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

Machine Learning is one of the subsets of AI, which focuses on the use of data and algorithms to imitate the way that humans learn and improve accuracy during learning. It enables computers to learn automatically from history of data and find the patterns and relations between them to predict new data.

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

Classification problems such as spam detection in email, regression problems such home price prediction, clustering problem such as partitioning customers to different group based on their interest, time series problem such as forecasting future cost of share in share market.

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

In training dataset if we have a column or columns which present target or ground truth of data are labels and in this case our training dataset is labelled such as cancer classification when we have some photo that their cancer types of them exist and based on these data we want to train models that can predict new photos without labels.

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

Classification when the target column is categorical such as spam/not spam and regression problems when the target column has continuous data type such as price or income.

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

Customer segmentation which is clustering problem, dimensionality reduction for feature selection, anomaly detection in finance and also in recommendation problems and association rules

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

unfamiliar terrains?

Reinforcement learning will be best model, in this method there is am agent that learn from environment and do some action then if the action is correct get reward else get punishment and by rewards and punishment try to improve itself and future actions.

7.Which algorithm will you use to divide your customers into different groups?

If we do not know the group of customers the best option is using cluttering algorithms such as Kmeans, but if we know the groups of customers then it will be supervised problem and with the help of classification algorithm such as decision tree can be solved.

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

It is a supervised problem because, we can recognize which data are spam or not spam and can easily labelled past data.

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

online machine learning is a method of machine learning in which data becomes available in a sequential order and is used to update the best predictor for future data at each step. It is used where it is computationally infeasible to train over the entire dataset and a out-of-core algorithm is required. It is also used in situations where it is necessary for the algorithm to dynamically adapt to new patterns in the data, or when the data itself is generated as a function of time

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

Out-of-core learning is used when a dataset is too large to fit into a computer's memory but can easily fit into some data storage, such as a local hard disk or web repository. The algorithm loads part of the data, runs a training step, then repeats the process until it has run on all the data.

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

Instance-based learning algorithms use a measure of similarity to generalize to new cases. In an instance-based learning system, the algorithm learns the examples by heart, then uses the similarity measure to generalize new instances based on some similarity measure. It builds the hypotheses from the training instances. It is also known as memory-based learning or lazy-learning (because they delay processing until a new instance must be classified).

12.What’s the difference between a model parameter and a hyperparameter in a learning

algorithm?

Models have some parameters that must be determined using the training data set. While there are some adjustable parameters that must be tuned in order to obtain a model with optimal performance. Then during training, we try to find the value which give the best performance such as max\_depth of tree in decision tree algorithm, these are called hyperparameters.

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?

Model-based learning or structure-based learning takes a different approach by constructing models from the training data that can generalize better and create an underlying model from which predictions can be made for new data points. This involves using algorithms like linear regression, logistic regression, random forest, etc. Model-based system learn by minimizing a cost function which is summation of errors and measures how bad the system is at making predictions on new data, plus a penalty for model complexity if the model is regularized.

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

- Not gathering enough data, or sampling noise. Sampling noise means we'll have non-representative data as a result of chance.

- Using a dataset that is not representative of the cases you want to generalize to. This is called sampling bias. For example, if we want to train an algorithm to classify different flowers but all the data are different family of rose flower.

- Dataset is full of missing values, outliers, and noise.

- When model performs well on the training data, but not on test data, and face overfitting problem and model cannot generalize well.

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?

Here the overfitting problem occur, we can collect more data if the sample data is small, using feature selection techniques to remove sum unnecessary columns, try different algorithms or using cross validation techniques to prepare different set of data for algorithms. Hyperparameter tuning also helps to reduce overfitting in some cases.

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

Test data, is a dataset which is unseen for model, and when we want to control how good is the generalizing of our model, we try to predict target value of test dataset and check how much our model predict correctly. If we do not have any test dataset, we can split our data to train and test dataset with for example 80:20 and use that 20% for testing the model. It is important that all the pre-processing step also perform separately on train and test dataset to keep it completely unseen for final model.

17.What is a validation set’s purpose?

Validation dataset is a sample of data separated from training data and is used to give an estimate of model skill while tuning model’s hyperparameters. It is used to give an unbiased estimate of the final tuned model when comparing final models.

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

The Development set is used to select the parameters, tune them and then use them to choose the best model of a training algorithm. Nevertheless, it also helps in avoiding or minimizing overfitting and simultaneously controls the learning rate. It is the quantity and quality of the dataset that determines the picking of the best performance model and its precision. Development sets develop machine learning solutions and help one find the best model of all the different models. It allows one to choose the number of layers (Depth), neurons per layer (width), activation function (ReLU, ELU, etc.), optimizer (SGD, Adam, etc.), learning rate, batch size, and more in the algorithm. It is created by dividing data and most of the time 20% of data selected as development set.

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

If we use test data for hyperparameter tuning, it means, we give the chance of seeing test data to the model them the model can not generalize well and also the calculated performance of model on test data will be high because model may saw them and learn them then can predict their label with higher accuracy.