1. **What is marginal probability**

It is a probability of one event occurred calculation by ignoring all other variables.





1. **What are the probability axioms?**

It is called rules or principles of probability theory, has three rules





This calculation will used in **Naive Bayes** where naïve mean assumption, so in **Naive Bayes** is a **probabilistic classifier** based on **Bayes' Theorem** that assumes features are independent — it predicts the class with the highest probability.

1. **What is Conditional probability ?**

Conditional probability is the chance of event **A** happening **if** event **B** has already happened — like the chance it’s raining (**A**) given that it’s cloudy (**B**).

**Example**: If 30 out of 50 cloudy days are rainy, then  
**P(Rain | Cloudy) = 30 / 50 = 0.6**.

1. **What is Bayes’ Theorem and when is it used in data science?**

Bayes' Theorem helps us find the probability of a cause given an observed result. For example, if a person tests positive for a disease, Bayes' Theorem helps calculate how likely they truly have it, considering the test’s accuracy and disease rate. It's useful in areas like medical diagnosis, spam detection, and predictive modeling.

 Bayes' Theorem = Formula for conditional probability

 Naive Bayes = A classification model built using Bayes’ Theorem

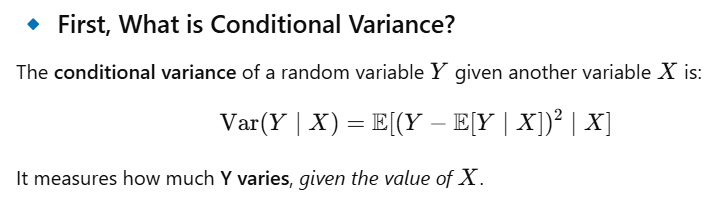
Naive Bayes is widely used for spam detection, sentiment analysis, and text classification.

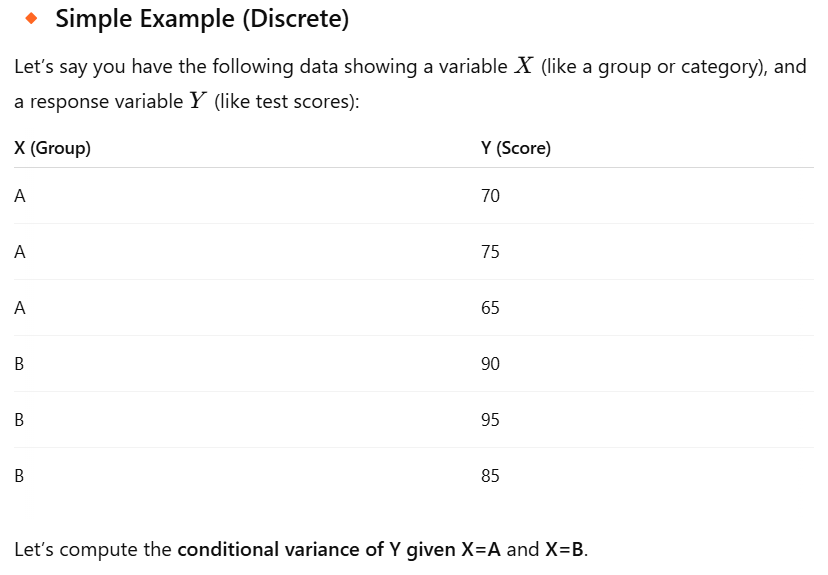
1. **Define variance and conditional variance.**

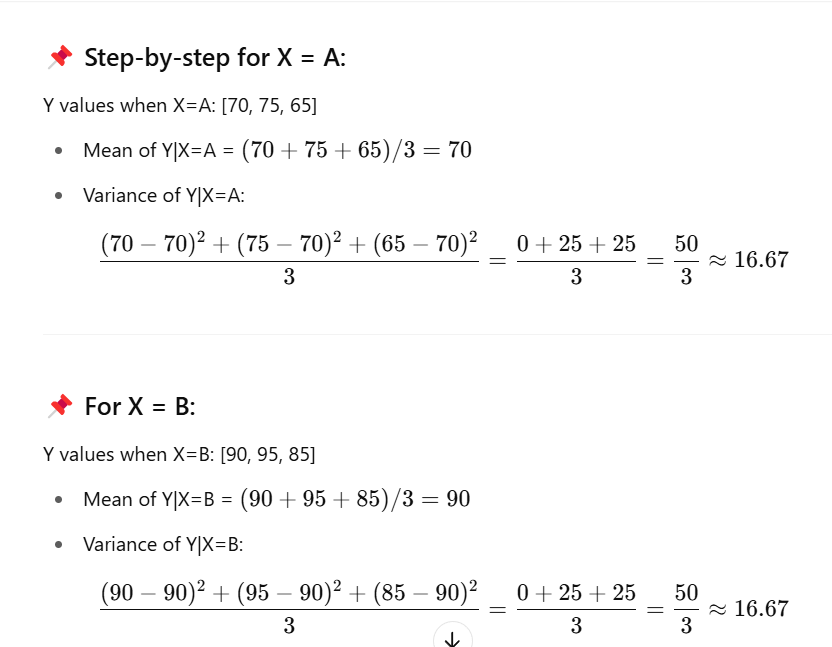
**Variance**: It shows how spread out numbers are from the average — for example, in the numbers [2, 4, 6], the variance is small because they’re close to the mean (4).

