1. Explain logistic regression.

Logistic Regression in Machine Learning

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
* Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



**Note:** Logistic regression uses the concept of predictive modelling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Logistic Function (Sigmoid Function):

* The sigmoid function is a mathematical function used to map the predicted values to probabilities.
* It maps any real value into another value within a range of 0 and 1.
* The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
* In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

* The dependent variable must be categorical in nature.
* The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

* We know the equation of the straight line can be written as:

Logistic Regression in Machine Learning

* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

Logistic Regression in Machine Learning

* But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

Logistic Regression in Machine Learning

The above equation is the final equation for Logistic Regression.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

* **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
* **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
* **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Advantages

**Easy to Understand:**

Simple to use and easy to interpret results.

You can see the influence of each feature on the outcome.

**Efficient**:

Quick to train and works well with smaller datasets.

It calculates results quickly, making it good for real-time use.

**Probabilities:**

Provides probabilities for each outcome, which helps in understanding the confidence of predictions.

**No Scaling Needed:**

Generally, works without needing to adjust the scale of the input features.

**Handles Linear Data Well:**

Performs well if the relationship between the features and the outcome is linear.

**Prevents Overfitting:**

Can use techniques to avoid overfitting (making it fit too closely to the training data).

Disadvantages

**Linear Assumption:**

Assumes a straight-line relationship, which might not fit all types of data well.

**Binary Focus:**

Primarily designed for two-class problems; handling more than two classes is more complex.

**Sensitive to Outliers:**

Can be affected by extreme values, which may skew results.

**Needs Large Datasets:**

Works best with larger amounts of data to make reliable predictions.

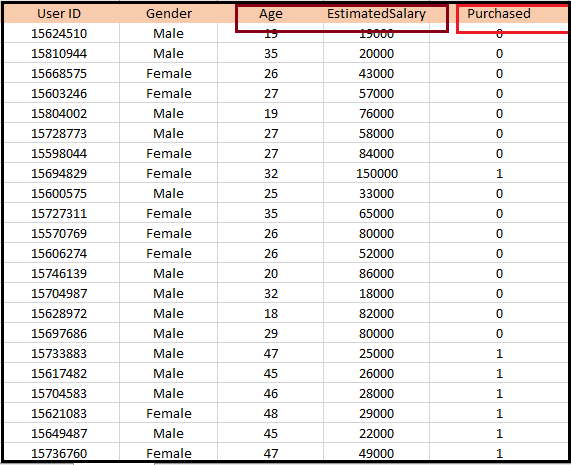
**Feature Independence:**

Assumes features don’t influence each other, which isn’t always true.

**Limited for Complex Patterns:**

Not the best for data with complex patterns and interactions between features; other models might work better for such cases.

**Example:** There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So, the company wanted to check how many users from the dataset, wants to purchase the car.

For this problem, we will build a Machine Learning model using the Logistic regression algorithm. The dataset is shown in the below image. In this problem, we will predict the **purchased variable (Dependent Variable)** by using **age and salary (Independent variables)**.

**7.What is Over fitting &Under fitting**

Overfitting and Underfitting in Machine Learning

* Overfitting and Underfitting are the two main problems that occur in machine learning and degrade the performance of the machine learning models.
* The main goal of each machine learning model is **to generalize well**. Here **generalization** defines the ability of an ML model to provide a suitable output by adapting the given set of unknown input.
* It means after providing training on the dataset, it can produce reliable and accurate output.
* Hence, the underfitting and overfitting are the two terms that need to be checked for the performance of the model and whether the model is generalizing well or not.

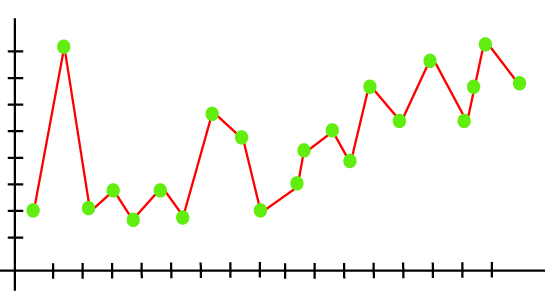
Before understanding the overfitting and underfitting, let's understand some basic term that will help to understand this topic well:

* **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.
* **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.
* **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.
* **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

Overfitting

* Overfitting occurs when our [machine learning](https://www.javatpoint.com/machine-learning) model tries to cover all the data points or more than the required data points present in the given dataset.
* Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.
* The overfitted model has **low bias** and **high variance.**
* The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.
* Overfitting is the main problem that occurs in [supervised learning](https://www.javatpoint.com/supervised-machine-learning).

