Project Report

On

# SUPERVISED AND UNSUPERVISED MACHINE LEARNING

Submitted

In partial fulfillment of the requirements for the degree of

**BACHELOR OF SCIENCE**

**In**

**MATHEMATICS**

**By**


**FIA JAISON**

**(Register No.AB21BMAT032)**


**Under the Supervision of**

**Dr. ELIZABETH RESHMA M T**



**DEPARTMENT OF MATHEMATICS**

**ST. TERESA'S COLLEGE (AUTONOMOUS)**

**ERNAKULAM, KOCHI - 682011**


**APRIL 2024**

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

# CERTIFICATE

This is to certify that the dissertation entitled, **SUPERVISED AND UNSUPERVISED MACHINE LEARNING** is a bonafide record of the work done by **Ms. Fia Jaison** under my guidance as partial fulfillment of the award of the degree of **Bachelor of Science in Mathematics at** St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 16/02/2024

Place: Ernakulam

**Dr. ELIZABETH RESHMA M T**

Assistant Professor,

Department of Mathematics,

St. Teresa's College (Autonomous),

Ernakulam.

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

**Dr. URSALA PAUL**

Assistant Professor and HOD,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous),

Ernakulam.

**External Examiners**

**1:** .....

**2:** .....

# DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of Dr. ELIZABETH RESHMA M T , Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date: 16/02/2024

**FIA JAISON**

**AB21BMAT032**

## ACKNOWLEDGEMENT

I extend my heartfelt appreciation to Dr. ELIZABETH RESHMA M T, Assistant Professor at St. Teresa's College, for her invaluable coordination, guiding us through the project's stages, and to my group members for their engaging defense and insightful contributions. Gratitude goes to Dr. Ursala Paul, Dr. Elizabeth Reshma M. T, and divine guidance for steering me through challenges, culminating in the completion of this project.

Ernakulam.

**FIA JAISON**

Date: 16/02/2024

**AB21BMAT032**

# INDEX

# CHAPTER 1

# INTRODUCTION TO MACHINE LEARNING

## 1.1 Machine learning

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems. Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or Past experience. The model may be predictive to make predictions in the future, or descriptive to gain Knowledge from data, or both.

Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, Coined the term "Machine Learning" in 1959 while at IBM. He defined machine learning as "the field of Study that gives computers the ability to learn without being explicitly programmed." However, there is no universally accepted definition for machine learning. Different authors define the term differently.

**Definition of learning**

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E.

Examples

I) Handwriting recognition learning problem

- Task T: Recognizing and classifying handwritten words within images
- Performance P: Percent of words correctly classified
- Training experience E: A dataset of handwritten words with given classifications.

II) A robot driving learning problem

- Task T: Driving on highways using vision sensors

- Performance measure P: Average distance traveled before an error
- training experience: A sequence of images and steering commands recorded while observing a human driver

III) A learning problem

- Task T: Playing chess
- Performance measure P: Percent of games won against opponents
- Train chess experience E: Playing practice games against itself

A computer program which learns from experience is called a machine learning program or simply a learning program. Such a program is sometimes also referred to as a learner.

Machine learning has played a progressively central role in human society since its beginnings in the mid-20$^{th}$ century, when AI pioneers like Walter Pitts, Warren McCulloch, Alan Turing and John von Neumann laid the groundwork for computation. The training of machines to learn from data and improve over time has enabled organizations to automate routine tasks that were previously done by humans – in principle, freeing us up for more creative and strategic work.

Machine learning also performs manual tasks that are beyond our ability to execute at scale – for example, processing the huge quantities of data generated today by digital devices. Machine learning's ability to extract patterns and insights from vast data sets has become a competitive differentiator in fields ranging from finance and retail to healthcare and scientific discovery. Many of today's leading companies, including Facebook, Google and Uber, make machine learning a central part of their operations. Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions.

There are four basic types of machine learning: supervised learning, unsupervised learning, semi supervised learning and reinforcement learning. In this project we are focusing on supervised and unsupervised machine learning.

### 1.2 Supervised machine learning

Supervised learning is a machine learning technique that is widely used in various fields such as finance, healthcare, marketing, and more. It is a form of machine learning in which the algorithm is trained on labeled data to make predictions or decisions based on the data inputs. In supervised learning, the algorithm learns a mapping between the input and output data. This mapping is learned from a labeled dataset, which consists of pairs of input and output data.

The algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on new, unseen data. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).Supervised learning models can require certain levels of expertise to structure accurately.

Training supervised learning models can be very time intensive. Datasets can have a higher likelihood of human error, resulting in algorithms learning incorrectly.

Unlike unsupervised learning models, supervised learning cannot cluster or classify data on its own are challenges. In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

#### Working principle

The first step is to gather a labeled training dataset, which contains examples paired with their corresponding correct outputs. This dataset serves as the foundation for training the model. Once the training dataset is collected, it's essential to split it into three subsets: the training dataset itself, the test dataset, and optionally, the validation dataset. The training dataset is used to train the model, while the test dataset is reserved for evaluating the model's performance on unseen data. The validation dataset, if utilized, helps fine-tune the model's parameters and prevent over fitting during training. Determining the input features of the training dataset is crucial; these features should encapsulate sufficient information to enable the model to accurately predict the output. Based on the nature of the problem and dataset characteristics, the appropriate algorithm, such as support vector machine or decision tree, is selected for training the model. Once the algorithm is executed on the training dataset, the model learns from the input features and their corresponding labels. Evaluation of the model's accuracy is conducted by providing it with the test set and comparing its predictions with the actual outputs. If the model accurately predicts the outputs for the test dataset, it signifies that the model has

successfully learned from the training data and can generalize well to new instances, thus demonstrating its accuracy.

**Applications**

Supervised learning is used to predict house prices. Data having details about the size of the house, price, the number of rooms in the house, garden and other features are needed. We need data about various parameters of the house for thousands of houses and it is then used to train the data.

Supervised learning, fueled by labeled training data, proves invaluable in various applications such as predictive analytics for house and stock prices, text recognition, spam detection, customer sentiment analysis, and object detection like face models. The advantages lie in its ability to enable models to intricately learn patterns and relationships, leading to accurate predictions and classifications for new data. This approach spans a wide spectrum of applications, encompassing classification, regression, and tackling complex challenges like image recognition and natural language processing. The well-established evaluation metrics, including accuracy, precision, recall, and F1-score, play a crucial role in efficiently assessing the performance of supervised learning models across diverse tasks.
Supervised learning is typically divided into two main categories: regression and classification

**1.3 Unsupervised machine learning**

Unsupervised learning is a type of machine learning that learns from unlabeled data. This means that the data does not have any pre-existing labels or categories. A machine leveraging unclassified and unlabeled data, enabling the algorithm to operate autonomously without explicit guidance.The task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

Unsupervised learning is helpful for finding useful insights from the data and it is much similar as a human learns to think by their own experiences, which makes it closer to the real AI Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important. In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Unsupervised learning allows the model to discover patterns and relationships in unlabeled data. Clustering algorithms group similar data points together based on their inherent characteristics. Feature extraction captures essential information from the data, enabling the model to make meaningful distinctions. Label association assigns categories to the clusters based on the extracted patterns and characteristics.