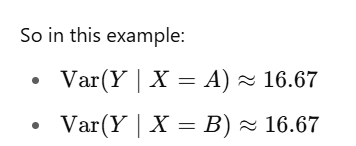
**Conditional Variance**: It’s the variance of a variable given some condition — for example, if we know it’s a rainy day, the variance of people carrying umbrellas may be less than on random days.

**EX**:







`

1. **Explain the concepts of mean, median, mode, and standard deviation.**

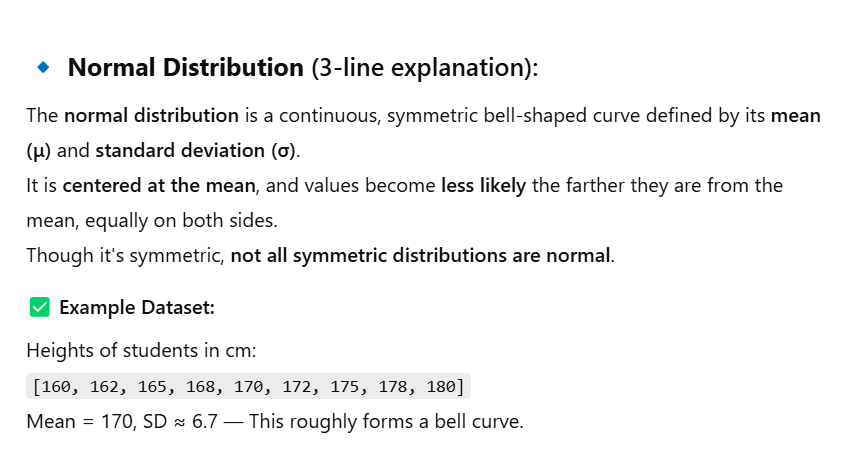
**Mean:**The mean, often referred to as the average, is calculated by summing up all the values in a dataset and then dividing by the total number of values.

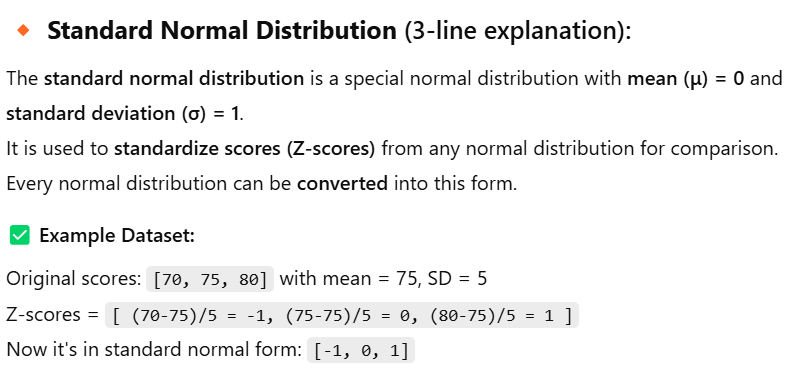
**Median:**When data are sorted in either ascending or descending order, the median is the value in the middle of the dataset. The median is the average of the two middle values when the number of data points is even. In comparison to the mean, the median is less impacted by extreme numbers, making it a more reliable indicator of central tendency.

