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.

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.

Overfitting

occurs when a machine learning model learns the training data too well, capturing noise and specific patterns that do not generalize to new, unseen data.

High Variance: The model performs well on the training data but poorly on the test data, indicating it has learned specific details of the training set that don't apply generally.

Underfitting

Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in the data, resulting in poor performance on both the training and test datasets. (too simple to explain the variance)

High Bias: The model makes large errors on both the training and test datasets because it can't capture the data's complexity.

Training Set	Test 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	• 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

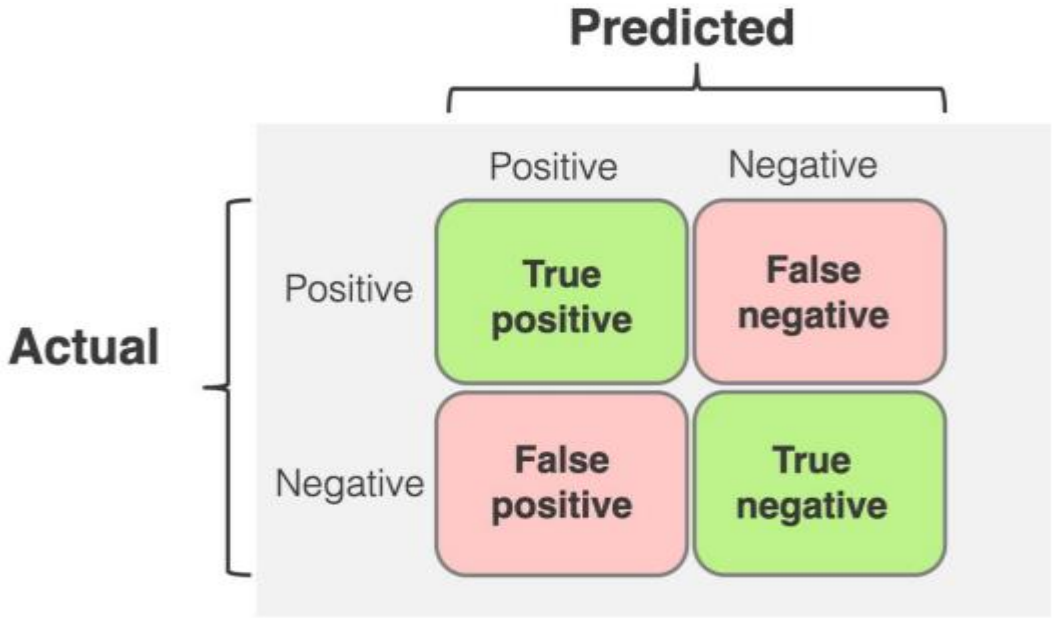
Check Null

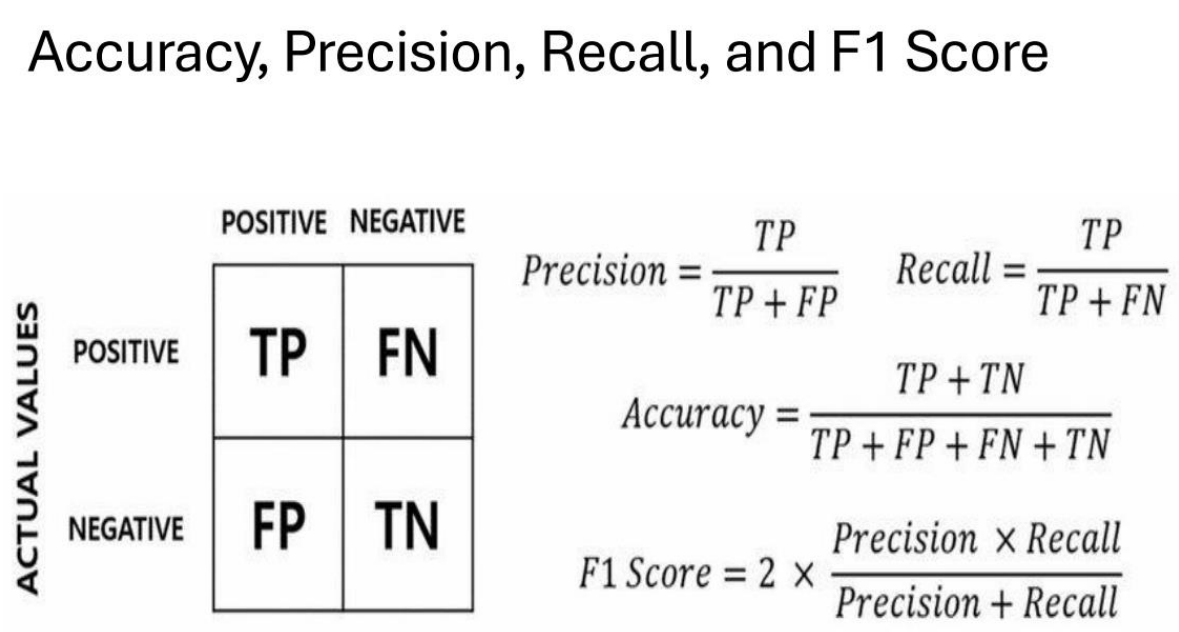
- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

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.

Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It provides insight into the true positives, false positives, true negatives, and false negatives made by the model, helping to calculate metrics like accuracy, precision, recall, and F1 score





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). If precision is low it means then your model has too many false positives

A **recall** is the ratio of the number of events you can recall the number of total events. If recall is low it means then your model has too many false negatives

The **F1 score** is a metric that combines both Precision and Recall. It is also the weighted average of precision and recall.

7. 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.

Example: In a medical test for a disease, a false positive would mean that the model predicts a person has the disease (positive) when they do not (negative).

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

Example: In the same medical test scenario, a false negative would mean the model predicts a person does not have the disease (negative) when they actually do (positive).

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

- 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

Machine Learning	Deep Learning
• Enables machines to take decisions on their own, based on past data • It needs only a small amount of data for training • Works well on the low-end system, so you don't need large machines • Most features need to be identified in advance and manually coded • The problem is divided into two parts and solved individually and then combined	• Enables machines to take decisions with the help of artificial neural networks • It needs a large amount of training data • Needs high-end machines because it requires a lot of computing power • The machine learns the features from the data it is provided • The problem is solved in an end-to-end manner

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

Applications of supervised machine learning include:

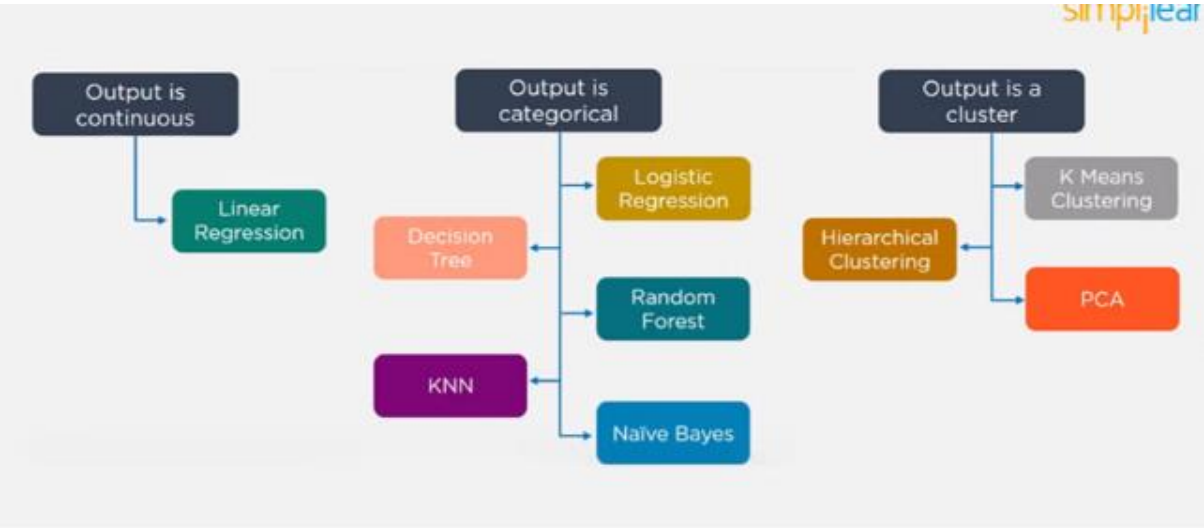
- Email Spam Detection
- Healthcare Diagnosis
- Fraud Detection
- Sentiment Analysis

14. 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.

21. 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.



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

Bias

Bias in a machine learning model occurs when the predicted values are far from the actual values. **High Bias** leads to **underfitting**, as the model makes strong assumptions and fails to capture the complexity of the data.

Example: A linear model trying to fit non-linear data will have high bias, as it's too simple to capture the underlying patterns.