**Working principle**

An unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Unlabeled input data is fed to the machine learning model in order to train it. It will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

**Applications**

Unsupervised learning can identify unusual patterns or deviations from normal behavior in data, enabling the detection of fraud, intrusion, or system failures and can uncover hidden relationships and patterns in scientific data, leading to new hypotheses and insights in various scientific fields. Identify patterns and similarities in user behavior and preferences to recommend products, movies, or music that align with their interests also can identify groups of customers with similar characteristics, allowing businesses to target marketing campaigns and improve customer service more effectively. Then unsupervised learning can group images based on their content, facilitating tasks such as image classification, object detection, and image retrieval.

A machine learning model trained on a large dataset of unlabeled images, containing both dogs and cats. The model has never seen an image of a dog or cat before, and it has no pre-existing

labels or categories for these animals. Your task is to use unsupervised learning to identify the dogs and cats in a new, unseen image.

It does not require training data to be labeled. Dimensionality reduction can be easily accomplished using unsupervised learning. Capable of finding previously unknown patterns in data. Unsupervised learning can help gain insights from unlabeled data that you might not have been able to get otherwise and it is good at finding patterns and relationships in data without being told what to look for. This can help to learn new things about the data are the advantages of unsupervised learning.

# CHAPTER 2
# CLASSIFICATION ALGORITHMS IN SUPERVISED MACHINE LEARNING

Classification in supervised machine learning is a predictive modeling task where the goal is to assign predefined categories or labels to input data. It involves training a model on a labeled dataset, learning the patterns in the data, and then using that knowledge to classify new, unseen instances into predefined classes. Common algorithms for classification include logistic regression, decision trees, and random forest. This powerful technique finds applications in various domains, such as spam detection, image recognition, and medical diagnosis.

Key concepts in classification include features, which are the input variables used for prediction, and the decision boundary, a delineation between different classes. The choice of algorithm depends on the nature of the data and the problem at hand. Classification in supervised machine learning is a vital tool for automating decision-making processes, making predictions, and solving problems where assigning predefined categories to data is essential.

In our exploration of classification algorithms in supervised learning, our focus narrowed to logistic regression, decision tree, and random forest for specific reasons. Logistic regression was chosen for its simplicity in binary tasks, while decision trees provided transparency in decision-making. The inclusion of random forest, an ensemble method of decision trees, aimed to enhance accuracy through collective strength. It's essential to note that other classification algorithms, such as Support Vector Machines, Naïve Bayes, and k-Nearest Neighbors, exist with their unique merits. Our selection aimed to strike a balance between foundational understanding (logistic regression) and the more advanced ensemble approach (random forest), showcasing the versatility within the classification landscapes.

## 2.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not. For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).

**Working principle**

Logistic regression uses a logistic function called a sigmoid function to map predictions and their probabilities. The sigmoid function refers to an S-shaped curve that converts any real value to a range between 0 and 1.

Moreover, if the output of the sigmoid function (estimated probability) is greater than a predefined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability is less than the predefined threshold, the model predicts that the instance does not belong to the class. Sigmoid function is referred to as an activation function for logistic regression Equation of logistic regression

$$f(x) = \frac{1}{1 + e^{-x}}$$

In this equation of logistic regression e is the base of natural logarithms, x is the input value and value of f(x) gives the transformed numerical value between the range of 0 and 1.

If the output of the sigmoid function is above 0.5, the output is considered as 1. On the other hand, if the output is less than 0.5, the output is classified as 0. Also, if the graph goes further to the negative end, the predicted value of y will be 0 and vice versa. In other words, if the output of the sigmoid function is 0.65, it implies that there are 65% chances of the event occurring.

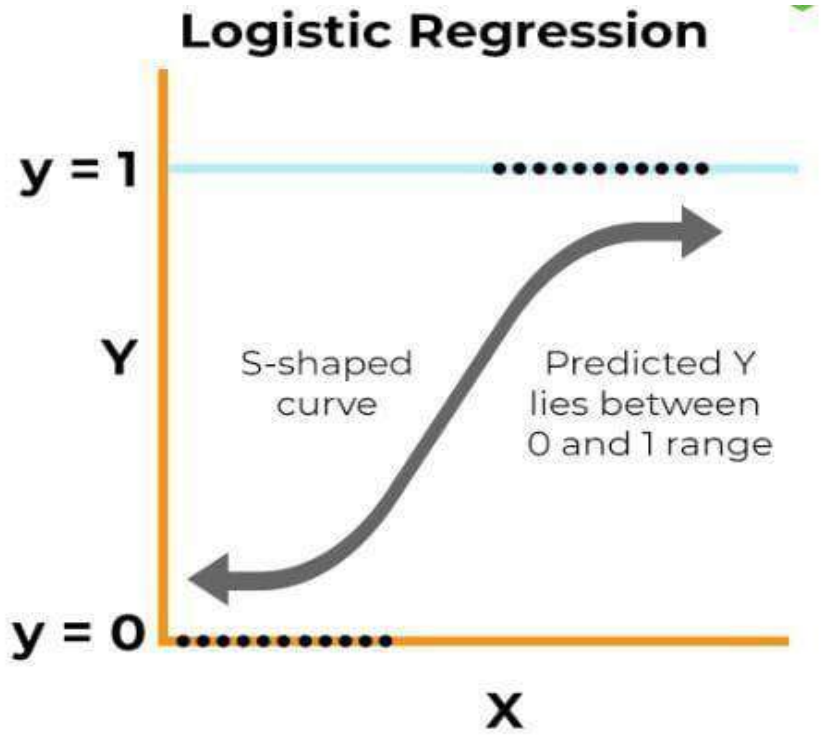


Figure 2.1: Logistic regression curve that shows the probability of a binary outcomes, based on two predictors variables X and Y

**Application**

Consider a medical scenario. Suppose you have a dataset of patients, and you want to predict whether a patient is at risk of developing a certain medical condition (1) or not (0) based on features like age, blood pressure, and cholesterol levels. The logistic regression model would analyze these features, assigning weights to each, and then use the logistic function to calculate the probability of a patient being at risk. If the model predicts a probability of 0.7, it implies a 70% chance that the patient is at risk. By setting a threshold (e.g., 0.5), you can classify patients—above the threshold means at risk, below means not at risk. This application of logistic regression is common in healthcare for predicting disease risks, helping identify individuals who might need further medical attention or preventive measures.

Advantage of logistic regression is that it is easier to implement, interpret, and very efficient to train. Also it makes no assumptions about distributions of classes in feature space. But if the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to over fitting. Also it can easily be outperformed by other complex algorithms. Due to its simplicity, it is can be used as a good baseline to compare with the performance of other more complex algorithms.

## 2.2 Decision Tree

Decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.