**Example:** The concept of the overfitting can be understood by the below graph of the linear regression output:



As we can see from the above graph, the model tries to cover all the data points present in the scatter plot. It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

How to avoid the Overfitting in Model

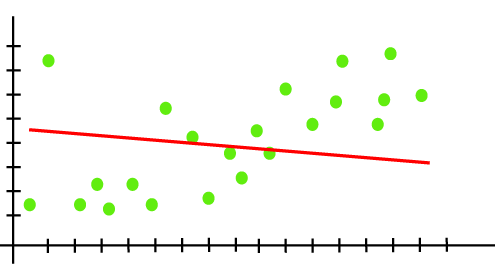
Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.

* **Cross-Validation**
* **Training with more data**
* **Removing features**
* **Early stopping the training**
* **Regularization**
* **Ensembling**

Underfitting

* Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data.
* To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.
* In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.
* An underfitted model has high bias and low variance.

**Example:** We can understand the underfitting using below output of the linear regression model:



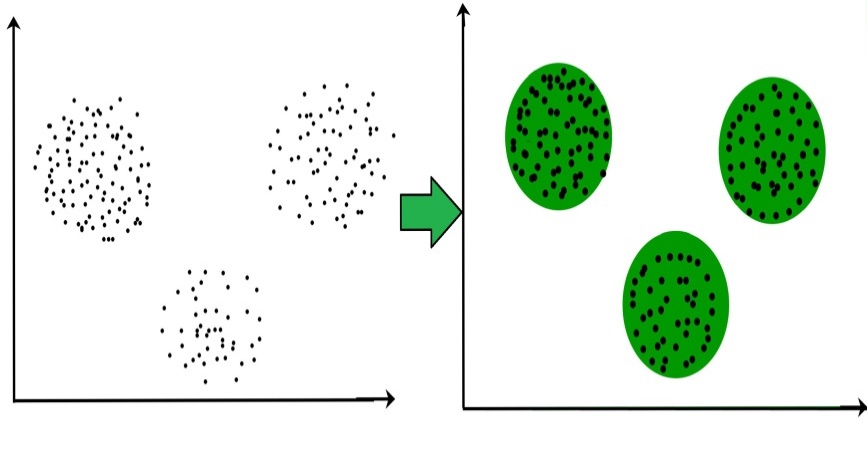
As we can see from the above diagram, the model is unable to capture the data points present in the plot.

How to avoid underfitting:

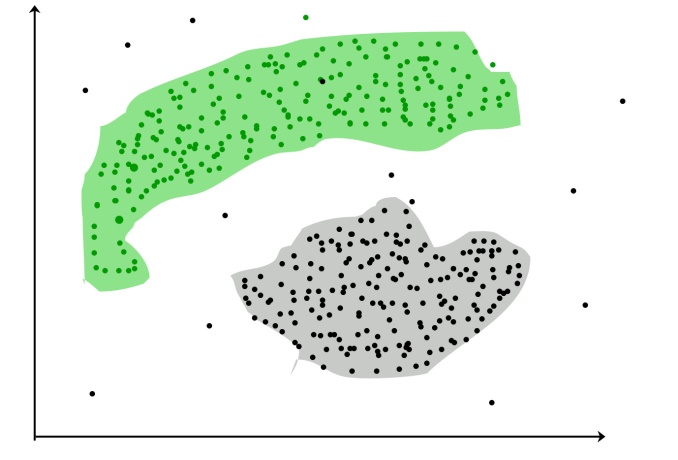
* By increasing the training time of the model.
* By increasing the number of features.

**1. What is clustering? Explain in detail.**

* The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis.
* This method is defined under the branch of [Unsupervised Learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/), which aims at gaining insights from unlabelled data points, that is, unlike [supervised learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) we don’t have a target variable.
* Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset.
* It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc. and then group the points with highest similarity score together.
* For Example, In the graph given below, we can clearly see that there are 3 circular clusters forming on the basis of distance.



