COMPREHENSIVE GUIDE TO INTERVIEWS FOR MACHINE LEARNING

ZEP ANALYTICS

ZEP ANALYTICS

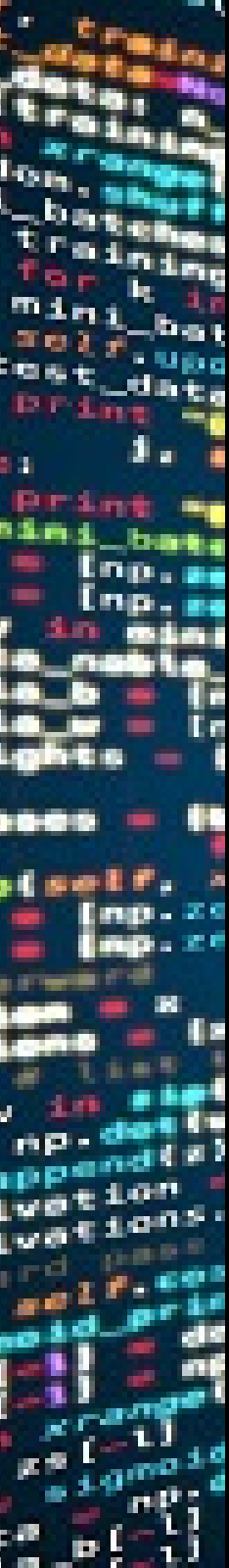
Introduction

We've curated this series of interview guides to accelerate your learning and your mastery of data science skills and tools.

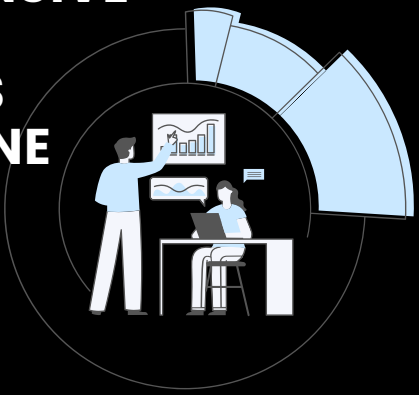
From job-specific technical questions to tricky behavioral inquires and unexpected brainteasers and guesstimates, we will prepare you for any job candidacy in the fields of data science, data analytics, or BI analytics.

These guides are the result of our data analytics expertise, direct experience interviewing at companies, and countless conversations with job candidates. Its goal is to teach by example - not only by giving you a list of interview questions and their answers, but also by sharing the techniques and thought processes behind each question and the expected answer.

Become a global tech talent and unleash your next, best self with all the knowledge and tools to succeed in a data analytics interview with this series of guides.



COMPREHENSIVE GUIDE TO INTERVIEWS FOR MACHINE LEARNING



Machine learning is one of the most important part of the Data Analytics interview. Most of the companies prefer to have a person with skills of Machine Learning.

So here we have brought you the intensive Machine Learning interview E-book which will provide you the most important Machine Learning interview questions that you can encounter during your interviews.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

1. What Are the Different Types of Machine Learning?

There are three types of machine learning:

Supervised Learning

In supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.

Unsupervised Learning

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.

Reinforcement Learning

Using reinforcement learning, the model can learn based on the rewards it received for its previous action.

Consider an environment where an agent is working. The agent is given a target to achieve. Every time the agent takes some action toward the target, it is given positive feedback. And, if the action taken is going away from the goal, the agent is given negative feedback.

2. What is Overfitting, and How Can You Avoid It?

The Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data. When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.

There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used
- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

3. What is 'training Set' and 'test Set' in a Machine Learning Model? How Much Data Will You Allocate for Your Training, Validation, and Test Sets?

There is a three-step process followed to create a model:

1. Train the model
2. Test the model
3. Deploy the model

Training Set

- The training set is examples given to the model to analyze and learn.
- 70% of the total data is typically taken as the training dataset.
- This is labeled data used to train the model.

Test Set

- The test set is used to test the accuracy of the hypothesis generated by the model
- Remaining 30% is taken as testing dataset
- We test without labeled data and then verify results with labels

Consider a case where you have labeled data for 1,000 records. One way to train the model is to expose all 1,000 records during the training process. Then you take a small set of the same data to test the model, which would give good results in this case.

But, this is not an accurate way of testing. So, we set aside a portion of that data called the 'test set' before starting the training process. The remaining data is called the 'training set' that we use for training the model. The training set passes through the model multiple times until the accuracy is high, and errors are minimized.

Now, we pass the test data to check if the model can accurately predict the values and determine if training is effective. If you get errors, you either need to change your model or retrain it with more data.

Regarding the question of how to split the data into a training set and test set, there is no fixed rule, and the ratio can vary based on individual preferences.

4. How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.

There are two useful methods in Pandas:

- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

5. How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit.

For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

6. What Are the Three Stages of Building a Model in Machine Learning?

The three stages of building a machine learning model are:

- Model Building

Choose a suitable algorithm for the model and train it according to the requirement

- Model Testing

Check the accuracy of the model through the test data

- Applying the Model

Make the required changes after testing and use the final model for real-time projects

Here, it's important to remember that once in a while, the model needs to be checked to make sure it's working correctly. It should be modified to make sure that it is up-to-date.

7. What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:

- Email Spam Detection

Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.

- Healthcare Diagnosis

By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.

- Sentiment Analysis

This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral, or negative in sentiment.

8. What is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.
In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data.

9. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

Clustering
Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

Association
In an association problem, we identify patterns of associations between different variables or items.
For example, an e-commerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

10. What is the Difference Between Supervised and Unsupervised Machine Learning?

- Supervised learning - This model learns from the labeled data and makes a future prediction as output
- Unsupervised learning - This model uses unlabeled input data and allows the algorithm to act on that information without guidance.

11. What Is 'naive' in the Naive Bayes Classifier?

The classifier is called 'naive' because it makes assumptions that may or may not turn out to be correct.
The algorithm assumes that the presence of one feature of a class is not related to the presence of any other feature (absolute independence of features), given the class variable.
For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

12. Explain How a System Can Play a Game of Chess Using Reinforcement Learning.

Reinforcement learning has an environment and an agent. The agent performs some actions to achieve a specific goal. Every time the agent performs a task that is taking it towards the goal, it is rewarded. And, every time it takes a step that goes against that goal or in the reverse direction, it is penalized.

Earlier, chess programs had to determine the best moves after much research on numerous factors. Building a machine designed to play such games would require many rules to be specified.

With reinforced learning, we don't have to deal with this problem as the learning agent learns by playing the game. It will make a move (decision), check if it's the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

13. How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

14. When Will You Use Classification over Regression?

Classification is used when your target is categorical, while regression is used when your target variable is continuous. Both classification and regression belong to the category of supervised machine learning algorithms.

Examples of classification problems include:

- Predicting yes or no
- Estimating gender
- Breed of an animal
- Type of color

Examples of regression problems include:

- Estimating sales and price of a product
- Predicting the score of a team
- Predicting the amount of rainfall

15. What is a Random Forest?

A 'random forest' is a supervised machine learning algorithm that is generally used for classification problems. It operates by constructing multiple decision trees during the training phase. The random forest chooses the decision of the majority of the trees as the final decision.

16. What is Bias and Variance in a Machine Learning Model?

Bias

Bias in a machine learning model occurs when the predicted values are further from the actual values. Low bias indicates a model where the prediction values are very close to the actual ones.

Underfitting: High bias can cause an algorithm to miss the relevant relations between features and target outputs.

Variance
Variance refers to the amount the target model will change when trained with different training data. For a good model, the variance should be minimized.

Overfitting: High variance can cause an algorithm to model the random noise in the training data rather than the intended outputs.

17. What is the Trade-off Between Bias and Variance?

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, variance, and a bit of irreducible error due to noise in the underlying dataset.
Necessarily, if you make the model more complex and add more variables, you'll lose bias but gain variance. To get the optimally-reduced amount of error, you'll have to trade off bias and variance. Neither high bias nor high variance is desired.
High bias and low variance algorithms train models that are consistent, but inaccurate on average.
High variance and low bias algorithms train models that are accurate but inconsistent.

