1. How would define Machine Learning?

-> Machine Learning is the subset of Artificial intelligence that makes a machine learn from the data making the human tedious work easy.

2. Can you name four types of problems where it shines?

-> Problem for which existing solutions require a lot of fine-tuning or a long list of rules

Complex problems with no known algorithm

Fluctuating environment

Getting insights about complex problems and large amounts of data

3. What is labeled training set?

-> Labeled training set is the set of data that have already defined target and used to train a machine learning model.

4. What are the two most common supervised tasks?

-> Classification

Regression (Predict a target numeric value)

5. Can you name four common unsupervised tasks?

-> Anomaly Detection

Novelty Detection

Association rule learning

Visualization (Clustering)

6. What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?

-> Reinforcement learning or Online learning

7. What type of algorithm would you use to segment your customers into multiple groups?

-> Clustering Algorithm

8. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?

-> Supervised Learning problem

9. What is an online learning system?

-> Online learning system is a system which is an incremental learning which means it learns by itself.

10. What is out-of-core learning?

-> Out-of-core learning means training on a huge dataset that cannot fit in one machine's memory.

11. What type of learning algorithm relies on a similarity measure to make predictions?

-> Instance-based learning

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

-> Model parameters are parameters that are used to fit the model in the training set while the learning algorithm's hyperparameter is used to control the amount of regularization to apply during learning.

13. What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?

-> Model-based learning algorithms search for the model parameter values that minimize the cost function.

The most common strategy they use to succeed is to tune some parameters to fit the model to the training set

They make predictions by specifying a performance measure and selecting the optimal parameter to generalize the model.

14. Can you name four of the main challenges in Machine Learning?

-> Lack of data

Overfitting the training data

Underfitting the training data

Irrelevant features

15. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

-> Overfitting of training data

Possible solution are:

Simplyfying the model

Increase the dataset

Reducing the noise

16. What is a test set and why would you want to use it?

-> The test set is some percentage of the original data which is used to check the accuracy of our model.

17. What is the purpose of a validation set?

-> The purpose of a validation set is to select best model and tune the hyperparameter.

18. What can go wrong if you tune hyperparameters using the test set?

-> Overfitting of the test set

19. What is repeated cross-validation and why would you prefer it to using a single validation set?

-> Repeated cross validation technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the supportive subset of the data. I prefer it to using a single validation set because it increases the training time multiplied by the number of validation sets.

Interview Questions

• Scenario: You are working on a dataset that includes patient records in a healthcare database. After initial analysis, you identify that 15% of the patient age data is missing. The dataset is fairly large, with over 100,000 records. The age data is important for your analysis to track disease prevalence across age groups.

• Question 1: Describe how you would handle these missing values. Discuss the techniques you would consider and explain your choice. How would your approach change if missing values were identified in a more critical feature, such as diagnosis information?

-> I will handle these missing values in two choices:

I prefer choosing deletion because 15% of 100,000 is not the big dataset so deleting it won't affect the Machine learning model.

Secondly we can also apply Mean/Median Imputation cause its the numerical value and missing data is not significant as well.

If missing data is critical, I will handle missing data by predication models because it works great on clear pattern and diagnosis information have clear patterns as well.

- **Scenario: You are tasked with cleaning a financial dataset used for forecasting stock prices. The dataset contains some apparent outliers in the volume of trades, which could potentially skew the predictive models. Preliminary analysis shows that these outliers represent days with significant market news.**
- **Question 2: What methods would you use to detect these outliers, and how would you decide whether to remove or adjust these data points in the dataset? Outline the potential impacts of your decision on the forecast model's performance.**

-> I would use the Interquartile Range method to detect these outliers because it works well with skewed distribution. If outliers are deemed errors or irrelevant, I will remove the datapoints in the dataset otherwise I will adjust the data points in the dataset. I this case i think I am going to adjust the data point using the Transformation technique. It will help to reduce the skewness of the predictive model and reduce the impact of the outliers.

- **Scenario: Imagine you are preparing a dataset for a machine learning model that predicts real estate prices. The dataset features have varying scales and distributions, including property size in square feet and local crime rate per 1,000 residents.**
- **Question 4: Would you choose to normalize or standardize these features, and why? Provide a detailed explanation of how each process would affect the data and the model's learning process. What might be the implications of choosing one method over the other in terms of model performance and accuracy?**

I would standardize these features because linear regression, the likely algorithm for this task, benefits from standardized data. Standardization is less affected by outliers compared to normalization. While normalizing the features the compresses outliers so it would be generalizability as well but by doing standardize everything will transform and there will be low impact of outliers and it leading better generalization of the features.

Read Lesson 2 and summarize it.

The lesson 2 deals with the regression which is useful for showing the relationship between dependent variables and one or more independent variables. We get equation from regression which is used for prediction for dependent variables. The best line is the least

amount of error.
Mainly three type of regression is mentioned in the lesson.

Linear regression:
Formula for calculating linear regression.
No iteration

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Polynomial regression:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon.$$

Iterable
providing a better fit for datasets with complex relationships.

Non-linear regression:
model will overfit very quicky for non linear regression.

Two methods for evaluating linear regression
Mean Squared Error

Average squared between the estimated values and the actual values

Calculation: $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Root Mean Squared Error:

Providing measure of the magnitude of the error

Calculation: $RMSE = \sqrt{MSE}$