* Now it is not necessary that the clusters formed must be circular in shape. The shape of clusters can be arbitrary.
* There are many algorithms that work well with detecting arbitrary shaped clusters.
* For example, In the below given graph we can see that the clusters formed are not circular in shape.



**Types of Clustering**

Broadly speaking, there are 2 types of clustering that can be performed to group similar data points:

* **Hard Clustering:**In this type of clustering, each data point belongs to a cluster completely or not. For example, Let’s say there are 4 data point and we have to cluster them into 2 clusters. So, each data point will either belong to cluster 1 or cluster 2.

| **Data Points** | **Clusters** |
| --- | --- |
| A | C1 |
| B | C2 |
| C | C2 |
| D | C1 |

* **Soft Clustering:**In this type of clustering, instead of assigning each data point into a separate cluster, a probability or likelihood of that point being that cluster is evaluated.
* For example, Let’s say there are 4 data point and we have to cluster them into 2 clusters. So, we will be evaluating a probability of a data point belonging to both clusters. This probability is calculated for all data points.

| Data Points | Probability of C1 | Probability of C2 |
| --- | --- | --- |
| A | 0.91 | 0.09 |
| B | 0.3 | 0.7 |
| C | 0.17 | 0.83 |
| D | 1 | 0 |

**Uses of Clustering**

Now before we begin with types of clustering algorithms, we will go through the use cases of Clustering algorithms. Clustering algorithms are majorly used for:

* [**Market Segmentation**](https://www.geeksforgeeks.org/customer-segmentation-using-unsupervised-machine-learning-in-python/)– Businesses use clustering to group their customers and use targeted advertisements to attract more audience.

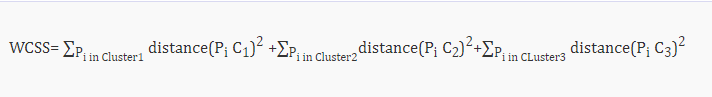
1. [**Market Basket Analysis**](https://www.geeksforgeeks.org/market-basket-analysis-in-data-mining/) – Shop owners analyze their sales and figure out which items are majorly bought together by the customers. For example, In USA, according to a study diapers and beers were usually bought together by fathers.
2. [**Social Network Analysis**](https://www.geeksforgeeks.org/social-network-analysis-using-r-programming/) – Social media sites use your data to understand your browsing behaviour and provide you with targeted friend recommendations or content recommendations.
3. **Medical Imaging** – Doctors use Clustering to find out diseased areas in diagnostic images like X-rays.
4. [**Anomaly Detection**](https://www.geeksforgeeks.org/machine-learning-for-anomaly-detection/) – To find outliers in a stream of real-time dataset or forecasting fraudulent transactions we can use clustering to identify them.
5. **Simplify working with large datasets** – Each cluster is given a cluster ID after clustering is complete. Now, you may reduce a feature set’s whole feature set into its cluster ID. Clustering is effective when it can represent a complicated case with a straightforward cluster ID. Using the same principle, clustering data can make complex datasets simpler.

There are many more use cases for clustering but there are some of the major and common use cases of clustering. Moving forward we will be discussing Clustering Algorithms that will help you perform the above tasks.

**3. Explain Elbow Method in K Means clustering**

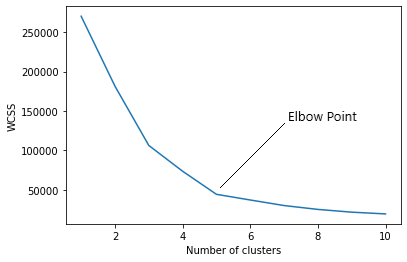
* The Elbow Method is a technique used in data analysis and machine learning for determining the optimal number of clusters in a dataset.
* It involves plotting the variance explained by different numbers of clusters and identifying the “elbow” point, where the rate of variance decreases sharply levels off, suggesting an appropriate cluster count for analysis or model training.
* The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value.
* WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.

The formula to calculate the value of WCSS (for 3 clusters) is given below:

****

**K Means Clustering Using the Elbow Method**

1. In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10.
2. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square).
3. WCSS is the sum of the squared distance between each point and the centroid in a cluster.
4. When we plot the WCSS with the K value, the plot looks like an Elbow.
5. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1.
6. When we analyze the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape.
7. From this point, the graph moves almost parallel to the X-axis. The K value corresponding to this point is the optimal value of K or an optimal number of clusters.



