hyperparameters based on its performance, you may end up with an overly optimistic evaluation of your model's ability to generalize to new instances.
3. Lack of Generalization:
   * Machine learning models are designed to generalize well to new, unseen data. If hyperparameters are tuned using the test set, the model may be overly tuned to specific characteristics of the test set rather than learning general patterns from the training data. This can lead to poor performance on real-world data.
4. Invalid Model Evaluation:
   * The primary purpose of the test set is to provide an independent evaluation of the final model after all model development and hyperparameter tuning are complete. If you use the test set during hyperparameter tuning, you lose this independence and compromise the validity of your final model evaluation.

Best Practices:

To avoid these pitfalls, follow these best practices:

* Split Data Properly: Divide your dataset into three parts: training set, validation set (or development set), and test set. Use the training set for model training, the validation set for hyperparameter tuning, and the test set for final model evaluation.
* Single Use of Test Set: Reserve the test set exclusively for the final evaluation of your model after all hyperparameter tuning and model selection decisions have been made. Do not use it for any form of model development or tuning.
* Cross-Validation: If your dataset is small and you need to perform hyperparameter tuning without compromising the test set's integrity, consider using cross-validation techniques on the training set to estimate the model's performance and tune hyperparameters.