**Terminologies**

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

**Working principle**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

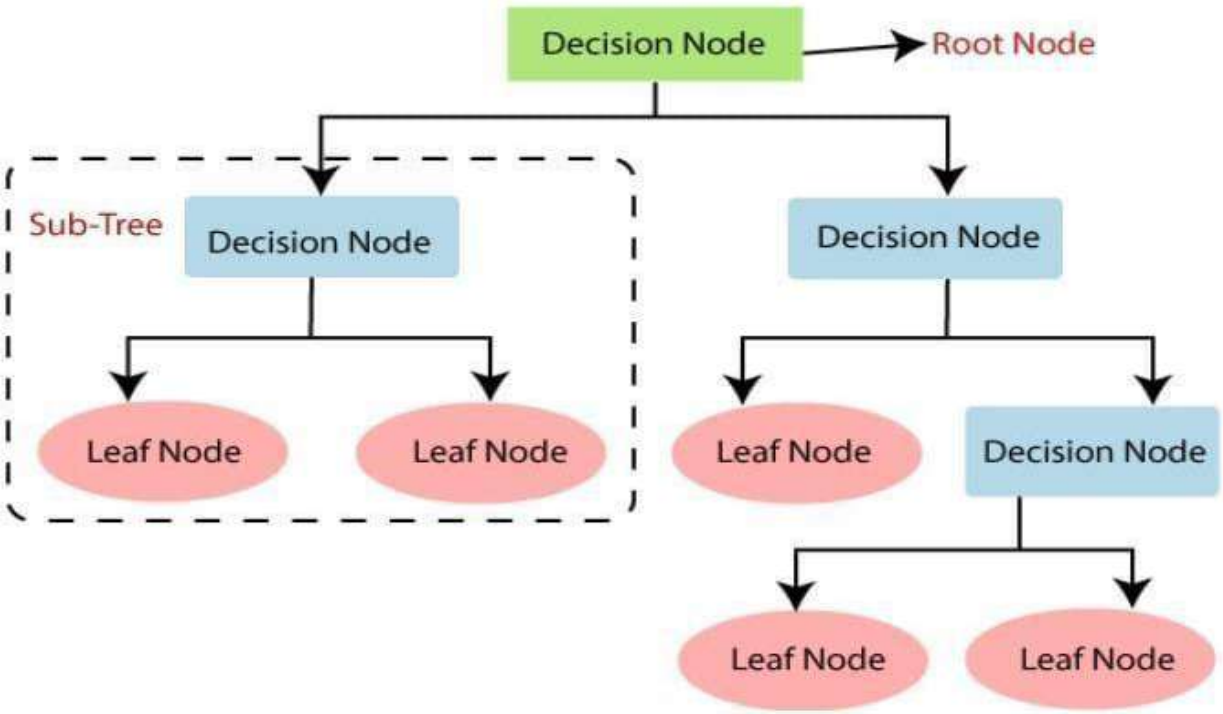


Figure 2.2: Structure of a decision tree.

**Application**

Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node, Salary attribute. The root node splits further into the next decision node, distance from the office, and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node, Cab facility, and one leaf node. Finally, the decision node splits into two leaf nodes Accepted offers and Declined offer.

A key advantage of decision trees lies in their simplicity to understand, as it follows the same process which a human follow while making any decision in real-life and helps to think about all the possible outcomes for a problem. But decision tree contains lots of layers, which makes it complex & for more class labels, the computational complexity of the decision tree may increase.

Banking sector mostly uses this algorithm for the identification of loan risk. In medical field with the help of this algorithm, disease trends and risks of the disease can be identified.

We can identify the areas of similar land use by this algorithm. Marketing trends can be identified using this algorithm are the applications of random forest.

### 2.3 Random Forest

Random forest is most popular machine learning algorithm which can be used in both classification and regression. It is a collection of decision trees, where each tree is trained using a randomly selected subset of the data. When random forest algorithm combines multiple decision trees in order to reduce the risk of over fitting. The result is a much more accurate and stable prediction. It is one of the most popular and widely used machine learning algorithms. It can be used for both regression and classification tasks. It is also used for feature selection and to identify important variables in a dataset. It is an efficient and effective tool for complex data analysis. It takes less training time as compared to other algorithms.

It predicts output with high accuracy, even for the large dataset it runs efficiently. Random Forest is a classifier that comprises a number of decision trees on various subsets of the provided dataset and takes the average to predicted accuracy of that dataset, as the name implies. Instead of depending on a single decision tree, the random forest collects the predictions from each tree and predicts the final output based on the majority vote of predictions. It can also maintain accuracy when a large proportion of data is missing. Requires more resources, more complex, Time-consuming process are the Challenges. Reduced risk of over fitting, Provides flexibility, Easy to determine feature importance are the benefits of random forest. Some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations are two assumptions for a better random forest classifier.

Random Forest is capable of performing both Classification and Regression tasks and it is capable of handling large datasets with high dimensionality. It enhances the accuracy of the model and prevents the over fitting issue and it works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

**Working principle**

Select random K data points from the training set. Build the decision trees associated with the selected data points (Subsets).Choose the number N for decision trees that you want to build. Repeat Step 1 & 2. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

**Application**

The dataset, consisting of various fruit images, is fed into a Random Forest classifier where it is partitioned into subsets and distributed among individual decision trees. Throughout the training process, each decision tree generates its own prediction outcome. When confronted with a new data point, the Random Forest classifier combines the results from all decision trees to make a final decision, relying on the majority consensus.

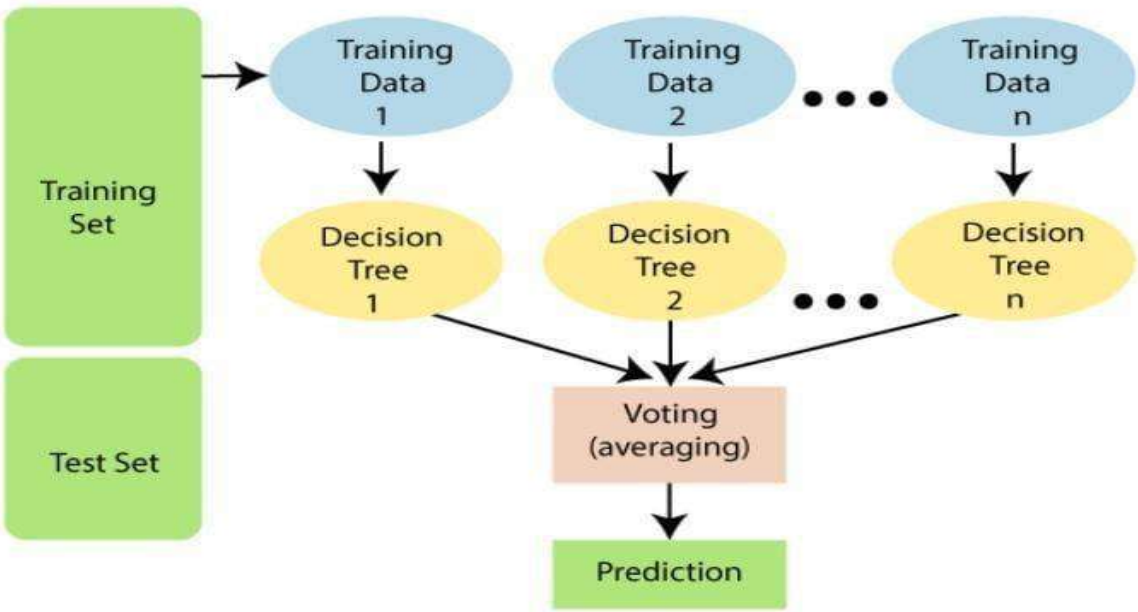


Figure 2.3: Structure of a random forest.