18. What is a Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

19. What is Pruning in Decision Trees, and How Is It Done?

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning can occur in:

- Top-down fashion. It will traverse nodes and trim subtrees starting at the root
- Bottom-up fashion. It will begin at the leaf nodes

There is a popular pruning algorithm called reduced error pruning, in which:

- Starting at the leaves, each node is replaced with its most popular class
- If the prediction accuracy is not affected, the change is kept
- There is an advantage of simplicity and speed

20. Briefly Explain Logistic Regression.

Logistic regression is a classification algorithm used to predict a binary outcome for a given set of independent variables.

The output of logistic regression is either a 0 or 1 with a threshold value of generally 0.5. Any value above 0.5 is considered as 1, and any point below 0.5 is considered as 0.

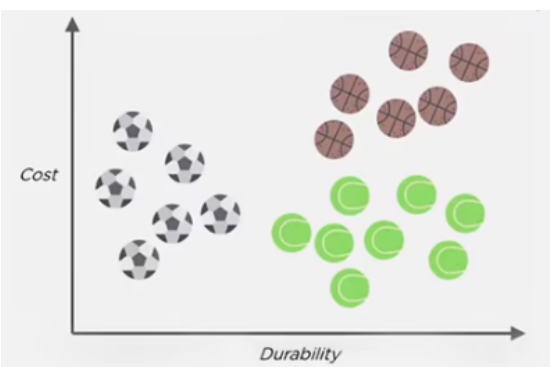
21. Explain the K Nearest Neighbor Algorithm.

K nearest neighbor algorithm is a classification algorithm that works in a way that a new data point is assigned to a neighboring group to which it is most similar.

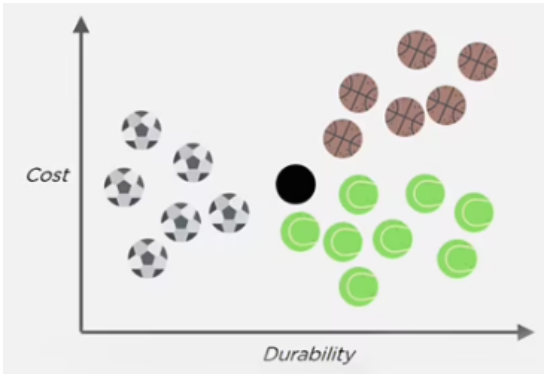
In K nearest neighbors, K can be an integer greater than 1. So, for every new data point, we want to classify, we compute to which neighboring group it is closest.

Let us classify an object using the following example. Consider there are three clusters:

- Football
- Basketball
- Tennis ball



Let the new data point to be classified is a black ball. We use KNN to classify it. Assume K = 5 (initially).
Next, we find the K (five) nearest data points, as shown.



Observe that all five selected points do not belong to the same cluster. There are three tennis balls and one each of basketball and football.

When multiple classes are involved, we prefer the majority. Here the majority is with the tennis ball, so the new data point is assigned to this cluster.

22. What is a Recommendation System?

Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system: It's an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.

23. What is Kernel SVM?

Kernel SVM is the abbreviated version of the kernel support vector machine. Kernel methods are a class of algorithms for pattern analysis, and the most common one is the kernel SVM.

24. What Are Some Methods of Reducing Dimensionality?

You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.
Now that you have gone through these machine learning interview questions, you must have got an idea of your strengths and weaknesses in this domain.

25. What is Principal Component Analysis?

Principal Component Analysis or PCA is a multivariate statistical technique that is used for analyzing quantitative data. The objective of PCA is to reduce higher dimensional data to lower dimensions, remove noise, and extract crucial information such as features and attributes from large amounts of data.

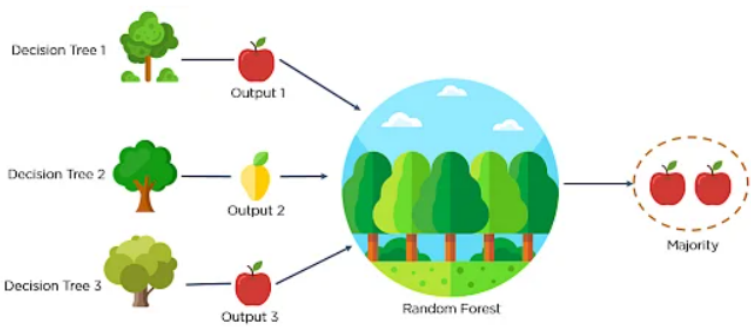
26. What are Support Vectors in SVM?

Support Vectors are data points that are nearest to the hyperplane. It influences the position and orientation of the hyperplane. Removing the support vectors will alter the position of the hyperplane. The support vectors help us build our support vector machine model.

27. What is Ensemble learning?

Ensemble learning is a combination of the results obtained from multiple machine learning models to increase the accuracy for improved decision-making.

Example: A Random Forest with 100 trees can provide much better results than using just one decision tree.



28. What is Cross-Validation?

Cross-Validation in Machine Learning is a statistical resampling technique that uses different parts of the dataset to train and test a machine learning algorithm on different iterations. The aim of cross-validation is to test the model's ability to predict a new set of data that was not used to train the model. Cross-validation avoids the overfitting of data.

K-Fold Cross Validation is the most popular resampling technique that divides the whole dataset into K sets of equal sizes.

29. What are the different methods to split a tree in a decision tree algorithm?

Variance: Splitting the nodes of a decision tree using the variance is done when the target variable is continuous.

Information Gain: Splitting the nodes of a decision tree using Information Gain is preferred when the target variable is categorical.

Gini Impurity: Splitting the nodes of a decision tree using Gini Impurity is followed when the target variable is categorical.

30. How does the Support Vector Machine algorithm handle self-learning?

The SVM algorithm has a learning rate and expansion rate which takes care of self-learning. The learning rate compensates or penalizes the hyperplanes for making all the incorrect moves while the expansion rate handles finding the maximum separation area between different classes.

31. What is the difference between Lasso and Ridge regression?

Lasso(also known as L1) and Ridge(also known as L2) regression are two popular regularization techniques that are used to avoid overfitting of data. These methods are used to penalize the coefficients to find the optimum solution and reduce complexity. The Lasso regression works by penalizing the sum of the absolute values of the

coefficients. In Ridge or L2 regression, the penalty function is determined by the sum of the squares of the coefficients.

32. What are the assumptions you need to take before starting with linear regression?

There are primarily 5 assumptions for a Linear Regression model:

- Multivariate normality
- No auto-correlation
- Homoscedasticity
- Linear relationship
- No or little multicollinearity

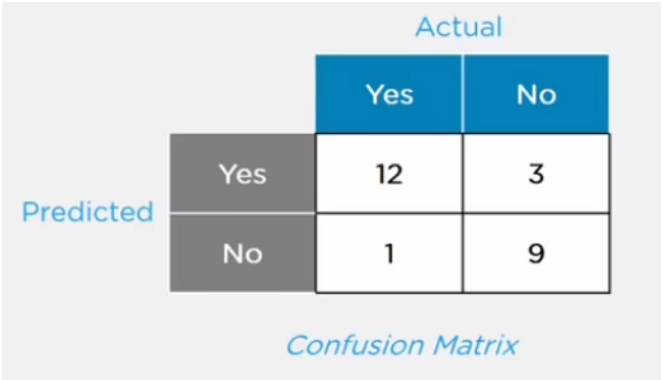
33. Explain the Confusion Matrix with Respect to Machine Learning Algorithms.

A confusion matrix (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it's called the matching matrix.

The confusion matrix has two parameters:

- Actual
- Predicted

It also has identical sets of features in both of these dimensions.