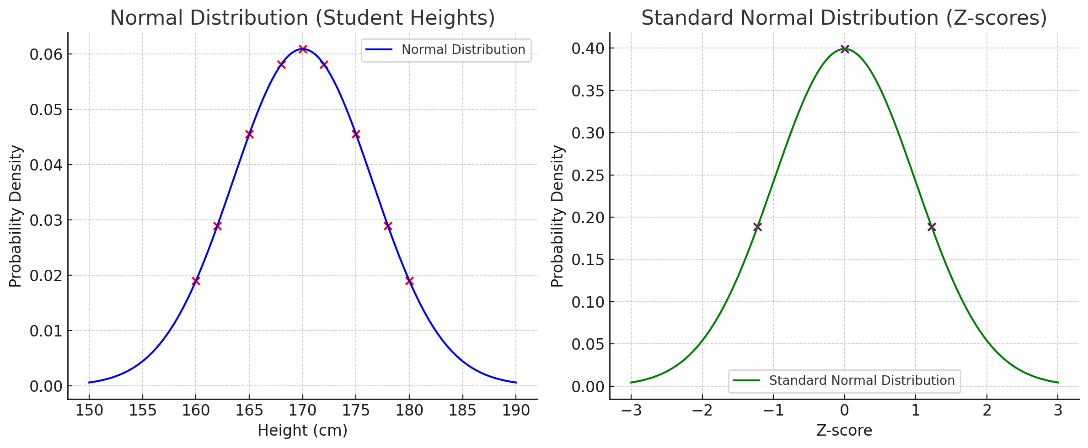
**Mode:** The value that appears most frequently in a dataset is the mode. One mode (unimodal), several modes (multimodal), or no mode (if all values occur with the same frequency) can all exist in a dataset.

**Standard deviation:** The spread or dispersion of data points in a dataset is measured by the standard deviation. It quantifies the variance between different data points

1. **Difference between normal distribution and standard normal distribution**

****

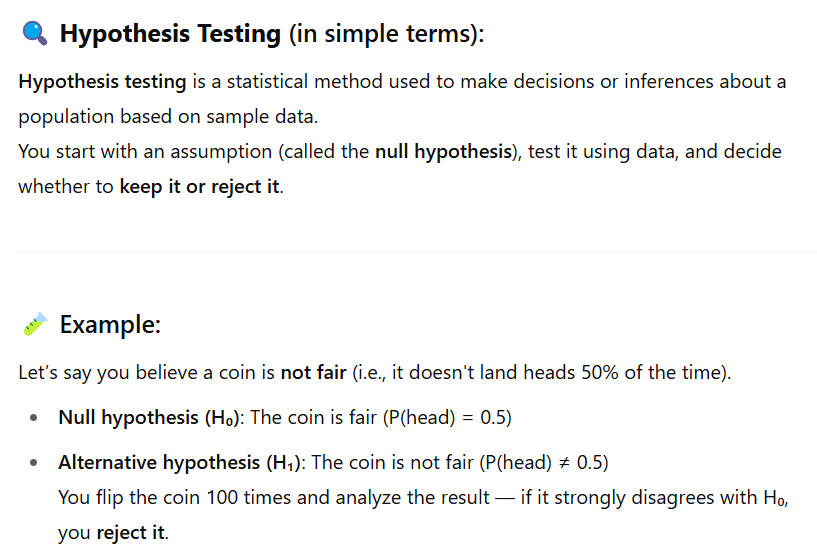
****

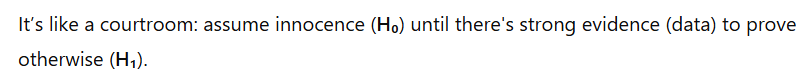


we use normal and standard normal distributions primarily to understand how data is distributed,

And also it has the following usage too.

The **normal distribution** models real-world data, estimates probabilities, supports hypothesis testing, and underpins the Central Limit Theorem.  
The **standard normal distribution** standardizes data for easy comparison, simplifies probability calculations, detects outliers, and improves machine learning performance.





1. **What is SQL, and what does it stand for?**

SQL stands for Structured Query Language.It is a specialized programming language used for managing and manipulating relational databases. It is designed for tasks related to database management, data retrieval, data manipulation, and data definition.

1. **SQL vs NoSQL Databases – Key Differences**

* Data Structure:
  + SQL: Uses structured, relational data (tables with rows and columns).
  + NoSQL: Uses flexible formats like documents (JSON/BSON), key-value pairs, columns, or graphs.
* Schema:
  + SQL: Has a fixed schema; structure must be defined before data is added.
  + NoSQL: Schema-less or dynamic; data can be added without a predefined structure.
* Query Language:
  + SQL: Uses standard SQL language; supports complex queries (joins, subqueries, aggregations).
  + NoSQL: Uses varied, model-specific query languages or APIs.
* Use Cases:
  + SQL: Best for structured data and complex queries.
  + NoSQL: Suitable for unstructured data, scalability, and flexible application needs.

