ML 50 Interview Question

Q1. Differentiate between Training Sets and Test Sets?

Ans. Training set is the labelled data that is used to train the model. The data in the training set are the examples provided to the model to train that particular model

Test Sets is the data in the test are used to test the model accuracy of the already trained model. For testing purpose , labelled data is not used at all , the results are further verified with the labels.

Q2. Define Bias and Variance

Ans. Bias 🡺 When a model makes predictions , the difference between model’s prediction values and actual values arises is called bias.

Variance 🡺 The variance would describe the degree of variation in the prediction.

Q3. You have come across some missing data in your dataset. How will you handle it.

Ans. Is null () – is used to find missing values in a dataset

Fill na () – is used to fill missing values with 0’s

Q4. Explain Decision Tree Classification?

Ans. A decision tree uses a tree structure to generate any regression and classification models. While the decision tree is developed, the datasets are split up into ever – smaller subsets in a tree like manner with branches and nodes. It can handle both numerical data and categorical data

Q5. To start Linear Regression, you would need to make some assumptions. What are those assumptions?

Ans. To start a Linear Regression model, there are some fundamental assumptions that are:

🡺The model should have a multivariate normal distribution

🡺There should be no auto – correlation

🡺There should be a linear relationship

Q6. What is multicollinearity?

Ans. If there is a correlation between the independent variables in a regression model, it is known as multicollinearity.

Q7. What are Machine Learning and Artificial Intelligence?

Ans. Machine Learning is a branch of AI that develops algorithms by learning the hidden patterns of the datasets used it to make predictions on new similar type data .

Artificial Intelligence is the ability of a machine to display human like intelligence to perform various tasks , like planning , reasoning , creativity and planning.

Q8. Differentiate between Deep learning and machine learning?

Ans. ML adopts algorithms to learn from data sets and apply this to future decision making

DL is a subset of machine learning that uses large amounts of data and complex algorithms to create neural networks that can learn and make decisions on their own

Q9. What is cross validation?

Ans. It is a concept used to evaluate model’s performance to avoid overfitting

Q10. What are three type of Machine learning?

Ans. There are mainly three types of machine learning:

1. Reinforcement learning: ML technique that trains software to make decisions to achieve the most optimal results.
2. Supervised learning: Using labelled datasets to train algorithms to classify data easily for predicting accurate outcomes.
3. Unsupervised learning: It uses ML to analyze and cluster unlabeled datasets. It discover hidden pattern and structures in data

Q10. What is Selection Bias?

Ans. Selection bias occurs if a data set’s examples are chosen in a way that is not reflective of their real world distribution.

Q11. What is the difference between correlation and causality?

Ans. Correlation is the relation of one action (A) to another action (B) when A does not necessarily lead to B, but Causality is the situation where one action (A) causes a result (B)

Q12. What is the difference between Correlation and Covariance?

Ans. Correlation qualifies the relationship between two random variables with three values 0, 1, -1.

Covariance is the measure of how two different variables are related and how changes in one impact the other.

Q13.What different targets to classification and regression algorithms require?

Ans. Regression algorithms require categorical and numerical targets. Regression finds correlations between dependent and independent variables. Example – market trends and weather patterns

Classification is an algorithm that segregates the datasets into classes on various parameter. Example – loans, email, or spam classification.

Q14. Difference between KNN and K-means clustering?

Ans. KNN🡺 useful for classification and regression

This is a supervised technique

K-means🡺 used for clustering

This is an unsupervised learning

Q15. Where are semi – supervised learning applied?

Ans. Some areas it is applied include labelling data, fraud detection and machine translation. It happens when a small amount of labelled data is introduce to an algorithm The algorithm studies that data and uses it on unlabeled data. It combines the efficiency of unsupervised learning and the performance of supervised learning.

Q16. What are support vectors in SVM?

Ans. Support vectors are the data points in a dataset that are closest to the hyperplane and are used to build the classifier.

Q17. What is a neural network?

Ans. It is like a human brain, the neural network is a network of different neurons connected in a way that helps information flow from one neuron to the other.

Q18. What is Outlier?

Ans. An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection or simply show variables not considered when collecting the data.