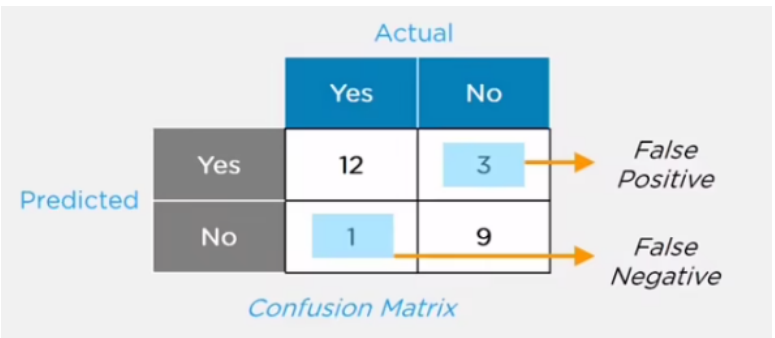
Here,
For actual values:
Total Yes = 12+1 = 13
Total No = 3+9 = 12
Similarly, for predicted values:
Total Yes = 12+3 = 15
Total No = 1+9 = 10
For a model to be accurate, the values across the diagonals should be high. The total sum of all the values in the matrix equals the total observations in the test data set.
For the above matrix, total observations = 12+3+1+9 = 25
Now, accuracy = sum of the values across the diagonal/total dataset
= (12+9) / 25
= 21 / 25
= 84%

34. What Is a False Positive and False Negative and How Are They Significant?

False positives are those cases that wrongly get classified as True but are False.

False negatives are those cases that wrongly get classified as False but are True.

In the term 'False Positive,' the word 'Positive' refers to the 'Yes' row of the predicted value in the confusion matrix. The complete term indicates that the system has predicted it as a positive, but the actual value is negative.



So, looking at the confusion matrix, we get:

False-positive = 3

True positive = 12

Similarly, in the term 'False Negative,' the word 'Negative' refers to the 'No' row of the predicted value in the confusion matrix. And the complete term indicates that the system has predicted it as negative, but the actual value is positive.

So, looking at the confusion matrix, we get:

False Negative = 1

True Negative = 9

35. Define Precision and Recall.

Precision

Precision is the ratio of several events you can correctly recall to the total number of events you recall (mix of correct and wrong recalls).

Precision = (True Positive) / (True Positive + False Positive)

Recall

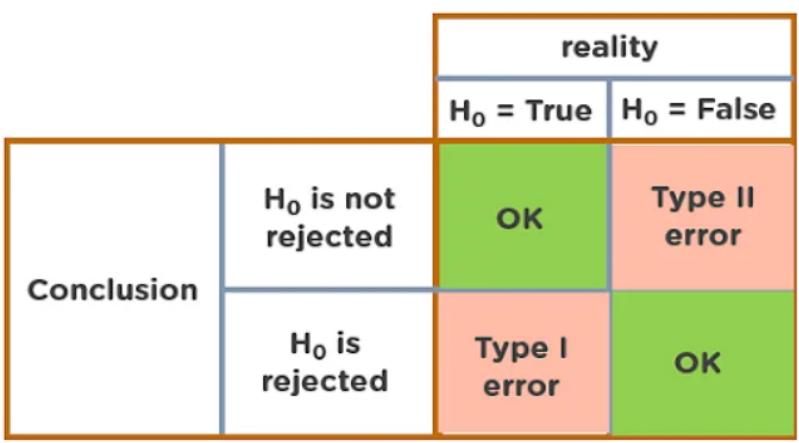
A recall is the ratio of the number of events you can recall the number of total events.

Recall = (True Positive) / (True Positive + False Negative)

36. What do you understand by Type I vs Type II error?

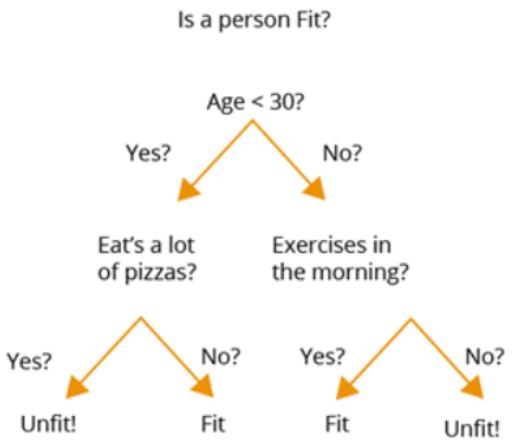
Type I Error: Type I error occurs when the null hypothesis is true and we reject it.

Type II Error: Type II error occurs when the null hypothesis is false and we accept it.



37. What is a Decision Tree in Machine Learning?

A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.



An algorithm can be created for a decision tree on the basis of the set hierarchy of actions.

In the above decision-tree diagram, a sequence of actions has been made for driving a vehicle with or without a license.

38. What is Hypothesis in Machine Learning?

Machine Learning allows the use of available dataset to understand a specific function that maps input to output in the best possible way. This problem is known as function approximation. Here, approximation needs to be used for the

the unknown target function that maps all plausible observations based on the given problem in the best manner. Hypothesis in Machine learning is a model that helps in approximating the target function and performing the necessary input-to-output mappings. The choice and configuration of algorithms allow defining the space of plausible hypotheses that may be represented by a model.
In the hypothesis, lowercase h (h) is used for a specific hypothesis, while uppercase h (H) is used for the hypothesis space that is being searched. Let us briefly understand these notations:

- Hypothesis (h): A hypothesis is a specific model that helps in mapping input to output; the mapping can further be used for evaluation and prediction.
- Hypothesis set (H): Hypothesis set consists of a space of hypotheses that can be used to map inputs to outputs, which can be searched. The general constraints include the choice of problem framing, the model, and the model configuration.

39. What are the differences between Deep Learning and Machine Learning?

Deep Learning: Deep Learning allows machines to make various business-related decisions using artificial neural networks, which is one of the reasons why it needs a vast amount of data for training. Since there is a lot of computing power required, Deep Learning requires high-

-end systems as well. The systems acquire various properties and features with the help of the given data, and the problem is solved using an end-to-end method.

Machine Learning: Machine Learning gives machines the ability to make business decisions without any external help, using the knowledge gained from past data. Machine Learning systems require relatively small amounts of data to train themselves, and most of the features need to be manually coded and understood in advance. In Machine Learning, a given business problem is dissected into two and then solved individually. Once the solutions of both have been acquired, they are then combined.

40. What is Entropy in Machine Learning?

Entropy in Machine Learning measures the randomness in the data that needs to be processed. The more entropy in the given data, the more difficult it becomes to draw any useful conclusion from the data. For example, let us take the flipping of a coin. The result of this act is random as it does not favor heads or tails. Here, the result for any number of tosses cannot be predicted easily as there is no definite relationship between the action of flipping and the possible outcomes.

41. What is Epoch in Machine Learning?

Epoch in Machine Learning is used to indicate the count of passes in a given training dataset where the Machine Learning algorithm has done its job. Generally, when there is a large chunk of data, it is grouped into several batches. All these batches go through the given model, and this process is referred to as iteration. Now, if the batch size comprises the complete training dataset, then the count of iterations is the same as that of epochs.

In case there is more than one batch, d*e=i*b is the formula used, wherein d is the dataset, e is the number of epochs, i is the number of iterations, and b is the batch size.

42. How is the suitability of a Machine Learning Algorithm determined for a particular problem?

To identify a Machine Learning Algorithm for a particular problem, the following steps should be followed:

Step 1: Problem classification: Classification of the problem depends on the classification of input and output:

- Classifying the input: Classification of the input depends on whether there is data labeled (supervised learning) or unlabeled (unsupervised learning), or whether a model has to be created that interacts with the environment and improves itself (reinforcement learning.)
- Classifying the output: If the output of a model is required as a class, then some classification techniques need to be used.

If the output is a number, then regression techniques must be used; if the output is a different cluster of inputs, then clustering techniques should be used.

Step 2: Checking the algorithms in hand: After classifying the problem, the available algorithms that can be deployed for solving the classified problem should be considered.