# CHAPTER 3
# REGRESSION ALGORITHMS IN SUPERVISED MACHINE LEARNING

Regression in supervised machine learning is a type of algorithm used for predicting a continuous outcome or numerical value based on input features. The goal is to establish a relationship between the input variables and the target variable, allowing the model to make predictions for new, unseen data.

Let's consider a scenario where you want to predict a student's final exam score based on the number of hours they spend studying per week. Collect data on students, noting the number of hours each student studies per week and their corresponding final exam scores. Using this dataset, train a regression model. The model learns the relationship between the input feature (study hours) and the output (final exam score). Now, with a trained model, you can input the number of hours a new student studies, and the model will predict their likely final exam score. For instance, if a student studies 10 hours per week, the regression model might predict a final exam score of 85.

What makes regression different from classifications is that regression is primarily about making predictions, while association is about discovering patterns or relationships within the data. In regression, there is a clear distinction between dependent (outcome to be predicted) and independent variables (features used for prediction). Association typically deals with finding connections between variables without a strict emphasis on predicting one from another. Regression predicts a continuous numerical outcome, whereas association often deals with categorical variables or binary relations.

Linear regression, along with its regularized counterparts, ridge regression and lasso regression, are employed in supervised learning to forecast a variable's value based on the input of another variable, facilitating predictive analysis. Lasso and Ridge Regression are two of the most popular techniques for regularizing linear models, which often yield more accurate predictions than traditional linear models. These methods reduce the model’s complexity by introducing shrinkage or adding a penalty to complex coefficients. It’s essential to note that

other classification algorithms, such as Random Forest Regression, Decision Tree Regression, Support Vector Regression, Polynomial Regression, and Logistic Regression. Our selection aimed to capture the relationships between independent and dependent variables, with the main purpose of predicting an outcome.

### 3.1 Linear Regression:

Linear Regression is a supervised learning method, it is one of the most popular and easiest machine learning algorithms. It is a statistical method that is used for predictive analysis. The main idea of regression is to examine two things. First, does a set of predictor variables do a good job in predicting an outcome (dependent) variable. The second thing is which variables are significant predictors of the outcome variable. Linear regression makes predictions for continuous/real or numeric variables. Such as price prediction, marks prediction, salary, age and so on. For example we're given a dataset of used cars, which contains the name of the car, year, selling price, present price, number of kilometers it has driven, type of fuel, type of the seller, transmission, and if the seller is the owner. Our goal is to predict the selling price of the car.

Linear regression helps to determine the strength of predicators, Forecast the effect through Prediction, Trend Analysis / Forecasting, preventing mistakes, increasing efficiency. Importance of linear regression are used to predict the output variable based on input variables, used to analyses the relationship between two or more variables it helps in identifying the significant variables that affect the outcome variable, to identify the most important features in a dataset it helps in selecting the relevant features that are important for prediction. Evaluate the performance of a model and it helps in determining the accuracy of the model and identifying areas for improvement.

#### Working principle

Linear regression is a fundamental statistical and machine learning technique. It shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. A linear regression model provides a sloped straight line representing the relationship

between the dependent and independent variables is called a regression line. Linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing.

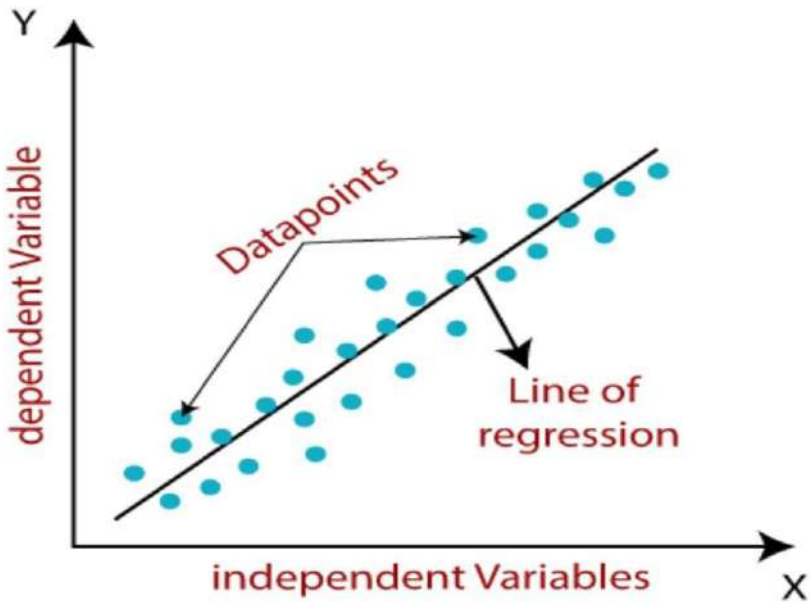


Figure 3.1: Linear regression model provides a sloped straight line representing the relationship between the variables.

Mathematically, we can represent a linear regression as:

$y = a_0 + a_1 x + \varepsilon$

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

$a_0$= intercept of the line

$a_1$ = Linear regression coefficient

ε = random error

For example, here we are taking a dataset that has two variables: salary (dependent variable) and experience (Independent variable).

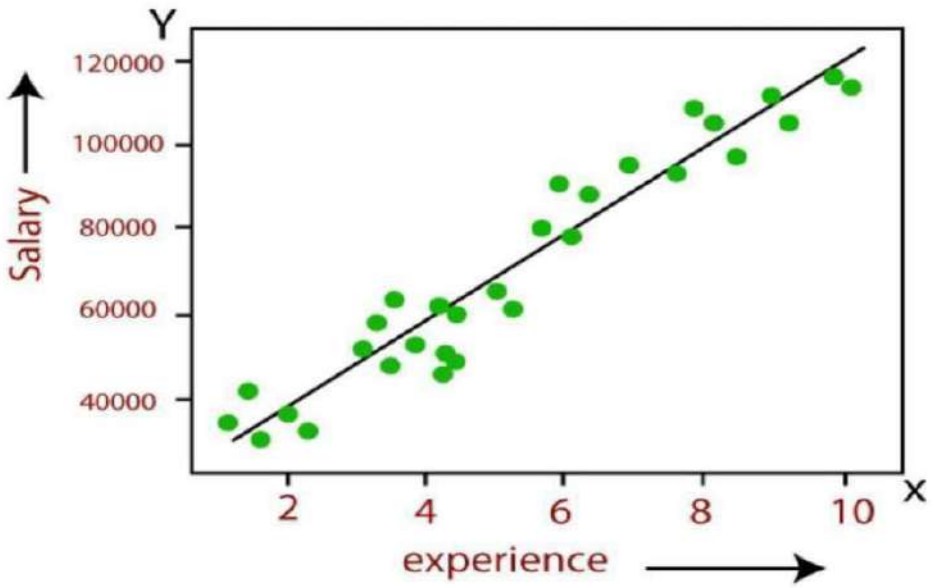


Figure 3.2: Relationship between the dependent variable and independent variable.