Q19. What is ensemble learning?

Ans. It is a method that merges multiple machine learning models to create various powerful models. The aim is to provide better performance by combining models rather than sticking to a single model.

Q20. What is clustering?

Ans. Clustering is a process of grouping sets of items into

Several groups. Items or objects must be similar within the cluster and different from other objects in other clusters.

Q21. What is overfitting?

Ans. Overfitting models produce good predictions for data points in the training set but perform poorly on new samples.

Q22. What is the Bayesian Network?

Ans. Bayesian network represents a graphical model between sets of variables. A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.

Q23. What is underfitting?

Ans. It is a type of error in ML models where the model fails to capture the underlying pattern of the data.

Q24. Why is the Naïve Bayes Method ‘NAÏVE’?

Ans. The Navie Bayes is called naïve because, as a supervised learning algorithms, it makes assumptions by applying the Bayes Theorem that all attributes are independent of each other.

Q25. What are some types of clustering?

Ans. 1. Fuzzy Clustering: Each data point can belong to multiple clusters.

2. Hierarchical Clustering: It is an algorithm that groups similar objects into segments called clusters.

3. Density Based Clustering: It clusters data based on their density and hence requires uniform density within the cluster.

Q26. What is Loss function and Cost functions?

Ans. Loss function shows the difference between the actual and predicted values for a single record

Cost function is calculating the sum of errors for multiple data.

Q27. How do you handle outlier values?

Ans. 1. Dropping them

2. Marking them as outliers and including them as a feature

3.Transforming the features to reduce the effect of the outlier

Q28. What is a random forest, and how does it work?

Ans Random forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is built using a different subset of the training data and a random subset of the features. This randomness helps to prevent overfitting and improve the generalization of the model.

Q29. What ensembles techniques can be used to aggregate multiple models?

Ans. Bagging: The bagging technique combines multiple models trained on different subsets of data to reduce error and optimize the model

Boosting: The booting trains the model sequentially, focusing on the error made by the previous model.

Q30. What difference exist between soft max and sigmoid functions?

Ans: The sigmoid function is a mathematical function that maps any input value to a value between 0 and 1 and it is used for binary classification.

Soft max function is used for multi – classification and the sum of the probabilities will be 1.

Q31. How overfitting can be avoided in ML?

Ans. 1. Cross Validation

2. Custom Feature selection

3. Data augmentation

4. Using larger datasets

5. Data Simplification

Q32. What is Gradient Boosting?

Ans. In gradient boosting , the trees are trained sequentially one after another, thereby correcting errors of the previous one. It gives better performance when there is unbalanced data.

Q33. What are some classification methods that SVM can handle?

Ans. 1. Binary Classification

2. Multiclass classification

3. Multilabel classification

Q34. How would you handle a dataset suffering from high variance?

Ans. The bagging algorithm can be used to split the data into subgroups sampling replicated from random data.

Q35. Why do we perform normalization?

Ans. To achieve stable and fast training of the model we use normalization techniques to bring all the features to a certain scale or range value.

Q36. What is one shot learning?

Ans. It is a concept in machine learning where the model is trained to recognize the patterns in datasets from a single example instead of training on large datasets.

Q37. What is One hot encoding and Ordinal encoding?

Ans. One hot encoding and ordinal encoding both are different methods to convert categorical data into numerical data. In one hot encoding, we create a separate column for each category and add 0 or 1 as per the value corresponding to that row. Hence In ordinal encoding we replace the categories with numbers from 0 to n-1 based on the rank where n is the number of unique categories present in the dataset.

Q38. How to handle missing values in that data?

Ans. 🡺 Removing the rows with null values may lead to the loss of some important information

🡺Removing the column having null values if it has very less valuable information, it may lead to the loss of some important information.

🡺Imputing null values with descriptive statistical measures like mean, mode and median

🡺Using methods like KNN imputer to impute the null values in a more sophisticated way.

Q39. What is gradient descent?

Ans. Gradient descent is the method of minimizing the cost function. The form of the cost function will depend on the type of supervised model.

Q40. What is the apriori algorithms?

Ans. It is a traditional data mining technique for association rules mining in transactional databases or datasets.