Step 3: Implementing the algorithms: If there are multiple algorithms available, then all of them are to be implemented. Finally, the algorithm that gives the best performance is selected.

43. What is the Variance Inflation Factor?

Variance inflation factor (VIF) is the estimate of the volume of multicollinearity in a collection of many regression variables.

VIF = Variance of the model / Variance of the model with a single independent variable

This ratio has to be calculated for every independent variable. If VIF is high, then it shows the high collinearity of the independent variables.

44. When should Classification be used over Regression?

Both classification and regression are associated with prediction. Classification involves the identification of values or entities that lie in a specific group. Regression entails predicting a response value from consecutive sets of outcomes.

Classification is chosen over regression when the output of the model needs to yield the belongingness of data points in a dataset to a particular category.

For example, If you want to predict the price of a house, you should use regression since it is a numerical variable. However, if you are trying to predict whether a house situated in a particular area is going to be high-, medium-, or low-priced, then a classification model should be used.

45. Why is rotation required in PCA? What will happen if the components are not rotated?

Rotation is a significant step in principal component analysis (PCA.) Rotation maximizes the separation within the variance obtained by the components. This makes the interpretation of the components easier.

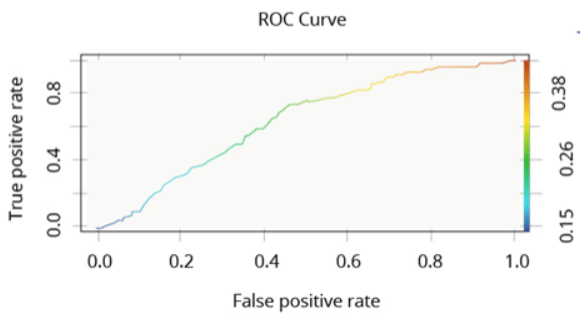
The motive behind conducting PCA is to choose fewer components that can explain the greatest variance in a dataset. When rotation is performed, the original coordinates of the points get changed. However, there is no change in the relative position of the components.

If the components are not rotated, then there needs to be more extended components to describe the variance.

46. What is ROC Curve and what does it represent?

ROC stands for receiver operating characteristic. ROC Curve is used to graphically represent the trade-off between true and false positive rates.

In ROC, area under the curve (AUC) gives an idea about the accuracy of the model.



The above graph shows an ROC curve. The greater the AUC, the better the performance of the model.

47. Why are Validation and Test Datasets Needed?

Data is split into three different categories while creating a model:

- **Training dataset:** Training dataset is used for building a model and adjusting its variables. The correctness of the model built on the training dataset cannot be relied on as the model might give incorrect outputs after being fed new inputs.

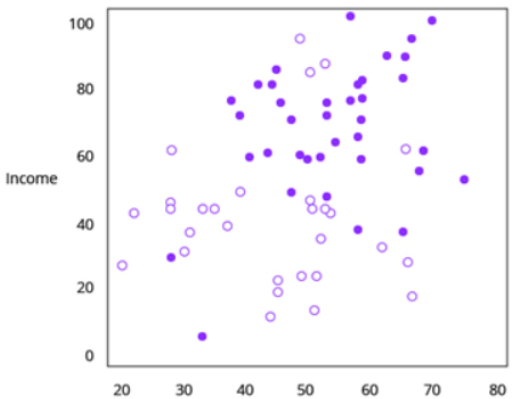
- **Validation dataset:** Validation dataset is used to look into a model's response. After this, the hyperparameters on the basis of the estimated benchmark of the validation dataset data are tuned. When a model's response is evaluated by using the validation dataset, the model is indirectly trained with the validation set. This may lead to the overfitting of the model to specific data. So, this model will not be strong enough to give the desired response to real-world data.
- **Test dataset:** Test dataset is the subset of the actual dataset, which is not yet used to train the model. The model is unaware of this dataset. So, by using the test dataset, the response of the created model can be computed on hidden data. The model's performance is tested on the basis of the test dataset.

Note: The model is always exposed to the test dataset after tuning the hyperparameters on top of the validation dataset.

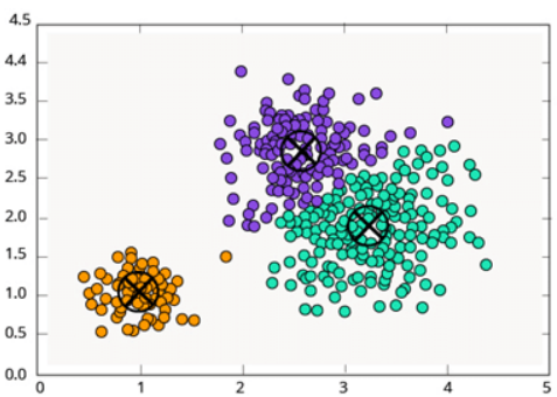
As we know, the evaluation of the model on the basis of the validation dataset would not be enough. Thus, the test dataset is used for computing the efficiency of the model.

48. Explain the difference between KNN and K-means Clustering.

K-nearest neighbors (KNN): It is a supervised Machine Learning algorithm. In KNN, identified or labeled data is given to the model. The model then matches the points based on the distance from the closest points.



K-means clustering: It is an unsupervised Machine Learning algorithm. In K-means clustering, unidentified or unlabeled data is given to the model. The algorithm then creates batches of points based on the average of the distances between distinct points.



49. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them.

This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables.

50. Both being Tree-based Algorithms, how is Random Forest different from Gradient Boosting Machine (GBM)?

The main difference between a random forest and GBM is the use of techniques. Random forest advances predictions using a technique called bagging. On the other hand, GBM advances predictions with the help of a technique called boosting.

- **Bagging:** In bagging, we apply arbitrary sampling and we divide the dataset into N. After that, we build a model by employing a single training algorithm. Following that, we combine the final predictions by polling. Bagging helps to increase the efficiency of a model by decreasing the variance to eschew overfitting.
- **Boosting:** In boosting, the algorithm tries to review and correct the inadmissible predictions at the initial iteration. After that, the algorithm's sequence of iterations for correction continues until we get the desired prediction.

Boosting assists in reducing bias and variance for strengthening the weak learners.

51. What is meant by Parametric and Non-parametric Models?

Parametric models refer to the models having a limited number of parameters. In case of parametric models, only the parameter of a model is needed to be known to make predictions regarding the new data.

Non-parametric models do not have any restrictions on the number of parameters, which makes new data predictions more flexible. In case of non-parametric models, the knowledge of model parameters and the state of the data needs to be known to make predictions.

52. Differentiate between Sigmoid and Softmax Functions.

Sigmoid and Softmax functions differ based on their usage in Machine Learning task classification. Sigmoid function is used in the case of binary classification, while Softmax function is used in case of multi-classification.

53. In Machine Learning, for how many classes can Logistic Regression be used?

Logistic regression cannot be used for more than two classes. Logistic regression is, by default, a binary classifier. However, in cases where multi-class classification problems need to be solved, the default number of classes can be extended, i.e., multinomial logistic regression.

54. What is meant by Correlation and Covariance?

Correlation is a mathematical concept used in statistics and probability theory to measure, estimate, and compare data samples taken from different populations. In simpler terms, correlation helps in establishing a quantitative relationship between two variables.

Covariance is also a mathematical concept; it is a simpler way to arrive at a correlation between two variables. Covariance basically helps in determining what change or affect does one variable has on another.

55. What are the Various Tests for Checking the Normality of a Dataset?

In Machine Learning, checking the normality of a dataset is very important. Hence, certain tests are performed on a dataset to check its normality. Some of them are:

- D'Agostino Skewness Test
- Shapiro-Wilk Test
- Anderson-Darling Test
- Jarque-Bera Test
- Kolmogorov-Smirnov Test

56. What are the Two Main Types of Filtering in Machine Learning? Explain.

The two types of filtering are:

- Collaborative filtering
- Content-based filtering

Collaborative filtering refers to a recommender system where the interests of the individual user are matched with preferences of multiple users to predict new content.
Content-based filtering is a recommender system where the focus is only on the preferences of the individual user and not on multiple users.