Mathematical equation for linear regression

$$Y = a X + b$$

In this equation of a linear regression, Salary of an employee based on year of experience. The recent company data, which indicates that the relationship between experience and salary. Here the experience is an independent variable, and the salary of an employee is a dependent variable and a & b are linear coefficients, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

.

There are two main types of linear regressions namely simple linear regression and multiple linear regression

**Simple Linear Regression:**

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression. In Simple Linear Regression is that the dependent variable must be a continuous/real value. However, the

independent variable can be measured on continuous or categorical values. For example the weight of the person is linearly related to their height. So, it shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase. It is not necessary that one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables.

**Multiple Linear regression:**

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm called Multiple Linear Regression. The dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form. For example when a real estate employee who wants to create a model to help predict the best time to sell homes at the maximum sales price, but multiple factors can affect the sales price. These variables include the age of the house, the value of other homes in the neighborhood, quantitative measurements of the public school system regarding student performance and the number of nearby parks, among other factors. They can build a prediction model off these four independent variables to predict the maximum sales price of homes and the variables if any of these factors change in terms of their coefficient values.

Linear Regressions are mainly the regression model to evaluate the relationship between natural elements, such as soil, water and air. An example is the relationship between the amount of water and plant growth. This can help environmentalists predict the effects of air or water pollution on environmental health. Medical researchers can use this regression model to determine the relationship between independent characteristics, such as age and body weight, and dependent ones, such as blood pressure. This can help reveal the risk factors associated with diseases and they can use this information to identify high-risk patients and promote healthy lifestyles. Financial analysts use linear models to evaluate a company's operational performance and forecast returns on investment also they use in the capital asset pricing model, which studies the relationship between the expected investment returns and the associated market risks. It shows companies if an investment has a fair price and contributes to decisions on whether or not to invest in the asset.

### 3.2 Ridge Regression:

Ridge Regression is a type of linear regression that adds a penalty term to the ordinary least squares (OLS) objective function. This penalty term, controlled by a regularization parameter, helps prevent over fitting and addresses multicollinearity in regression models and used to handle multicollinearity, where predictor variables are highly correlated, and to prevent over fitting in regression models with a large number of features. It helps stabilize coefficient estimates and improves the model's generalization to new data, it is primarily used for regression problems where the goal is to predict a continuous outcome variable. It may not be suitable for classification problems, where the outcome is categorical. Ridge regression is a linear regression type with the objective of analyzing multicollinearity in multiple regression data.

It aims to reduce the sum of squared errors between the actual and predicted values and actual values by adding a penalty term that diminishes the coefficients and brings them closer to zero it helps to analyze a data set where the predictor variables number is more than the number of observations. It is valuable for analyzing data sets comprising a more significant number of predictors than the number of observations. It restrains over fittings in the model and decreases its complexity. However, this regression is biased, and the L2 regularization decreases the regression coefficient values and brings them toward zero and it helps to analyze a data set where the predictor variables number is more than the number of observations.

Multicollinearity is when the data set contains over two predicted variables with high correlations. It is valuable for analyzing data sets comprising a more significant number of predictors than the number of observations. It restrains over fittings in the model and sexreses its complexity. The advantages are L2 penalty models will continue to work to decrease over fitting. Since the penalty reduces some coefficient values close to zero, it reduces over fitting. Additionally, it decreases the model's complications. Users can apply these models to data sets that comprise several correlated features. Generally, the correlated features are a drawback for regression models, but the L2 penalty's application into a regression model decreases the negative effect of correlated features.

The equation for ridge regression is :

$$L(x, y) = Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}(w_i)^2)$$

This equation is the objective function for ridge regression W represents the coefficients, is the Regularization parameter, $x_i$ represents the features, and $y_i$ represents the target values. The objective is to minimize this function with respect to the coefficients w to find the best fit for the data while penalizing large coefficients to prevent over fitting.

Applying this method is especially beneficial in cases with more features than observations. However, this method usually causes difficulties for standard regression models. Furthermore, it is appropriate if multicollinearity exists in the data set. Finally, it is suitable when many vast parameters have almost the same value, which means that most predictors influence the reaction. High Multicollinearity When your predictor variables are highly correlated with each other, Ridge regression can help stabilize coefficient estimates by adding a regularization term that discourages overly large coefficients. This is especially important because multicollinearity can lead to unstable and unreliable coefficient estimates in ordinary linear regression. Multicollinearity is when the data set contains over two predicted variables with high correlations. Advantages of Ridge Regression handles Multicollinearity, Reduces Over fitting, and Suitable for High-Dimensional Data.

Real life Application of ridge regression can be used to predict housing prices by considering various features like square footage, number of bedrooms, location, etc., while handling correlated predictor variables. In medical research, help build predictive models to diagnose diseases based on patient characteristics and medical measurements while mitigating the effects of correlated factors. In climate modeling to analyze the relationships between various climate variables and predict future climate patterns.

One of the most important things about ridge regression is that without wasting any information about predictions it tries to determine variables that have exactly zero effects. Ridge regression is popular because it uses regularization for making predictions and regularization is intended to resolve the problem of over fitting. In this method penalizes the model for aggregating the weight's squared value. As a result, the weights generally have smaller absolute values.

Furthermore, they penalize the extreme values of the weights, which leads to a group of weights that are more uniformly distributed.

**Applications**

A telecom company, they are analyzing the customers who stopped the services. The data set of customer information, including gender, age, customer service interactions, and usage patterns they build a model predicting which customers would end the services. The data set contained a vast number of features, and some features were unrelated to the study they used a regression model for adding a penalty term that would reduce the effects of the irrelevant features on the analysis.

### 3.3 Lasso Regression

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, is a regularization technique used in statistical modeling and machine learning. It aims to balance model simplicity and accuracy by adding a penalty term to the traditional linear regression model. The penalty encourages sparsity, leading some coefficients to be precisely zero. Lasso is particularly valuable for feature selection, automatically identifying and discarding irrelevant or redundant variables.

**Mechanism of Lasso Regression**

Lasso regression employs a linear regression model initially, assuming a linear relationship between independent variables and the dependent variable. It then introduces L1 regularization, a penalty term based on the absolute values of the coefficients. The regularization parameter $\lambda$ controls the strength of regularization. The objective is to minimize the sum of squared differences between predicted and actual values while minimizing the L1 regularization term. This process shrinks coefficients toward zero, with some becoming precisely zero when $\lambda$ is sufficiently large, aiding in feature selection.