**10. What are the primary SQL database management systems (DBMS)?** Relational database systems, both open source and commercial, are the main SQL (Structured Query Language) database management systems (DBMS), which are widely used for managing and processing structured data. Some of the most popular SQL database management systems are listed below: MySQL Microsoft SQL Server SQLite PostgreSQL Oracle Database 6. Amazon RDS.

**11. What is the ER model in SQL?**

The structure and relationships between the data entities in a database are represented by the Entity-Relationship (ER) model, a conceptual framework used in database architecture. The ER model is frequently used in conjunction with SQL for creating the structure of relational databases even though it is not a component of the SQL language itself.

**12 What is data transformation?**

The process of transforming data from one structure, format, or representation into another is referred to as data transformation. In order to make the data more suited for a given goal, such as analysis, visualisation, reporting, or storage, this procedure may involve a variety of actions and changes to the data. Data integration, cleansing, and analysis depend heavily on data transformation, which is a common stage in data preparation and processing pipelines.

**13 What are the main components of a SQL query?**

A relational database’s data can be retrieved, modified, or managed via a SQL (Structured Query Language) query. The operation of a SQL query is defined by a number of essential components, each of which serves a different function. 1. SELECT 2. FROM 3. WHERE 4. GROUP BY 5. HAVING 6. ORDER BY 7. LIMIT 8. JOIN

**14 What is a primary key?**

A relational database table’s main key, also known as a primary keyword, is a column that is unique for each record. It is a distinctive identifier. The primary key of a relational database must be unique. Every row of data must have a primary key value and none of the rows can be null

**15. What is the purpose of the GROUP BY clause, and how is it used?**

In SQL, the GROUP BY clause is used to create summary rows out of rows that have the same values in a set of specified columns. In order to do computations on groups of rows as opposed to individual rows, it is frequently used in conjunction with aggregate functions like SUM, COUNT, AVG, MAX, or MIN. we may produce summary reports and perform more in-depth data analysis using the GROUP BY clause.

**16. What is the WHERE clause used for, and how is it used to filter data?**

In SQL, the WHERE clause is used to filter rows from a table or result set according to predetermined criteria. It enables us to pick only the rows that satisfy particular requirements or follow a pattern. A key element of SQL queries, the WHERE clause is frequently used for data retrieval and manipulation.

**17. How do you retrieve distinct values from a column in SQL?**

Using the DISTINCT keyword in combination with the SELECT command, we can extract distinct values from a column in SQL. By filtering out duplicate values and returning only unique values from the specified column, the DISTINCT keyword is used