57. Outlier Values can be Discovered from which Tools?

The various tools that can be used to discover outlier values are scatterplots, boxplots, Z-score, etc.

58. What is meant by Ensemble Learning?

Ensemble learning refers to the combination of multiple Machine Learning models to create more powerful models. The primary techniques involved in ensemble learning are bagging and boosting.

59. What are the Various Kernels that are present in SVM?

The various kernels that are present in SVM are:

- Linear
- Polynomial
- Radial Basis
- Sigmoid

60. Suppose you found that your model is suffering from high variance. Which algorithm do you think could handle this situation and why?

Handling High Variance

- For handling issues of high variance, we should use the bagging algorithm.
- The bagging algorithm would split data into subgroups with a replicated sampling of random data.
- Once the algorithm splits the data, we can use random data to create rules using a particular training algorithm.
- After that, we can use polling for combining the predictions of the model.

61. What is Binarizing of Data? How to Binarize?

Converting data into binary values on the basis of threshold values is known as binarizing of data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when feature engineering has to be performed. This can also be used for adding unique features. Data can be binarized using Scikit-learn.

62. How to Standardize Data?

Standardization is the method that is used for rescaling data attributes. The attributes are likely to have the mean value as 0 and the value of standard deviation as 1. The main objective of standardization is to prompt the mean and standard deviation for the attributes.

63. We know that one-hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

When one-hot encoding is used, there is an increase in the dimensionality of a dataset. The reason for the increase in dimensionality is that, for every class in categorical variables, it forms a different variable.

Example: Suppose there is a variable "Color." It has three sublevels, "Yellow," "Purple," and "Orange." So, one-hot encoding "Color" will create three different variables as Color.Yellow, Color.Purple, and Color.Orange.

In label encoding, the subclasses of a certain variable get the value as 0 and 1. So, label encoding is only used for binary variables.

This is why one-hot encoding increases the dimensionality of data and label encoding does not.

64. Imagine you are given a dataset consisting of variables having more than 30% missing values. Let's say, out of 50 variables, 16 variables have missing values, which is higher than 30%. How will you deal with them?

To deal with the missing values, we will do the following:

- We will specify a different class for the missing values.
- Now, we will check the distribution of values, and we will hold those missing values that are defining a pattern.
- Then, we will charge these values into yet another class while eliminating others.

65. Explain False Negative, False Positive, True Negative, and True Positive with a simple example.

True Positive (TP): When the Machine Learning model correctly predicts the condition, it is said to have a True Positive value.

True Negative (TN): When the Machine Learning model correctly predicts the negative condition or class, then it is said to have a True Negative value.

False Positive (FP): When the Machine Learning model incorrectly predicts a negative class or condition, then it is said to have a False Positive value.

False Negative (FN): When the Machine Learning model incorrectly predicts a positive class or condition, then it is said to have a False Negative value.

66. What is F1-score and How Is It Used?

F-score or F1-score is a measure of overall accuracy of a binary classification model. Before understanding F1-score, it is crucial to understand two more measures of accuracy, i.e., precision and recall.

Precision is defined as the percentage of True Positives to the total number of positive classifications predicted by the model. In other words,

Precision = (No. of True Positives / No. True Positives + No. of False Positives)

Recall is defined as the percentage of True Positives to the total number of actual positive labeled data passed to the model. In other words,

Precision = (No. of True Positives / No. True Positives + No. of False Negatives)

Both precision and recall are partial measures of accuracy of a model. F1-score combines precision and recall and provides an overall score to measure a model's accuracy.

F1-score = 2 × (Precision × Recall) / (Precision + Recall)

This is why, F1-score is the most popular measure of accuracy in any Machine-Learning-based binary classification model.

67. How can you avoid overfitting ?

By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as **cross validation**. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.

In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to "test" the model in the training phase.

68. What is inductive machine learning?

The inductive machine learning involves the process of learning by examples, where a system, from a set of observed instances tries to induce a general rule.

69. What is Genetic Programming?

Genetic programming is one of the two techniques used in machine learning. The model is based on the testing and selecting the best choice among a set of results.

70. What is Inductive Logic Programming in Machine Learning?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical programming representing background knowledge and examples.

71. What is Model Selection in Machine Learning?

The process of selecting models among different mathematical models, which are used to describe the same data set is known as Model Selection. Model selection is applied to the fields of statistics, machine learning and data mining.

72. What is the difference between heuristic for rule learning and heuristics for decision trees?

The difference is that the heuristics for decision trees evaluate the average quality of a number of disjointed sets while rule learners only evaluate the quality of the set of instances that is covered with the candidate rule.

73. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

74. What is bias-variance decomposition of classification error in ensemble method?

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

75.What is an Incremental Learning algorithm in ensemble?

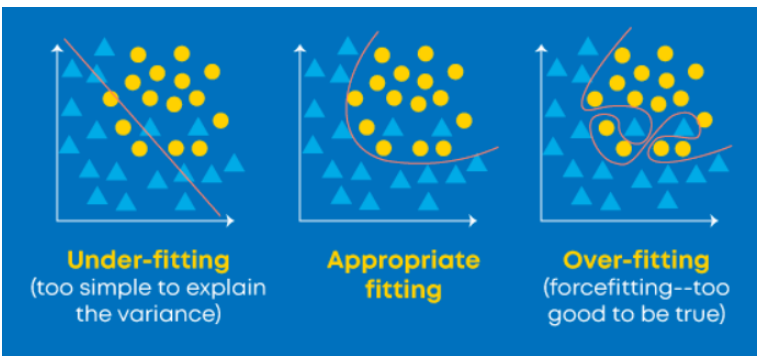
Incremental learning method is the ability of an algorithm to learn from new data that may be available after classifier has already been generated from already available dataset.

76. What is PCA, KPCA and ICA used for?

PCA (Principal Components Analysis), KPCA (Kernel based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

77. When does regularization come into play in Machine Learning?

At times when the model begins to underfit or overfit, regularization becomes necessary. It is a regression that diverts or regularizes the coefficient estimates towards zero. It reduces flexibility and discourages learning in a model to avoid the risk of overfitting. The model complexity is reduced and it becomes better at predicting.



78. How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. Variance is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

79. Is a high variance in data good or bad?

Higher variance directly means that the data spread is big and the feature has a variety of data. Usually, high variance in a feature is seen as not so good quality.

80. Explain the handling of missing or corrupted values in the given dataset.

An easy way to handle missing values or corrupted values is to drop the corresponding rows or columns. If there are too many rows or columns to drop then we consider replacing the missing or corrupted values with some new value.
Identifying missing values and dropping the rows or columns can be done by using IsNull() and dropna( ) functions in Pandas. Also, the Fillna() function in Pandas replaces the incorrect values with the placeholder value.

81. What is Time series?

A Time series is a sequence of numerical data points in successive order. It tracks the movement of the chosen data points, over a specified period of time and records the data points at regular intervals. Time series doesn't require any minimum or maximum time input. Analysts often use Time series to examine data according to their specific requirement.

82. What is a Box-Cox transformation?

Box-Cox transformation is a power transform which transforms non-normal dependent variables into normal variables as normality is the most common assumption made while using many statistical techniques. It has a lambda parameter which when set to 0 implies that this transform is equivalent to log-transform. It is used for variance stabilization and also to normalize the distribution.