**Key Steps in Lasso Regression:**

Lasso Regression starts with a linear regression model and then introduces L1 regularization, penalizing coefficients based on their absolute values. The objective function aims to minimize

the sum of squared differences and the L1 regularization term, facilitating automatic feature selection by driving some coefficients to zero. The tuning parameter λ crucially influences the regularization strength; larger values result in sparser models with more coefficients set to zero, while smaller values allow more non-zero coefficients. Optimization, often via Coordinate Descent, estimates coefficients by iteratively updating them. Lasso regression is integral for preventing over fitting in regularization, contributing to building robust and interpretable models, especially in high-dimensional datasets alongside Ridge regression and Elastic Net.

Mathematical equation of Lasso Regression

$$L(x, y) = Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + \lambda \sum_{i=1}^{n}|w_i|)$$

This equation represents the loss function, which measures the difference between the predicted values and the actual values.$w_i$ Denotes the weights associated with each feature $x_i$. These weights determine the contribution of each feature to the prediction. $y_i$ Represents the actual values of the target variable for each observation. $\lambda$ is a regularization parameter. It controls the strength of regularization, balancing the fitting of the data with keeping the model simple.

Lasso Regression is different from ridge regression as it uses absolute coefficient values for normalization. It uses a unique approach that sets some coefficients exactly to zero, making it useful for selecting important features and simplifying the model. This property helps in dealing with high-dimensional data by automatically excluding less relevant variables, offering a balance between accuracy and simplicity in predictions.

Lasso Regression is a valuable tool in machine learning, contributing to both prediction accuracy and automatic feature selection, making it particularly effective in scenarios with high-dimensional datasets.

# CHAPTER 4
# CLUSTERING ALGORITHMS IN UNSUPERVISED MACHINE LEARNING

Clustering in unsupervised machine learning involves grouping similar data points together based on certain characteristics, without predefined labels. The goal is to discover inherent patterns or structures within the data. Popular algorithms for clustering include K-means, hierarchical clustering, and DBSCAN. These methods help identify natural clusters, aiding in data exploration and pattern recognition without the need for labeled training data.

Clustering algorithms aim to minimize intra-cluster differences while maximizing inter-cluster dissimilarities. K-means, for instance, partitions data into K clusters by iteratively optimizing cluster centroids. Hierarchical clustering builds a tree-like structure of clusters, offering a visual representation of data relationships. DBSCAN, on the other hand, identifies dense regions separated by sparser areas. Clustering finds applications in various fields, from customer segmentation in marketing to anomaly detection in cyber security.

One of the primary objectives of clustering algorithms in unsupervised learning, including K-means clustering, Hierarchical clustering, DBSCAN, and the Mean Shift Algorithm, is to group data points into clusters based on similarity patterns. K-means clustering, a fundamental unsupervised machine learning technique, is just one among many powerful tools within the expansive landscape of Data Science methodologies and operations. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data. Hierarchical clustering is a very useful way of segmentation. The advantage of not having to pre-define the number of clusters gives it quite an edge over k-Means. However, it doesn’t work well when we have huge amount of data. DBSCAN is a versatile clustering algorithm with several advantages for uncovering hidden patterns within datasets. Through its density based approach, DBSCAN can identify clusters of arbitrary shapes and handle noisy data, making it well-suited for various data analysis tasks.

### 4.1 K-means clustering

K-means clustering is a widely used unsupervised machine learning algorithm designed for partitioning a dataset into K distinct, non-overlapping subsets or clusters. The "K" in K-means represents the predetermined number of clusters that the algorithm seeks to create within the dataset. The algorithm works iteratively to assign each data point to one of the K clusters based on the similarity of features.

K-means clustering is a widely utilized unsupervised machine learning algorithm aimed at partitioning a dataset into K distinct, non-overlapping subsets or clusters. The algorithm iteratively assigns each data point to one of the K clusters based on feature similarity, with “K” representing the predetermined number of clusters sought within the dataset. K-means is known for its simplicity, efficiency, and scalability, making it suitable for various applications such as customer segmentation, image compression, and data preprocessing.

**Working principle**

K-means begins by placing K centroids randomly in the feature space, with each centroid representing the center of a potential cluster. It then iterates between two main steps. First, it assigns each data point to the nearest centroid, forming initial clusters. Next, it recalculates the centroids as the mean of all data points assigned to each cluster. This iterative process continues until convergence, where the centroids stabilize, and the clusters become more defined.

**Applications**

Imagine having a dataset of customer purchasing behavior, where each data point represents a customer’s features like age, income, and spending habits. By employing K-means clustering with K=3, the algorithm partitions the dataset into three clusters based on similarities in these features. This segmentation enables businesses to understand distinct customer segments and tailor marketing strategies accordingly, enhancing customer engagement and satisfaction.

Consider a dataset of student performance, with features like exam scores in different subjects. By applying K-means clustering with K=2, the algorithm partitions the dataset into two clusters representing high-performing and low-performing students. This segmentation allows

educators to identify factors contributing to student success and tailor interventions accordingly, ultimately improving academic outcomes. Thus, K-means clustering proves valuable in uncovering meaningful patterns and structures within data for informed decision making across various domains.

## 4.2 Hierarchical clustering

Hierarchical clustering is a versatile unsupervised machine learning technique that organizes data into a tree-like structure, where clusters of similar items are successively grouped together. This approach can be either agglomerative, starting with individual data points and merging them into clusters, or divisive, beginning with one cluster and progressively dividing it. One of the defining features of hierarchical clustering is its ability to visualize relationships within the data through dendrogram, offering insights into the hierarchical structure of the clusters.

### Working principle

When a dataset is given with individual data points, at the beginning, each point is considered its own cluster. The algorithm looks at all pairs of data points and calculates their similarity or dissimilarity. The two most similar points or clusters are merged into a new cluster. This process continues iteratively, and the similarity between clusters is recalculated at each step. As clusters merge, a tree structure (dendrogram) is formed. The height at which two clusters merge in the dendrogram represents their dissimilarity. Depending on the analysis goals, you can cut the dendrogram at a certain height to obtain a specific number of clusters. A lower cut results in fewer, larger clusters, while a higher cut leads to more, smaller clusters.

