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

Solving problems with the help of data using different algorithms backed by statistics.

Machine itself will be continuously learning how to solve the particular set of problem with the help of model developed.

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

* Where there is a continuous chance of automation for eg, say, we need output to predict weather on daily basis. It can be automated by developing the model and feeding the data on regular basis.
* To give suggestions for the human for eg, say, we need to know the maintenance status and life of the machine part. To classify the data to two or more classes.
* ChatGPT is the best example of power of Machine Learning. Enabling the user to make decisions/ Learning anything easily/ using algorithms to make day to day work activities easy.
* To improve the performance of anything based on suggestions or analysis using data and Machine Learning.

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

Labelled data set comes with a tag of names associated for the feature. For the raw data adding tags or labels is called Labelling. Labelled training set will have labels and help model to know what it should predict. Labelled data is used for Supervised Machine Learning and unlabelled data is used for Unsupervised Machine Learning.

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

* Regression and Classification.
* There should be past data available for estimating or predicting output.
* Relation between X and y should be established.

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

* Classifying/Categorising the images/objects/customers.
* Irregularity detection.
* Giving Recommendations.
* Clustering DNA / where output feature (y) is not known.

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

* Reinforcement Learning. (Learning through feedback)

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

* Logistic Regression
* Clustering Algorithms

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

Spam detection is a Supervised Machine Learning problem. Dataset should have labels.

1. What is the concept of an online learning system?
   * There two types of Machine Learning Systems. Online and Offline.
   * In online system, the data is trained and model is updated as and when the new data comes in.
   * In offline system, the data is available as a static data set.
2. What is out-of-core learning, and how does it differ from core learning?

Out of core learning means, the data is too large in memory that it cannot fit in local system memory for problem solving. In core machine learning, RAM can solve the entire data set which is fed.

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

Instance based learning algorithms are used where similarity is used to measure.

1. What's the difference between a model parameter and a hyperparameter in a learning algorithm?
   * Model parameters are the one which are calculated by the model itself using the given data while training and building the model.
   * Hyper parameters have to be set manually to obtain higher accuracy or performance.
   * Example, β0, β1 etc are model parameters whereas α (Learning Rate) is the hyper parameter in Linear regression.
2. 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 works on the fact that it should minimize the error or Cost Function.

Also, model looks for the best parameters that can be used to minimize the error.

Based on the best parameters the model has found, when the new data is provided it predicts the value.

1. Can you name four of the most important Machine Learning challenges?
2. Model has to be fit on the data. Thus, the data should be clean. There should be sufficient data available for training the model so that it generalises the condition.
3. Features of the data should be relevant and in good numbers to avoid underfitting.
4. EDA is the most important part since we see any missing values, outlier treatment, distribution of the data, feature engineering are done.
5. Selecting correct features and training the model is also important. Since more feature may cause overfitting. Model Evaluation also play an important part to check how it is performing.
6. 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?

* This is called as Overfitting condition. Bias is less and Variance is more. In this condition model cannot perform well to predict the output feature.
* Overfitting can be avoided by reducing the number of features used for model building.
* Check for Outliers, gather more data if required.
* If the model is not suitable for the given dataset, try to fit another suitable model.

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

Test data is required to check the performance of the model on the data which is not seen by the model to predict the output variable. Data is divided into training and testing set. Training set is used to train and build the model. Testing set is used to validate the performance of the model by checking the predicted value and actual value.

1. What is a validation set's purpose?

Data which is used to fine tune the model by changing the hyperparameters. Out of set of hyperparameters, the best hyperparameter is chosen finally and fed to the testing data set.

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

When we have different model in mind to fit for data, train-dev set helps in training and developing the correct accurate model out of all.

If we have 1000 dataset, then the splitting of Train-Dev-Test can be done by selecting 600-200-200 or 700-200-100. Instead of splitting the data into 3 sections, cross validation can be used so that more data can be used for training purpose.

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

The model will not be able to learn or generalise the data correctly.