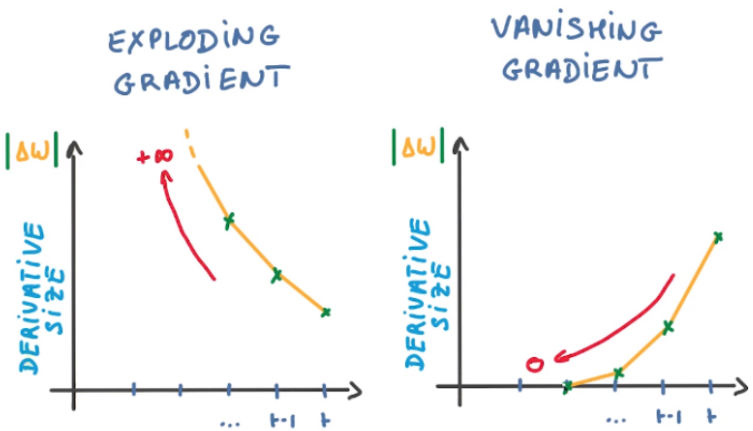
83. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Gradient Descent and Stochastic Gradient Descent are the algorithms that find the set of parameters that will minimize a loss function.

The difference is that in Gradient Descend, all training samples are evaluated for each set of parameters. While in Stochastic Gradient Descent only one training sample is evaluated for the set of parameters identified.

84. What is the exploding gradient problem while using back propagation technique?

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large as to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the vanishing gradient problem.



85. Explain the differences between Random Forest and Gradient Boosting machines.

Random forests are a significant number of decision trees pooled using averages or majority rules at the end. Gradient boosting machines also combine decision trees but at the beginning of the process unlike Random forests. Random forest creates each tree independent of the others while gradient boosting develops one tree at a time. Gradient boosting yields better outcomes than random forests if parameters are carefully tuned but it's not a good option if the data set contains a lot of outliers/anomalies/noise as it can result in overfitting of the model. Random forests perform well for multiclass object detection. Gradient Boosting performs well when there is data which is not balanced such as in real time risk assessment.

86. What's a Fourier transform?

Fourier Transform is a mathematical technique that transforms any function of time to a function of frequency. Fourier transform is closely related to Fourier series. It takes any time-based pattern for input and calculates the overall cycle offset, rotation speed and strength for all possible cycles. Fourier transform is best applied to waveforms since it has functions of time and space. Once a Fourier transform applied on a waveform, it gets decomposed into a sinusoid.

87. What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated. It is mostly used in Market-based Analysis to find how frequently an itemset occurs in a transaction. Association rules have to satisfy minimum support and minimum confidence at the very same time. Association rule generation generally comprised of two different steps:

- "A min support threshold is given to obtain all frequent item-sets in a database."
- "A min confidence constraint is given to these frequent item-sets in order to form the association rules."

Support is a measure of how often the "item set" appears in the data set and Confidence is a measure of how often a particular rule has been found to be true.

88. What is Marginalisation? Explain the process.

Marginalisation is summing the probability of a random variable X given joint probability distribution of X with other variables. It is an application of the law of total probability.

$$P(X=x) = \sum_Y P(X=x,Y)$$

Given the joint probability $P(X=x,Y)$, we can use marginalization to find $P(X=x)$. So, it is to find distribution of one random variable by exhausting cases on other random variables.

89. Explain the phrase "Curse of Dimensionality".

The Curse of Dimensionality refers to the situation when your data has too many features.

The phrase is used to express the difficulty of using brute force or grid search to optimize a function with too many inputs.

It can also refer to several other issues like:

- If we have more features than observations, we have a risk of overfitting the model.
- When we have too many features, observations become harder to cluster. Too many dimensions cause every observation in the dataset to appear equidistant from all others and no meaningful clusters can be formed.

Dimensionality reduction techniques like PCA come to the rescue in such cases.

90. What is the difference between regularization and normalisation?

Normalisation adjusts the data; regularisation adjusts the prediction function. If your data is on very different scales (especially low to high), you would want to normalise the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting. Regularization imposes some control on this by providing simpler fitting functions over complex ones.

91. Explain the difference between Normalization and Standardization.

Normalization and Standardization are the two very popular methods used for feature scaling. Normalization refers to re-scaling the values to fit into a range of [0,1]. Standardization refers to re-scaling data to have a mean of 0 and a standard deviation of 1 (Unit variance). Normalization is useful when all parameters need to have the identical positive scale however the outliers from the data set are lost. Hence, standardization is recommended for most applications.

92. List the most popular distribution curves along with scenarios where you will use them in an algorithm.

The most popular distribution curves are as follows- Bernoulli Distribution, Uniform Distribution, Binomial Distribution, Normal Distribution, Poisson Distribution, and Exponential Distribution.

Each of these distribution curves is used in various scenarios. Bernoulli Distribution can be used to check if a team will win a championship or not, a newborn child is either male or female, you either pass an exam or not, etc.

Uniform distribution is a probability distribution that has a constant probability. Rolling a single dice is one example because it has a fixed number of outcomes.

Binomial distribution is a probability with only two possible outcomes, the prefix 'bi' means two or twice. An example of this would be a coin toss. The outcome will either be heads or tails.

Normal distribution describes how the values of a variable are distributed. It is typically a symmetric distribution where most of the observations cluster around the central peak. The values further away from the mean taper off equally in both directions. An example would be the height of students in a classroom.

Poisson distribution helps predict the probability of certain events happening when you know how often that event has occurred. It can be used by businessmen to make forecasts about the number of customers on certain days and allows them to adjust supply according to the demand.

Exponential distribution is concerned with the amount of time until a specific event occurs. For example, how long a car battery would last, in months.

92. What is target imbalance? How do we fix it? A scenario where you have performed target imbalance on data. Which metrics and algorithms do you find suitable to input this data onto?

If you have categorical variables as the target when you cluster them together or perform a frequency count on them if there are certain categories which are more in number as compared to others by a very significant number. This is known as the target imbalance.

Example: Target column – 0,0,0,1,0,2,0,0,1,1 [0s: 60%, 1: 30%, 2:10%] 0 are in majority. To fix this, we can perform up-sampling or down-sampling. Before fixing this problem let's assume that the performance metrics used was confusion metrics. After fixing this problem we can shift the metric

system to AUC: ROC. Since we added/deleted data [up sampling or downsampling], we can go ahead with a stricter algorithm like SVM, Gradient boosting or ADA boosting.

93. When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

A place where the highest RSquared value is found, is the place where the line comes to rest. RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

94. Which machine learning algorithm is known as the lazy learner and why is it called so?

KNN is a Machine Learning algorithm known as a lazy learner. K-NN is a lazy learner because it doesn't learn any machine learnt values or variables from the training data but dynamically calculates distance every time it wants to classify, hence memorises the training dataset instead.

95. Is it possible to use KNN for image processing?

Yes, it is possible to use KNN for image processing. It can be done by converting the 3-dimensional image into a single-dimensional vector and using the same as input to KNN.

96. Explain the term instance-based learning.

Instance Based Learning is a set of procedures for regression and classification which produce a class label prediction based on resemblance to its nearest neighbors in the training data set. These algorithms just collects all the data and get an answer when required or queried. In simple words they are a set of procedures for solving new problems based on the solutions of already solved problems in the past which are similar to the current problem.

97. What is Bayes' Theorem? State at least 1 use case with respect to the machine learning context?

Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer than can be done without the knowledge of the person's age.

Chain rule for Bayesian probability can be used to predict the likelihood of the next word in the sentence.

98. What is Naive Bayes? Why is it Naive?

Naive Bayes classifiers are a series of classification algorithms that are based on the Bayes theorem. This family of algorithm shares a common principle which treats every pair of features independently while being classified.

Naive Bayes is considered Naive because the attributes in it (for the class) is independent of others in the same class. This lack of dependence between two attributes of the same class creates the quality of naiveness.

99. Explain the difference between Lasso and Ridge?

Lasso(L1) and Ridge(L2) are the regularization techniques where we penalize the coefficients to find the optimum solution. In ridge, the penalty function is defined by the sum of the squares of the coefficients and for the Lasso, we penalize the sum of the absolute values of the coefficients. Another type of regularization method is ElasticNet, it is a hybrid penalizing function of both lasso and ridge.