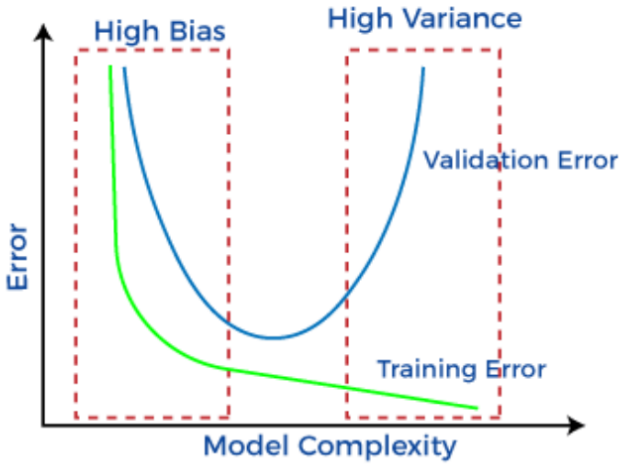
Variance

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

Example: A very deep decision tree might perform perfectly on the training data but fail on new data,

Bias-Variance Tradeoff

- ***Bias:*** The error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to underfit the data.
- ***Variance:*** The error due to excessive sensitivity to small fluctuations in the training data. High variance can cause the model to overfit the data.
- ***Tradeoff:*** There's a balance between bias and variance. Reducing one typically increases the other, and the goal is to find the right balance for the best model performance.



Interpretation of the Graph:

On the left side (High Bias): Both training and validation errors are high because the model is too simple (underfitting).

In the middle (Optimal): The validation error is at its lowest point, where the model is complex enough to capture data patterns but not so complex that it overfits.

On the right side (High Variance): Training error is very low (the model fits the training data well), but validation error is high because of overfitting.

The image conveys that model complexity needs to be carefully balanced to avoid high bias or high variance. Understanding this tradeoff is crucial when tuning a model, as it helps decide whether to increase model complexity or simplify it to achieve the best generalization on new data.

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.

28. What is a Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure that shows a map of possible outcomes of related decisions, with datasets broken up into 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.

improves predictive accuracy by the reduction of overfitting.

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.

KNN

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 the new data point to be classified is a black ball. We use KNN to classify it. Assume K = 5

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.

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

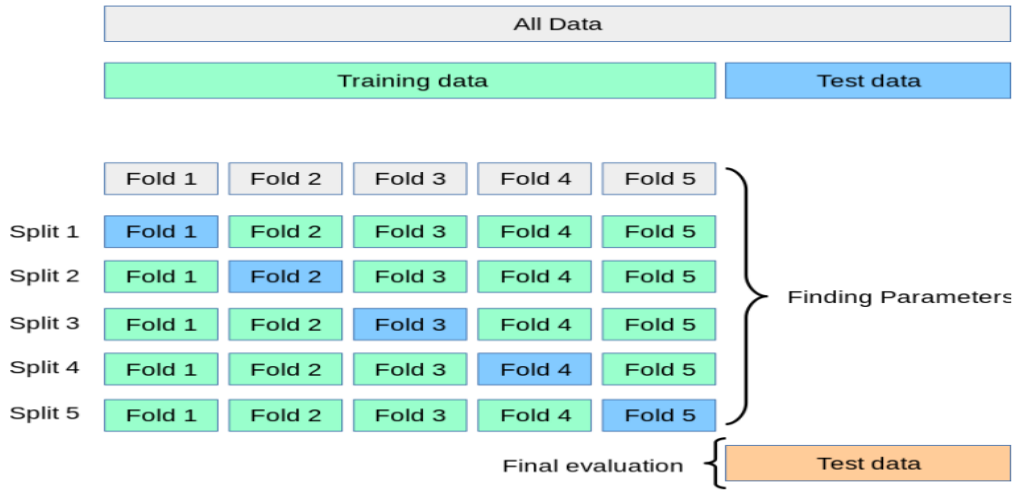
Type I Error: Type I error occurs when the null hypothesis is true and we reject it.

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

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.

Cross Validation



42. 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.

44. 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

Linear Regression is a supervised learning algorithm used for predicting a continuous numerical value based on one or more input features. It assumes a linear relationship between the input variables (independent variables) and the output variable (dependent variable). The goal of linear regression is to find the line (or hyperplane in multiple dimensions) that best fits the data points, minimizing the difference between the predicted and actual values.

When to Use Naïve Bayes Classifier:

Text Classification

Multiclass Classification:

Features are conditionally independent

When Not to Use Naive Bayes Classifier?

- **Strongly correlated features:**
- **Complex decision boundaries:**
- **Continuous Variables without discretization**

When to use decision trees:
Decision trees are ideal when you need an interpretable model that can handle both categorical and numerical data, and when the data might have non-linear relationships can capture complex patterns (**Handling Missing Values)**:.

Pros and Cons of Decision Trees:
Decision trees are easy to interpret and visualize, require little preprocessing, and can capture non-linear relationships well. However, they are prone to overfitting, sensitive to small changes in data, can be biased with imbalanced datasets, and generally perform less accurately than ensemble methods like Random Forests or Gradient Boosted Trees.