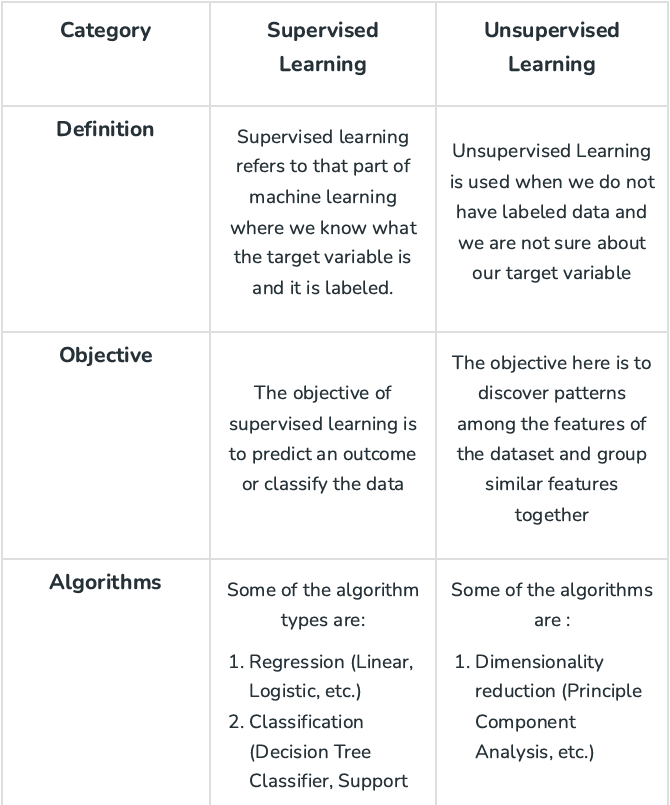
**18. What is the HAVING clause?**

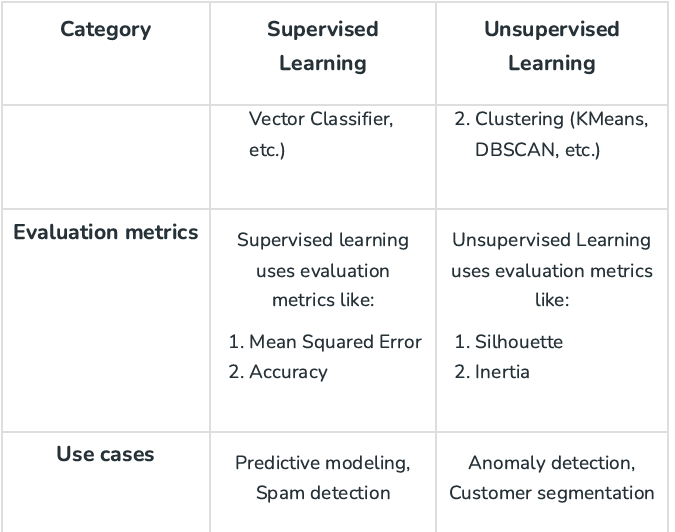
To filter query results depending on the output of aggregation functions, the HAVING clause, a SQL clause, is used along with the GROUP BY clause. The HAVING clause filters groups of rows after they have been grouped by one or more columns, in contrast to the WHERE clause, which filters rows before they are grouped

**19. How do you handle missing or NULL values in a database table?**

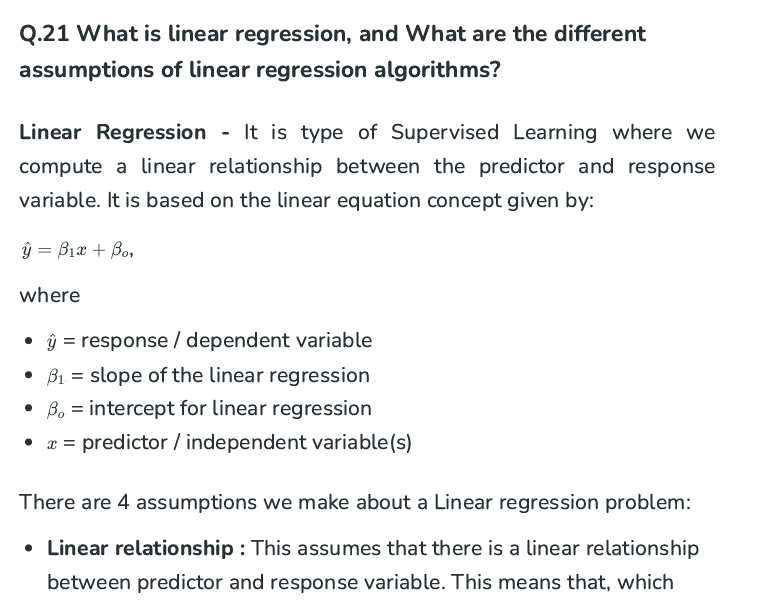
Missing or NULL values can arise due to various reasons, such as incomplete data entry, optional fields, or data extraction processes. 1. Replace NULL with Placeholder Values 2. Handle NULL Values in Queries 3. Use Default Value

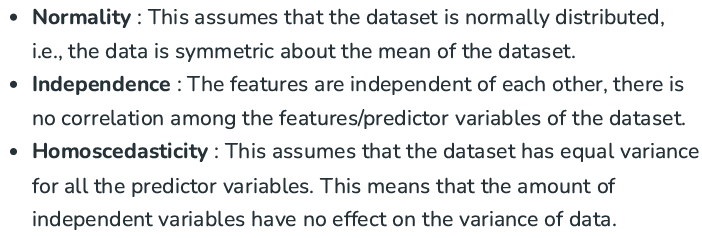
**20 What is the difference between supervised and unsupervised machine learning?**