100. Why would you Prune your tree?

In the context of data science or AIML, pruning refers to the process of reducing redundant branches of a decision tree. Decision Trees are prone to overfitting, pruning the tree helps to reduce the size and minimizes the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the tradeoff.

101. Model accuracy or Model performance? Which one will you prefer and why?

This is a trick question, one should first get a clear idea, what is Model Performance? If Performance means speed, then it depends upon the nature of the application, any application related to the real-time scenario will need high speed as an

important feature. Example: The best of Search Results will lose its virtue if the Query results do not appear fast.

If Performance is hinted at Why Accuracy is not the most important virtue – For any imbalanced data set, more than Accuracy, it will be an F1 score than will explain the business case and in case data is imbalanced, then Precision and Recall will be more important than rest.

102. Mention some of the EDA Techniques?

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the foundation of better models.

Visualization

- Univariate visualization
- Bivariate visualization
- Multivariate visualization

Missing Value Treatment – Replace missing values with Either Mean/Median

Outlier Detection – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the boundary for IQR

Transformation – Based on the distribution, apply a transformation on the features

Scaling the Dataset – Apply MinMax, Standard Scaler or Z Score Scaling mechanism to scale the data.

Feature Engineering – Need of the domain, and SME knowledge helps Analyst find derivative fields which can fetch more information about the nature of the data

Dimensionality reduction — Helps in reducing the volume of data without losing much information

103. Differentiate between Statistical Modeling and Machine Learning?

Machine learning models are about making accurate predictions about the situations, like Foot Fall in restaurants, Stock-Price, etc. where-as, Statistical models are designed for inference about the relationships between variables, as What drives the sales in a restaurant, is it food or Ambience.

104. Differentiate between Boosting and Bagging?

Bagging and Boosting are variants of Ensemble Techniques.

Bootstrap Aggregation or bagging is a method that is used to reduce the variance for algorithms having very high variance. Decision trees are a particular family of classifiers which are susceptible to having high bias.

Decision trees have a lot of sensitiveness to the type of data they are trained on. Hence generalization of results is often much more complex to achieve in them despite very high fine-tuning. The results vary greatly if the training data is changed in decision trees.

Hence bagging is utilised where multiple decision trees are made which are trained on samples of the original data and the final result is the average of all these individual models.

Boosting is the process of using an n-weak classifier system for prediction such that every weak classifier compensates for the weaknesses of its classifiers. By weak classifier, we imply a classifier which performs poorly on a given data set.

It's evident that boosting is not an algorithm rather it's a

process. Weak classifiers used are generally logistic regression, shallow decision trees etc.
There are many algorithms which make use of boosting processes but two of them are mainly used: Adaboost and Gradient Boosting and XGBoost.

105. What is the significance of Gamma and Regularization in SVM?

The gamma defines influence. Low values meaning 'far' and high values meaning 'close'. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. If gamma is very small, the model is too constrained and cannot capture the complexity of the data.
The regularization parameter (lambda) serves as a degree of importance that is given to miss-classifications. This can be used to draw the tradeoff with OverFitting.

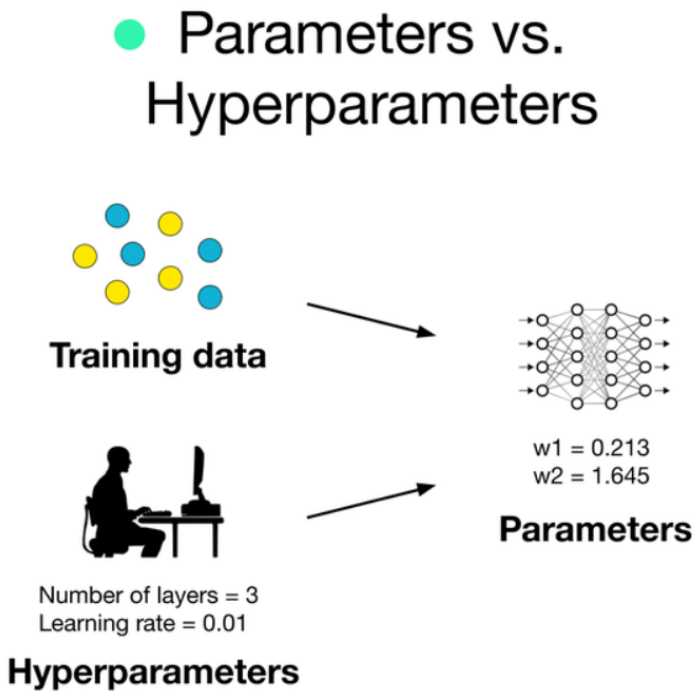
106. What is the difference between a generative and discriminative model?

A generative model learns the different categories of data. On the other hand, a discriminative model will only learn the distinctions between different categories of data. Discriminative models perform much better than the generative models when it comes to classification tasks.

107. What are hyperparameters and how are they different from parameters?

A parameter is a variable that is internal to the model and whose value is estimated from the training data. They are often saved as part of the learned model. Examples include weights, biases etc.

A hyperparameter is a variable that is external to the model whose value cannot be estimated from the data. They are often used to estimate model parameters. The choice of parameters is sensitive to implementation. Examples include learning rate, hidden layers etc.



108. Can logistic regression be used for classes more than 2?

No, logistic regression cannot be used for classes more than 2 as it is a binary classifier. For multi-class classification algorithms like Decision Trees, Naïve Bayes' Classifiers are better suited.

109. How to deal with multicollinearity?

Multi collinearity can be dealt with by the following steps:

- Remove highly correlated predictors from the model.
- Use Partial Least Squares Regression (PLS) or Principal Components Analysis

110. What is Heteroscedasticity?

It is a situation in which the variance of a variable is unequal across the range of values of the predictor variable.
It should be avoided in regression as it introduces unnecessary variance.

111. Is ARIMA model a good fit for every time series problem?

No, ARIMA model is not suitable for every type of time series problem. There are situations where ARMA model and others also come in handy.
ARIMA is best when different standard temporal structures require to be captured for time series data.

112. What is a voting model?

A voting model is an ensemble model which combines several classifiers but to produce the final result, in case of a classification-based model, takes into account, the classification of a certain data point of all the models and picks the most vouched/voted/generated option from all the given classes in the target column.

113. How to deal with very few data samples? Is it possible to make a model out of it?

If very few data samples are there, we can make use of oversampling to produce new data points. In this way, we can have new data points.

114. What is Pandas Profiling?

Pandas profiling is a step to find the effective number of usable data. It gives us the statistics of NULL values and the usable values and thus makes variable selection and data selection for building models in the preprocessing phase very effective.

115. When should ridge regression be preferred over lasso?

We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

116. What is a good metric for measuring the level of multicollinearity?

VIF or 1/tolerance is a good measure of measuring multicollinearity in models. VIF is the percentage of the variance of a predictor which remains unaffected by other predictors. So higher the VIF value, greater is the multicollinearity amongst the predictors.

A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

117. When can be a categorical value treated as a continuous variable and what effect does it have when done so?

A categorical predictor can be treated as a continuous one when the nature of data points it represents is ordinal. If the predictor variable is having ordinal data then it can be treated as continuous and its inclusion in the model increases the performance of the model.

118. What is the role of maximum likelihood in logistic regression.

Maximum likelihood equation helps in estimation of most probable values of the estimator's predictor variable coefficients which produces results which are the most likely or most probable and are quite close to the truth values.

119. What is a pipeline?

A pipeline is a sophisticated way of writing software such that each intended action while building a model can be serialized and the process calls the individual functions for the individual tasks. The tasks are carried out in sequence for a given sequence of data points and the entire process can be run onto n threads by use of composite estimators in scikit learn.

120. What do you understand by L1 and L2 regularization?

L2 regularization: It tries to spread error among all the terms. L2 corresponds to a Gaussian prior.

L1 regularization: It is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms.

121. What do you mean by AUC curve?

AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

122. Why does XGBoost perform better than SVM?

First reason is that XGBoos is an ensemble method that uses many trees to make a decision so it gains power by repeating itself.

