**Question 1: How would you define machine learning?**

Machine learning is a subset of Artificial Intelligence (AI). Artificial Intelligence is basically any technique that enables machines to solve a task in a way humans do. Machine learning is a set of algorithms that allow a computer to learn from examples provided to it without being explicitly programmed. Currently Machine Learning is used for

* E-mail Filtering,
* Sales analysis
* Fraud detection

**Question 2: What are the differences between Supervised and Unsupervised Learning? Specify example 3 algorithms for each of these.**

**Supervised Learning:** It is where you have the input and the output and you use an algorithm to map the function from the input to the output. If we have an accurate mapping function, even if the input changes, our model is capable of predicting the output. **Example:** Classification, Linear Regression, Logistic Regression.

**Unsupervised Learning:** This is where you have the input variable alone. There is no label given here. The goal is to discover hidden patterns in data. This is usually used to find customer segments in a market. **Example:** Clustering, Association.

**Question 3: What are the test and validation set, and why would you want to use them?**

Training dataset is sample data that is used to train a model. A Validation dataset is a data sample which is withheld from training the model. It is used to provide an unbiased evaluation of our trained model and is also used when tuning the model’s hyper parameters. The test dataset is the final data sample that is used to test our model’s skill after tuning.

Before choosing an algorithm, our dataset must be split into 2 or 3 parts. Machine learning algorithms have to be trained first, validated and tested before being deployed to deal with real-world data. This is done so that our model is skilled and accurate with its results.

**Question 4: What are the main pre-processing steps? Explain them in detail. Why we need to prepare our data?**

1. **Gathering data:** The quality and quantity of data that you gather will directly determine how good our predictive model can be.

2. **Preparing the data:** We load our datasets into a suitable place and prepare them for our machine learning algorithms.

3. **Pre-Processing**

* The duplicate values are removed so that no data object has an advantage or bias.
* An Imbalanced dataset is one where the number of instances of a class are significantly higher than another class.
* Eliminating missing values or filling the missing values with mean or median only if it is a continuous variable.
* **Outlier Detection**: The presence of outliers can result in lower predictive performance in our model. Outliers are data that don’t fit in our dataset. The different methods with which we can identify and remove outliers are Standard Deviation, IQR (Inter Quartile Range) Calculation and Isolation Forest.
* **Feature Scaling**: Once we have removed the noise from our data, we need to scale our data. This can be done by Standardization and Normalization. Standardization is usually used when our dataset is a normal distribution. Normalization can be classified into

--Linear Scaling: It is used when our feature is more or less uniformly distributed across a fixed range.

--Clipping: It is used when our feature contains extreme outliers.

--Log scaling: It is used when our feature conforms to the power law.

--Z score: It is used when the feature distribution does not contain extreme outliers.

* Bucketing or binning is used to minimise the effects of small observation errors.
* **Feature Extraction**

-Principle Components Analysis (PCA)

-Independent Component Analysis (ICA)

-Linear Discriminant Analysis (LDA)

-t-distributed Stochastic Neighbor Embedding (t-SNE)

* **Feature encoding :**  This is the process of performing transformations on the data so that it can be easily accepted as input for our machine learning algorithms.
* **Training, Validating and Testing** our model so that it can give its most efficient performance while dealing with real-world data.
* **Cross Validation:**  This method evaluates our model on a limited data set. It splits the data into k parts and hence it is called k-fold cross validation.

**Question 5: How can you explore continuous and discrete variables?**

A variable is classified into two types: Categorical and Numerical. The numerical variable is classified into: Discrete and Continuous Random variable.

A Discrete Random variable takes up a whole single numerical value. A probability distribution over discrete variables can be explained using a probability mass function (PMF).

A Continuous Random variable can take up any value within a particular range. Continuous variables are represented using probability density function.

**Question 6: Analyse the plot.**

Variable type: Continuous Random Variable

Plot type: Probability Density Function, Histogram

It is not a normal distribution and there are minimum and maximum values so the pre-processing scaling technique should be normalization. We can use the Z score technique in normalization since our feature does not contain extreme outliers.