**Pros of Elbow Method:**

1. **Simplicity:** It's easy to understand and implement. You just plot the WCSS vs number of clusters (k) and look for the elbow.
2. **Visualization:** The graph provides a visual aid for choosing k, making it easier to understand the relationship between cluster number and data fit.
3. **Efficiency:** It's computationally cheap compared to other methods for choosing k.

**Elbow Method Drawbacks:**

1. **Subjectivity:** The choice of the “elbow point” can be subjective and might vary between individuals analyzing the same data.
2. **Non-Gaussian Data:** It assumes that clusters are spherical and equally sized, which may not hold for complex datasets with irregularly shaped or differently sized clusters.
3. **Sensitivity to Initialization:** K-means itself is sensitive to initial cluster centroids, which can affect the WCSS values and, consequently, the choice of the optimal K.
4. **Inefficient for Large Datasets:** For large datasets, calculating WCSS for a range of K values can be computationally expensive and time-consuming.
5. **Unsuitable for All Distributions:** The elbow method is not suitable for all data distributions, especially when clusters have varying densities or are non-convex.
6. **Limited to K-means:** It specifically applies to K-means clustering and may not be suitable for other clustering algorithms with different objectives.
7. **Explain working of dendrogram in Hierarchical clustering.**

# **Hierarchical Clustering in Machine Learning**

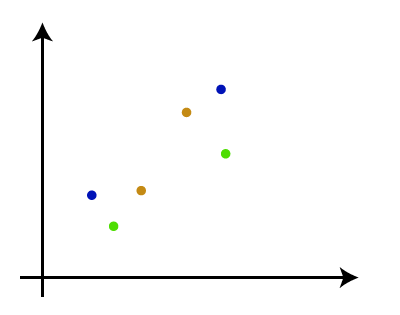
Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabelled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA.

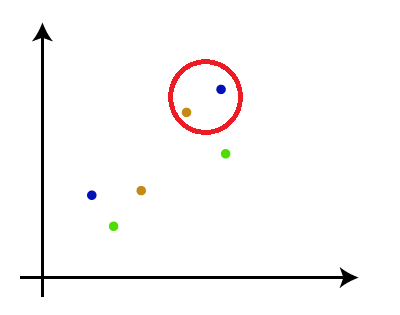
In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.

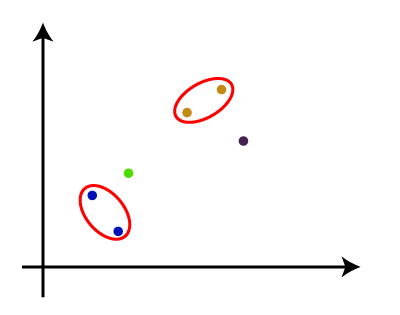
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

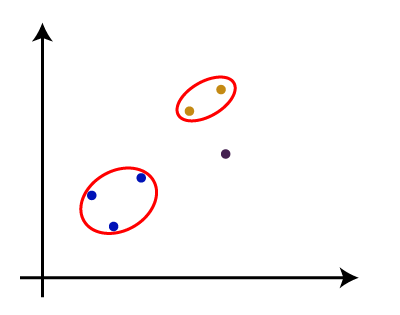
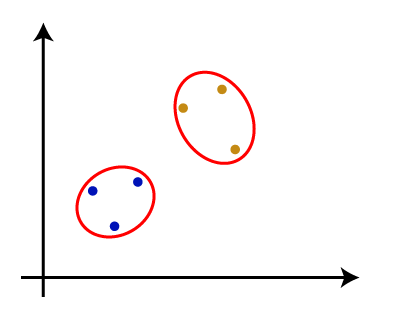
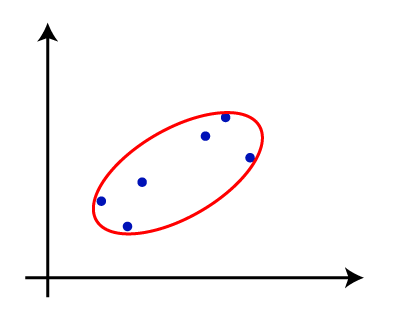
The hierarchical clustering technique has two approaches:

1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

**Step-1:** Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.  


**Step-2:** Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.  


**Step-3**: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.  


**Step-4:** Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:  
  
  