SVM is a linear separator, when data is not linearly separable SVM needs a Kernel to project the data into a space where it can separate it, there lies its greatest strength and weakness, by being able to project data into a high dimensional space SVM can find a linear separation for

almost any data but at the same time it needs to use a Kernel and we can argue that there's not a perfect kernel for every dataset.

123. What is the difference between SVM Rank and SVR (Support Vector Regression)?

One is used for ranking and the other is used for regression.
There is a crucial difference between regression and ranking. In regression, the absolute value is crucial. A real number is predicted.
In ranking, the only thing of concern is the ordering of a set of examples. We only want to know which example has the highest rank, which one has the second-highest, and so on. From the data, we only know that example 1 should be ranked higher than example 2, which in turn should be ranked higher than example 3, and so on. We do not know by how much example 1 is ranked higher than example 2, or whether this difference is bigger than the difference between examples 2 and 3.

124. What is the difference between the normal soft margin SVM and SVM with a linear kernel?

Hard-margin
You have the basic SVM – hard margin. This assumes that data is very well behaved, and you can find a perfect classifier – which will have 0 error on train data.
Soft-margin
Data is usually not well behaved, so SVM hard margins may not have a solution at all. So we allow for a little bit of error

on some points. So the training error will not be 0, but average error over all points is minimized.

Kernels

The above assume that the best classifier is a straight line. But what is it is not a straight line. (e.g. it is a circle, inside a circle is one class, outside is another class). If we are able to map the data into higher dimensions – the higher dimension may give us a straight line.

125. What are the advantages of using a naive Bayes for classification?

- Very simple, easy to implement and fast.
- If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
- Even if the NB assumption doesn't hold, it works great in practice.
- Need less training data.
- Highly scalable. It scales linearly with the number of predictors and data points.
- Can be used for both binary and mult-iclass classification problems.
- Can make probabilistic predictions.
- Handles continuous and discrete data.
- Not sensitive to irrelevant features.

126. Are Gaussian Naive Bayes the same as binomial Naive Bayes?

Binomial Naive Bayes: It assumes that all our features are binary such that they take only two values. Means 0s can represent "word does not occur in the document" and 1s as "word occurs in the document".

Gaussian Naive Bayes: Because of the assumption of the normal distribution, Gaussian Naive Bayes is used in cases when all our features are continuous. For example in Iris dataset features are sepal width, petal width, sepal length, petal length. So its features can have different values in the data set as width and length can vary. We can't represent features in terms of their occurrences. This means data is continuous. Hence we use Gaussian Naive Bayes here.

127. What is the difference between the Naive Bayes Classifier and the Bayes classifier?

Naive Bayes assumes conditional independence, $P(X|Y, Z)=P(X|Z)$

$P(X|Y,Z)=P(X|Z)$

$P(X|Y,Z)=P(X|Z)$, Whereas more general Bayes Nets (sometimes called Bayesian Belief Networks), will allow the user to specify which attributes are, in fact, conditionally independent.

For the Bayesian network as a classifier, the features are selected based on some scoring functions like Bayesian scoring function and minimal description length(the two are equivalent in theory to each other given that there is enough training data). The scoring functions mainly restrict the

structure (connections and directions) and the parameters(likelihood) using the data. After the structure has been learned the class is only determined by the nodes in the Markov blanket(its parents, its children, and the parents of its children), and all variables given the Markov blanket are discarded.

128. In what real world applications is Naive Bayes classifier used?

Some of real world examples are as given below

- To mark an email as spam, or not spam?
- Classify a news article about technology, politics, or sports?
- Check a piece of text expressing positive emotions, or negative emotions?
- Also used for face recognition software

129. Is naive Bayes supervised or unsupervised?

First, Naive Bayes is not one algorithm but a family of Algorithms that inherits the following attributes:

1.Discriminant Functions

2.Probabilistic Generative Models

3.Bayesian Theorem

4.Naive Assumptions of Independence and Equal Importance of feature vectors.

Moreover, it is a special type of Supervised Learning algorithm that could do simultaneous multi-class predictions (as depicted by standing topics in many news apps).

Since these are generative models, so based upon the assumptions of the random variable mapping of each feature vector these may even be classified as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, etc.

130. What do you understand by selection bias in Machine Learning?

Selection bias stands for the bias which was introduced by the selection of individuals, groups or data for doing analysis in a way that the proper randomization is not achieved. It ensures that the sample obtained is not representative of the population intended to be analyzed and sometimes it is referred to as the selection effect. This is the part of distortion of a statistical analysis which results from the method of collecting samples. If you don't take the selection bias into the account then some conclusions of the study may not be accurate.

The types of selection bias includes:

- Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
- Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

- Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
- Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

131. What Are the Three Stages of Building a Model in Machine Learning?

To build a model in machine learning, you need to follow few steps:

1. Understand the business model
2. Data acquisitions
3. Data cleaning
4. Exploratory data analysis
5. Use machine learning algorithms to make a model
6. Use unknown dataset to check the accuracy of the model

132. What is the difference between Entropy and Information Gain?

The **information gain** is based on the decrease in **entropy** after a dataset is split on an attribute. Constructing a decision tree is all about finding the attribute that returns the highest **information gain** (i.e., the most homogeneous branches). Step 1: Calculate **entropy** of the target.

133. What are collinearity and multicollinearity?

Collinearity is a linear association **between** two predictors. **Multicollinearity** is a situation where two or more predictors are highly linearly related.

134. What is A/B Testing?

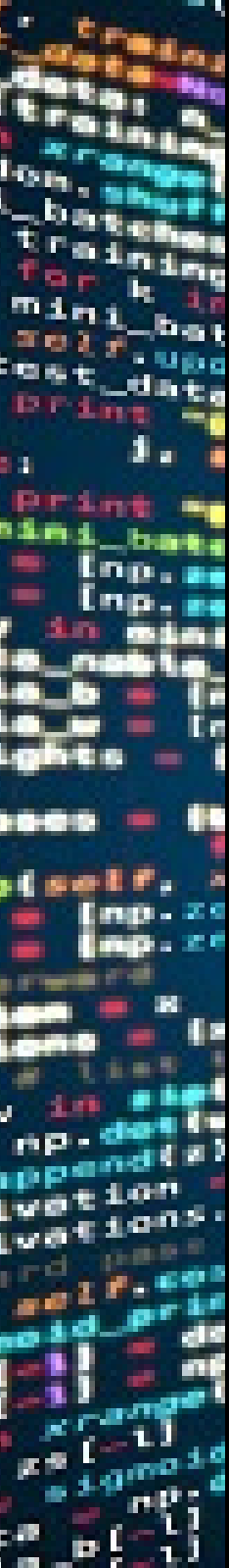
- A/B is Statistical hypothesis testing for randomized experiment with two variables A and B. It is used to compare two models that use different predictor variables in order to check which variable fits best for a given sample of data.
- Consider a scenario where you've created two models (using different predictor variables) that can be used to recommend products for an e-commerce platform.
- A/B Testing can be used to compare these two models to check which one best recommends products to a customer.

135. What is Cluster Sampling?

- It is a process of randomly selecting intact groups within a defined population, sharing similar characteristics.
- Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.
- For example, if you're clustering the total number of managers in a set of companies, in that case, managers (samples) will represent elements and companies will represent clusters.

136. What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data. In that sense, deep learning represents an unsupervised learning algorithm that learns representations of data through the use of neural nets.



This brings our list of 120+ Machine Learning interview questions to an end.

We believe these series of guides will help you “expect the unexpected” and enter your first data analytics interview with confidence.

We, at Zep provide a platform for Education, where your demand gets fulfilled. You demand we fulfil all your learning needs without costing you extra.

zepanalytics.com

Ready to take the next steps?

Zep offers a platform for education to learn, grow & earn.

Become a part of the team at Zep

Why don't you start your journey as a tech blogger and enjoy unlimited perks and cash prizes every month.

Explore

zepanalytics.com