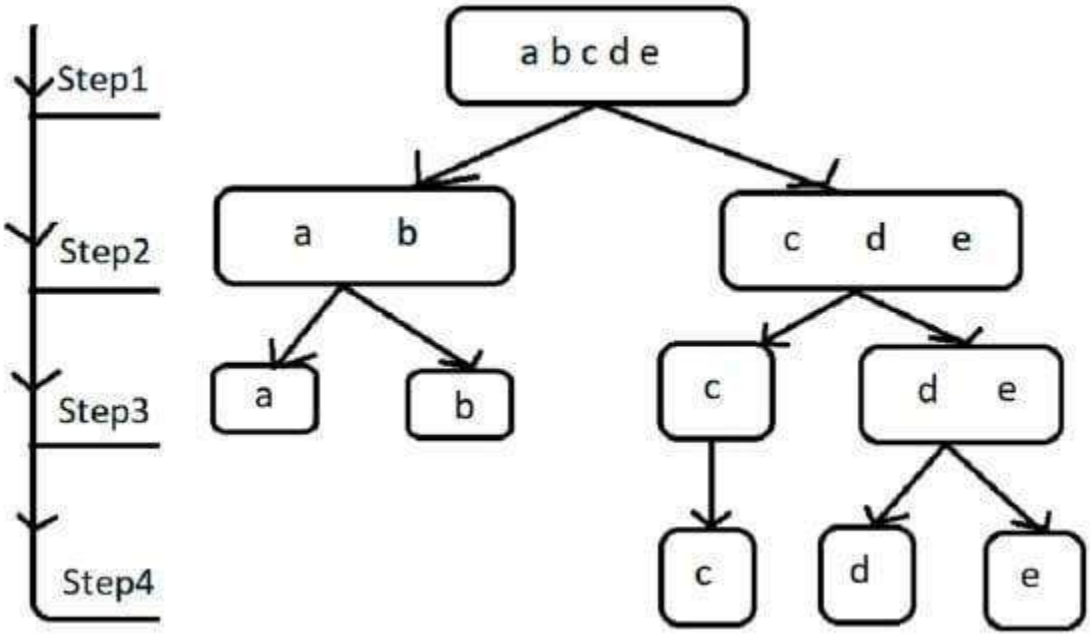


Figure 4.1: Structure of hierarchical clustering.

**Applications**

Consider you have a dataset of different animals based on their characteristics. Initially, each animal is its own cluster. The algorithm compares pairs of animals, merging the most similar ones into new clusters. For instance, it might first merge lions and tigers into a "big cats" cluster. Then, it might merge this new cluster with another containing leopards. This process continues until all animals are part of a single cluster or until you decide to stop based on a certain similarity threshold. Hierarchical clustering is powerful because it not only groups similar items but also provides a visual representation of their relationships in a tree structure.

### 4.3 DBSCAN clustering

DBSCAN, stands for Density-Based Spatial Clustering of Applications with Noise, is a clustering algorithm used in unsupervised machine learning. Its primary strength lies in identifying clusters of data points based on their density distribution within the feature space. DBSCAN groups data points based on their density, considering regions where points are closely packed together as potential clusters.

**Working principle**

When a dataset is given DBSCAN starts by picking a point and identifying nearby points within a specified distance. If there are enough nearby points, it marks the initial point as a core point and expands the cluster by connecting it to neighboring points. This process continues recursively, forming clusters based on the density of points. Points that are not part of any cluster or lack sufficient neighbors become noise. DBSCAN is a discovering clusters of varying shapes and sizes in data without requiring the number of clusters to be predefined, making it effective for datasets with irregular structures. The algorithm's flexibility and ability to handle noisy data make it widely used in various fields.

**Applications**

Consider a map of people in the city, and the objective is to find bustling neighborhoods. If you set DBSCAN to consider places where, let's say, at least 10 friends are within a 5-minute walk, it will identify clusters of friends hanging out together. These clusters could represent vibrant neighborhoods with a lot of social activity. In short DBSCAN helps you find groups of connected points (people in this analogy) based on how densely they're packed together. It's great for discovering natural groupings in data without needing to decide in advance how many groups you're looking for, just like finding lively neighborhoods without knowing how many there are in the city.

# CHAPTER 5
# ASSOCIATION ALGORITHMS IN UNSUPERVISED MACHINE LEARNING

Association in unsupervised machine learning refers to the task of discovering interesting relationships or patterns in large datasets without explicit labels. Unlike supervised learning, association rule mining doesn't involve predicting a target variable; instead, it focuses on finding associations among variables. One of the most common applications of association rule mining is in market basket analysis, where the goal is to uncover relationships between products frequently purchased together.

Association in unsupervised learning provides valuable insights into the underlying patterns and dependencies within a dataset. It finds applications not only in retail and market analysis but also in diverse fields such as healthcare, fraud detection, and recommendation systems, where understanding associations among variables is crucial for making informed decisions.

Association algorithms are essential tools in data mining and machine learning for discovering patterns and relationships within datasets. Among the commonly used algorithms, we explore Apriori algorithm, Éclat algorithm and FP-growth algorithm in this chapter. Apriori algorithm employs a level-wise approach to find frequent item sets by generating and pruning candidate item sets; the FP-Growth algorithm builds an FP-tree data structure to efficiently mine frequent item sets without candidate generation; and the Eclat algorithm uses a vertical data format and intersection-based approach to find frequent item sets directly. Additionally, algorithms like Prefix Span and SPADE are employed for sequential pattern mining, efficiently discovering frequent sequential patterns in sequential databases. Many of these algorithms utilize the FP tree structure for association rule mining, exploiting its efficiency in generating frequent item sets and association rules. Overall, association algorithms play a crucial role in uncovering valuable insights for various applications, including market basket analysis, sequential pattern mining, and recommendation systems.

## 5.1 Apriori algorithm

The Apriori algorithm is a classic method for discovering association rules in unsupervised machine learning, particularly applied to market basket analysis. It operates by iteratively

generating candidate item sets, evaluating their support, and extracting association rules based on predefined thresholds. The algorithm begins by identifying individual items and progressively forms larger item sets, leveraging the Apriori property that if an item set is frequent, all its subsets must also be frequent. This property helps prune the search space, enhancing computational efficiency.

The Apriori algorithm serves as a foundational technique in unsupervised machine learning, specifically tailored for discovering association rules, notably in market basket analysis. It functions through an iterative process of generating candidate item sets, evaluating their support, and extracting association rules based on predefined thresholds. Leveraging the Apriori property, which posits that if an item set is frequent, all its subsets must also be frequent, the algorithm efficiently prunes the search space, thereby enhancing computational efficiency.

**Working principle**

The algorithm commences by identifying individual items and progressively constructing larger item sets, harnessing the Apriori property to streamline the process. Subsequently, candidate item sets are evaluated for their support, which signifies the proportion of transactions containing those item sets. Any item sets failing to meet a specified minimum support threshold are discarded. Following this, the algorithm proceeds to generate association rules based on frequent item sets. These rules are assessed using confidence, which denotes the likelihood of the consequent occurring given the antecedent.

**Applications**

Apriori finds extensive application beyond retail, extending into diverse fields like healthcare and recommendation systems. In healthcare, it can unveil relationships within patient data, aiding in diagnosis or treatment planning. In recommendation systems, Apriori can discern patterns in user preferences, facilitating personalized recommendations.

Despite its effectiveness, Apriori may encounter scalability challenges with large datasets. To mitigate this, optimization techniques and alternative algorithms like FP-growth have been developed. However, Apriori's flexibility, enabled by user-defined parameters such as

minimum support and confidence levels, renders it suitable for a variety of datasets and applications. Nevertheless, striking a balance in parameter tuning is crucial, as overly stringent thresholds may overlook meaningful associations, while overly lax thresholds may yield numerous weak rules.

Apriori's association rules typically adhere to the "if-then" structure, offering valuable insights for decision-making. For instance, in a retail setting, a rule like "If a customer buys product A, then they are likely to purchase product B" can inform targeted marketing or product placement strategies.