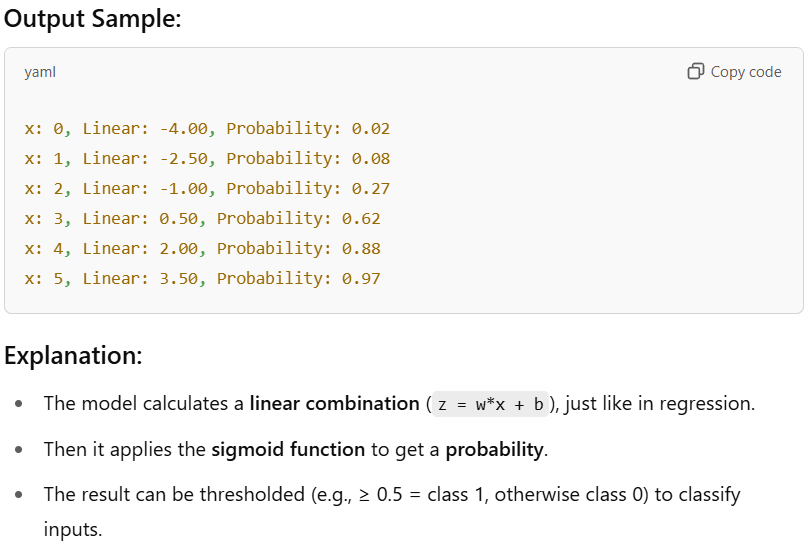
**21. What is linear regression, and What are the different assumptions of linear regression algorithms**

****

****

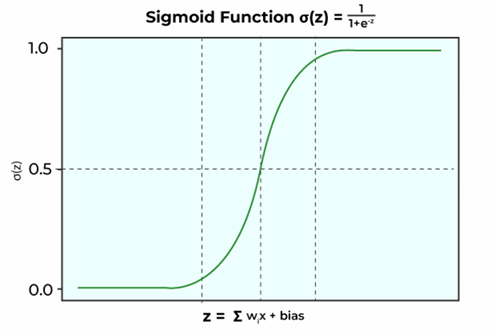
**28. Logistic regression is a classification technique, why its name is regressions, not logistic classifications**

**Logistic regression** is named so because it uses a **regression-like structure** to model the **probability** of a class. It applies a **logistic (sigmoid) function** to a linear combination of input features to produce a probability between 0 and 1. Although used for **classification**, it retains the underlying **regression framework**, which is why it's called "logistic regression" instead of "logistic classification.



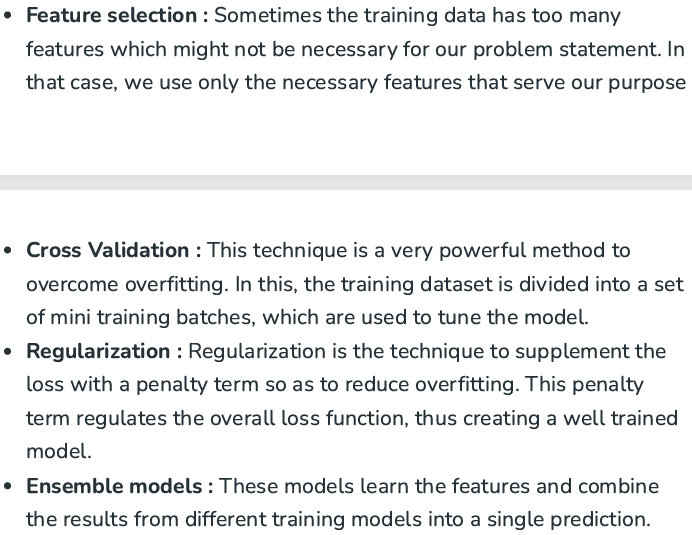
**29. What is the logistic function (sigmoid function) in logistic regression?**

Sigmoid Function: It is a mathematical function which is characterized by its S- shape curve. Sigmoid functions have the tendency to squash a data point to lie within 0 and 1. This is why it is also called Squashing function, which is given as: \sigma(x) = \frac{1}{1+e^{-x}



**30. What is overfitting and how can be overcome this?**

Overfitting refers to the result of analysis of a dataset which fits so closely with training data that it fails to generalize with unseen/future data. This happens when the model is trained with noisy data which causes it to learn the noisy features from the training as well. To avoid Overfitting and overcome this problem in machine learning, one can follow the following rules:



**31.** **What is a support vector machine (SVM), and what are its key components?**

Support Vector machines are a type of Supervised algorithm which can be used for both Regression and Classification problems. In SVMs, the main goal is to find a hyperplane which will be used to segregate different data points into classes. Any new data point will be classified based on this defined hyperplane. Support Vector machines are highly effective when dealing with high dimensionality space and can handle non linear data very well. But if the number of features are greater than number of data samples, it is susceptible to overfitting.

The key components of SVM are:

Kernels Function: It is a mapping function used for data points to convert it into high dimensionality feature space.

Hyperplane: It is the decision boundary which is used to differentiate between the classes of data points.

Margin: It is the distance between Support Vector and Hyperplane

C: It is a regularization parameter which is used for margin maximization and misclassification minimization