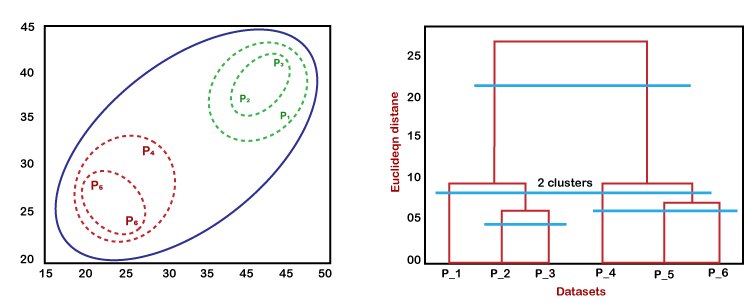
**Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

### **Woking of Dendrogram in Hierarchical clustering**

The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs.

In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

The working of the dendrogram can be explained using the below diagram:



In the above diagram, the left part is showing how clusters are created in agglomerative clustering, and the right part is showing the corresponding dendrogram.

* As we have discussed above, firstly, the datapoints P2 and P3 combine together and form a cluster, correspondingly a dendrogram is created, which connects P2 and P3 with a rectangular shape. The hight is decided according to the Euclidean distance between the data points.
* In the next step, P5 and P6 form a cluster, and the corresponding dendrogram is created. It is higher than of previous, as the Euclidean distance between P5 and P6 is a little bit greater than the P2 and P3.
* Again, two new dendrograms are created that combine P1, P2, and P3 in one dendrogram, and P4, P5, and P6, in another dendrogram.
* At last, the final dendrogram is created that combines all the data points together.

We can cut the dendrogram tree structure at any level as per our requirement.

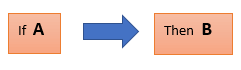
* Cutting the dendrogram at a lower height result in more clusters.
* Cutting at a higher height result in fewer clusters.

1. **Explain Association Rule mining.**

* Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.
* It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.
* The association rule learning is one of the very important concepts of machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production, etc.
* Here market basket analysis is a technique used by the various big retailer to discover the associations between items.
* We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.
* For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.  
    
  Association rule learning can be divided into three types of algorithms:

1. Apriori
2. Eclat
3. F-P Growth Algorithm

Association rule learning works on the concept of If and Else Statement, such as if A then B.



Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known *as single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

* **Support**
* **Confidence**
* **Lift**

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

Association Rule Learning

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

Association Rule Learning

Lift

It is the strength of any rule, which can be defined as below formula:

Association Rule Learning

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

* If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
* **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
* **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Types of Association Rule Lerning

Association rule learning can be divided into three algorithms:

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules.

It is designed to work on the databases that contain transactions.

This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together.

It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

* **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
* **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
* **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
* It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.

1. **Explain apriori algorithm.**

**Apriori Algorithm in Machine Learning:**

* The Apriori algorithm uses frequent item sets to generate association rules, and it is designed to work on the databases that contain transactions.
* With the help of these association rule, it determines how strongly or how weakly two objects are connected.
* This algorithm uses a breadth-first search and Hash Tree to calculate the item set associations efficiently.
* It is the iterative process for finding the frequent item sets from the large dataset.

This algorithm was given by the R. Agrawal and Srikant in the year 1994. It is mainly used for market basket analysis and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

To improve the efficiency of level-wise generation of frequent item sets, an important property is used called Apriori property which helps by reducing the search space.

**Apriori Property:**

All non-empty subset of frequent item set must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that

All subsets of a frequent item set must be frequent (Apriori property).

If an itemset is infrequent, all its supersets will be infrequent.

**Steps for Apriori Algorithm**

Below are the steps for the apriori algorithm:

**Step-1:** Determine the support of item sets in the transactional database, and select the minimum support and confidence.

**Step-2:** Take all supports in the transaction with higher support value than the minimum or selected support value.

**Step-3:** Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

**Step-4:** Sort the rules as the decreasing order of lift.

## Components of Apriori algorithm

The given three components comprise the apriori algorithm.

1. Support
2. Confidence
3. Lift

Let's take an example to understand this concept.

We have already discussed above; you need a huge database containing a large no of transactions. Suppose you have 4000 customers transactions in a Big Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together.

Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

### **Support**

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= 400/4000 = 10 percent.

### **Confidence**

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

Hence,

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

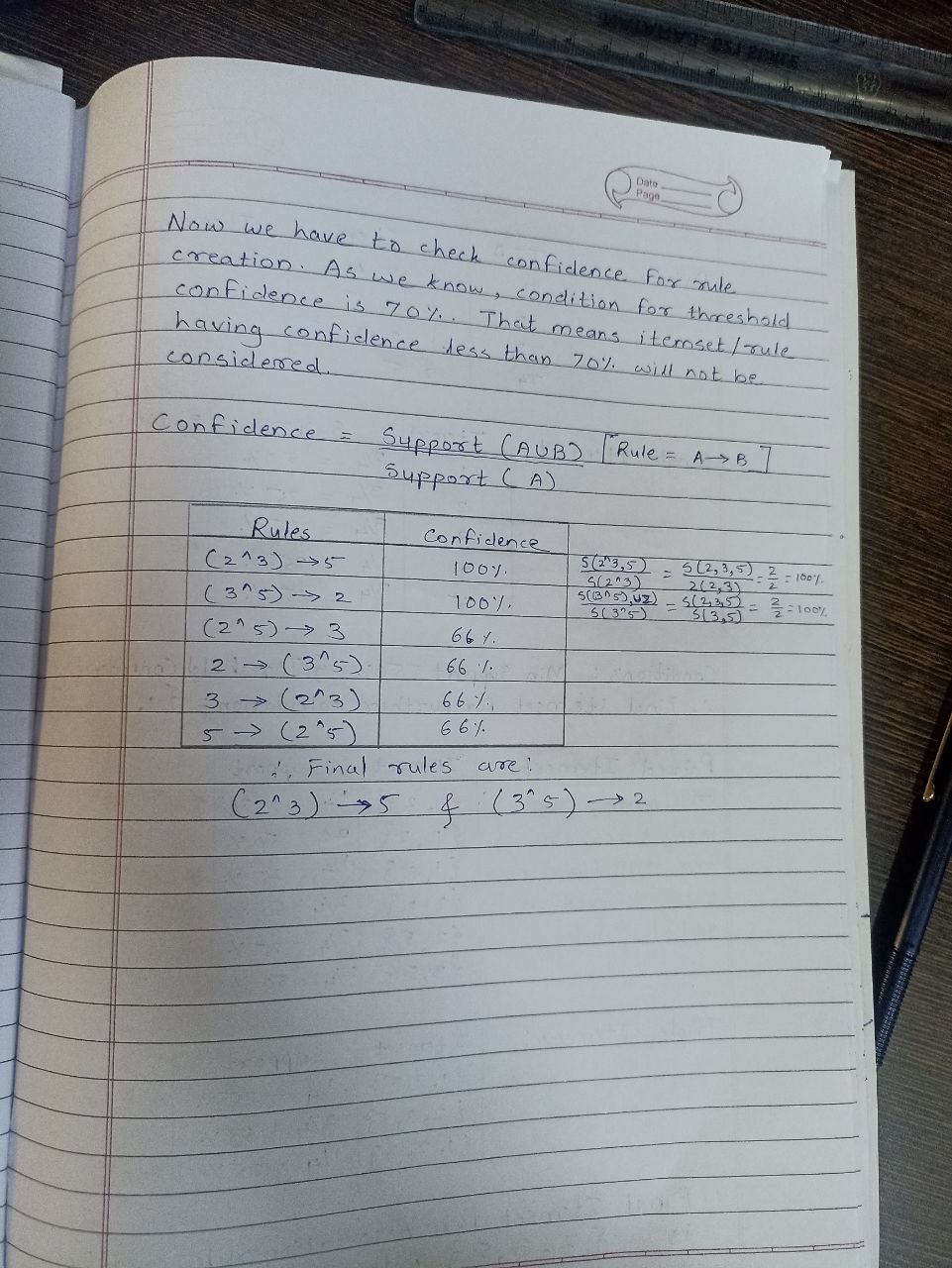
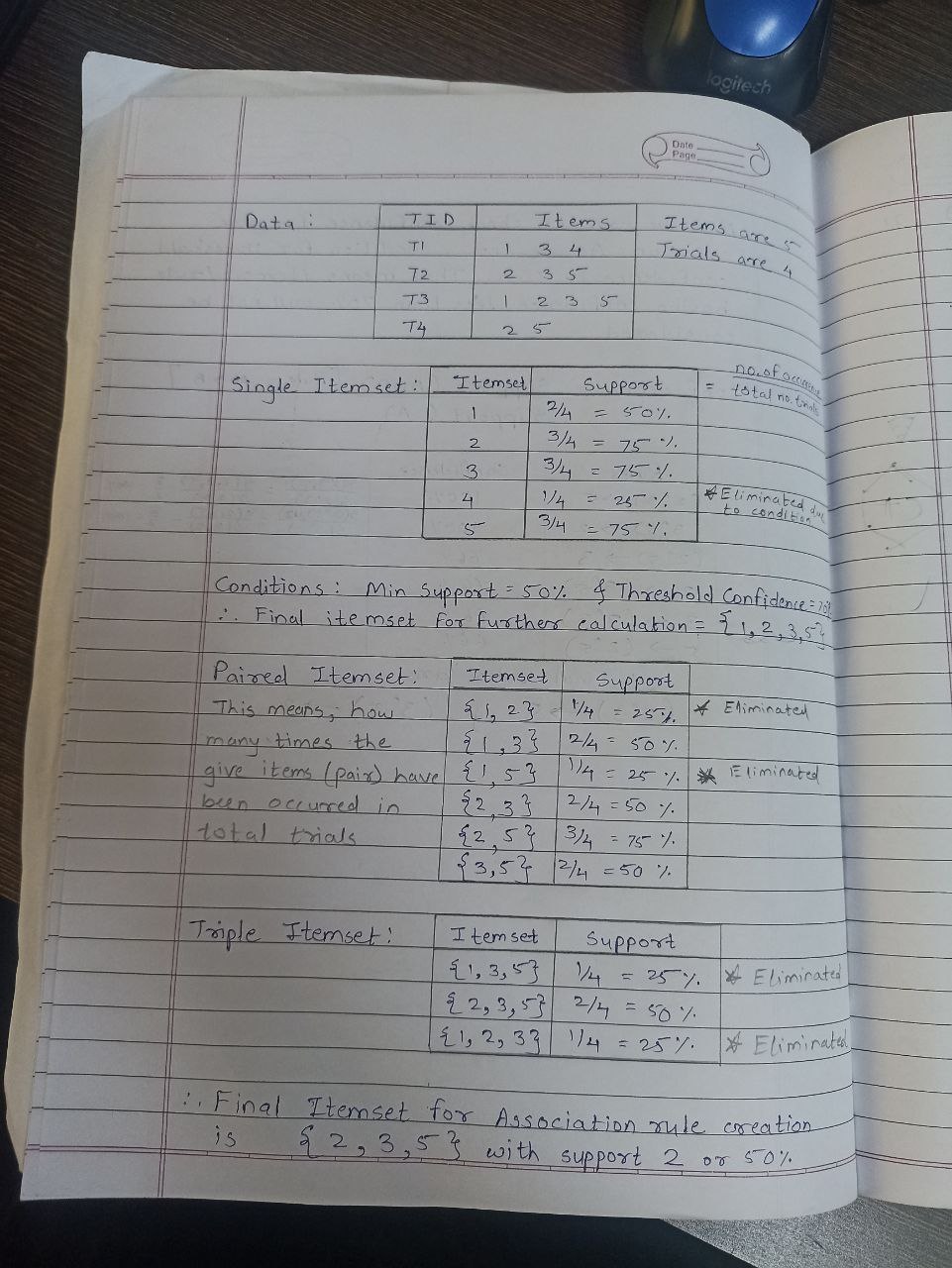
### **Lift**

Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.

Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

= 50/10 = 5

It means that the probability of people buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone. If the lift value is below one, it requires that the people are unlikely to buy both the items together. Larger the value, the better is the combination.

Example:  


Advantages of Apriori Algorithm

* This is easy to understand algorithm
* The join and prune steps of the algorithm can be easily implemented on large datasets.

Disadvantages of Apriori Algorithm

* The apriori algorithm works slow compared to other algorithms.
* The overall performance can be reduced as it scans the database for multiple times.
* The time complexity and space complexity of the apriori algorithm is O(2D), which is very high. Here D represents the horizontal width present in the database.