Consider a grocery store analyzing customer purchase data using Apriori. The dataset comprises transactions listing various items bought together. Applying Apriori, the algorithm identifies frequent item sets and generates association rules. For instance, it discovers that customer’s frequently purchasing bread and milk also tend to buy eggs. Armed with this insight, the store strategically positions eggs near the bread and milk section, prompting additional purchases. This not only enhances the shopping experience but also maximizes sales through targeted product placement, showcasing Apriori's ability to unveil hidden associations in transaction data for business optimization and customer satisfaction enhancement.

### 5.2 Eclat algorithm

ECLAT (Equivalence Class Clustering and bottom-up Lattice Traversal) is a data mining algorithm used for frequent item set mining. It efficiently discovers frequent item sets in a dataset by focusing on the intersections of transactions rather than candidate generation. This makes it particularly useful for sparse datasets.

ECLAT (Equivalence Class Clustering and bottom-up Lattice Traversal) stands as a prominent data mining algorithm utilized for frequent item set mining. Unlike Apriori, ECLAT efficiently discovers frequent item sets in datasets by focusing on the intersections of transactions rather than candidate generation. This approach makes ECLAT particularly advantageous for sparse datasets, simplifying the process of finding meaningful associations in real-world transaction data.

**Working principle**

ECLAT operates by utilizing a vertical structure for data representation, which differs from the horizontal structure used by Apriori. It avoids explicit candidate generation and instead directly works with intersections, making it more efficient, especially for sparse datasets. By leveraging equivalence class clustering and bottom-up lattice traversal, ECLAT efficiently identifies frequent item sets by examining transaction intersections, offering insights into item combinations that occur frequently together.

**Applications**

Imagine having a dataset of customer transactions at a bookstore, where each transaction records the books purchased by a customer. ECLAT would analyze the intersections of books across transactions to identify frequent combinations. For instance, if many customers who bought "Harry Potter" books also purchased "The Lord of the Rings," ECLAT would recognize this association. By focusing on commonalities in transactions, ECLAT efficiently discovers such patterns, helping understand which books are often bought together.

Consider a scenario where a bookstore employs ECLAT to analyze customer purchase data. Upon analyzing the dataset, ECLAT identifies frequent item sets, such as the frequent combination of "Harry Potter" books and "The Lord of the Rings." Armed with this insight, the bookstore can implement targeted marketing strategies, such as bundling these books together or offering discounts on related items. Additionally, the store can optimize its layout by strategically placing these frequently purchased items near each other, enhancing the shopping experience and potentially increasing sales. This real-life example illustrates how ECLAT efficiently uncovers meaningful associations in transaction data, enabling businesses to make informed decisions to improve customer satisfaction and drive revenue.

### 5.3 FP-growth algorithm

The FP-growth algorithm, widely used in unsupervised machine learning for association rule mining, begins by scanning the dataset to tally the frequency of each unique item, creating a frequency table. Sorting items in descending order based on their frequency facilitates the

identification of the most frequent items. The algorithm then constructs an FP-tree, a tree-like structure that succinctly represents relationships and frequencies of item sets in the dataset. Each path from the root to a leaf node in the FP-tree signifies a frequent item set.

The FP-growth algorithm is a widely utilized method in unsupervised machine learning for association rule mining. It starts by scanning the dataset to determine the frequency of each unique item, creating a frequency table. Sorting items based on their frequency enables the identification of the most frequent items. The algorithm then constructs an FP-tree, a compact structure representing item set relationships and frequencies. Each path from the root to a leaf node in the FP-tree represents a frequent item set.

**Working principle**

FP-growth operates by recursively removing each frequent item and its associated branches from the FP-tree to generate conditional pattern bases, creating subsets of the dataset. The algorithm is applied iteratively to these conditional pattern bases, generating sub problems until all frequent item sets are discovered. Finally, association rules are derived based on user defined minimum support and confidence thresholds, unveiling meaningful associations between items in the dataset.

**Applications**

Consider a grocery dataset with transactions like {bread, milk, eggs} and {bread, butter}. Using FP-growth, the algorithm efficiently identifies frequent item sets like {bread, butter} and {milk, bread}, revealing patterns such as the co-occurrence of bread and butter. This process makes FP-growth a powerful tool for tasks like market basket analysis, where understanding item associations aids in making informed decisions.

In a telecom company's call records, the FP-growth algorithm might reveal that customers frequently subscribing to both phone and internet services are also likely to have TV services. This insight can guide targeted promotions and service bundling, enhancing customer satisfaction and contributing to business growth. FP-growth enables the identification of such

meaningful associations, empowering businesses to make data-driven decisions and improve overall performance.

Association rules in unsupervised machine learning offer versatile applications across various domains. In retail, market basket analysis helps optimize product placement and promotions by revealing associations between frequently purchased items. Healthcare data analysis utilizes association rules to identify patterns in patient records, facilitating personalized healthcare interventions and treatment strategies. Web usage mining enhances user experience by analyzing navigation patterns on websites and providing relevant content recommendations. In network security, association rules contribute to threat detection by identifying patterns in network traffic, aiding in the prevention of cyber security threats. Additionally, applications extend to e-commerce, text mining, human resources analytics, and educational data mining, showcasing the adaptability of association rules in extracting meaningful insights from diverse datasets.

# CONCLUSION

In conclusion, our exploration of supervised and unsupervised machine learning has illuminated the distinct advantages and limitations each paradigm brings to the table. Supervised learning, with its reliance on labeled data, excels in tasks like classification and regression, offering precise predictions when training examples are abundant. However, its Achilles' heel lies in the need for labeled datasets, which can be labor-intensive and challenging to obtain. On the other hand, unsupervised learning, adept at discovering hidden patterns in unlabeled data, thrives in scenarios where explicit guidance is limited. Yet, the subjectivity in interpreting results poses a challenge. In real-life applications, supervised learning finds its stride in areas like healthcare diagnosis and recommendation systems, while unsupervised learning proves invaluable in customer segmentation and anomaly detection. Embracing both methodologies strategically, we can leverage their strengths to address diverse challenges, exemplifying the dynamic and complementary nature of supervised and unsupervised machine learning in real-world contexts.

# BIBLIOGRAPHY

**1. Baştanlar, Yalin, and Mustafa Özuysal. "Introduction to machine learning." miRNomics: MicroRNA biology and computational analysis (2014): 105-128**

**2. Machine Learning: What It is, Tutorial, Definition, Types - java point. (n.d.). Www.javatpoint.com. Retrieved February 19, 2024, from https://www.javatpoint.com/machine-learning .**

**3. Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.**

**4. Follow, M. (2020, September 1). Implementation of lasso regression from scratch using python. GeeksforGeeks. https://www.geeksforgeeks.org/implementation-of-lasso-regression-from-scratch-using-python/ .**

**5Gawali, Suvarna. "Linear Regression in Machine Learning." AnalyticsVidhya,9June2021, https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/.**

**6. Kassambara, Alboukadel. Practical guide to cluster analysis in R: Unsupervised machine learning. Vol. 1. Sthda, 2017.**

**7. Charu C. Aggarwal,"DATA CLUSTERING Algorithm and Application",CRC Press ,2014.**