**Visualisation And Visualisation Categories**

Visualisation is a way to represent information.

* Data visualisation helps data analysts understand complex relationships between variables such as helping to identify patterns and trends of data.
* Visualisation Chart Categories:
* Comparison
* Composition
* Distribution
* Relationship

**Comparison** (Temple University, 2023)

It suits for comparing variables:

* To find the highest and lowest points in data.
* To understand the increasing and decreasing values.

Comparison charts such as:

* Column chart - small data and oriented vertically
* Bar chart - complexed and detailed data, long x-axis labels, and oriented horizontally
* Line chart - analysing trends
* Choropleth - geographical data

**Composition** (Temple University, 2023)

* It is suited for comparing a part to a whole.
* It is in a percentage form.
* Composition charts such as:
  + Pie chart - small data and clear differences between variables
  + Stacked bar/column chart - multiple levels of the same variables

**Distribution** (Temple University, 2023)

* It is suited for analysing distribution analysis such as mean, range, and outliers.
* Distribution charts such as:
  + Line chart - analysing trends
  + Scatter plot - small data and highlighting the similarities between variables
  + Bar/Column histogram - relationship of a variable over a set of categories

**Visualisation Table and Charts**

**Table** (EasyBI, 2023)

* Table is suitable for illustrating a small number of variables and data points.
* It is often used when:
  + Represent comparison, composition, or relationship visualisation.
  + Data can be easily interpreted from a table.
  + Individual values are needed to be compared.
  + Require precise values

**Example: Table for Comparing Sales of Three Products by Quarter**

|  |  |  |  |
| --- | --- | --- | --- |
| **Quarter** | **Product A (units)** | **Product B (units)** | **Product C (units)** |
| Q1 | 500 | 300 | 450 |
| Q2 | 600 | 400 | 500 |
| Q3 | 550 | 350 | 480 |
| Q4 | 620 | 420 | 520 |

**Usage of this table**:

* It clearly shows a comparison of sales between the three products across quarters.
* Useful for analyzing which product’s sales increased or decreased during each period.
* Provides precise figures, making it easy to interpret and calculate averages or trends.

**Column Chart** (EasyBI, 2023)

* It is used for comparing different values.
* It shows the comparison of individual values between columns.
* It can be used to compare values between different categories or value changes in the same category.

A graph of green bars

Description automatically generated with medium confidence

**Data suited for a Column Chart: ( เหมาะกับข้อมูลแบบไหน )**

* Comparing sales over different time periods (daily, monthly, yearly)
* Comparing the performance of different products or services
* Comparing statistics across different categories, such as regional sales or user demographics by age group

A Column Chart is ideal for data that involves distinct categories or groups where comparing values visually makes it easier to interpret patterns and trends.

**Column/Bar Histogram Chart** (EasyBI, 2023)

Column and Bar Histogram Charts are suitable for presenting data that is comparative and shows distributions. The choice of chart type depends on the nature of the data and the purpose of the presentation.

* It is used to represent the distribution and relationships of a variable over a set of categories.
* For example, a distribution of the sizes of pumpkins in a pumpkin festival.

A graph of pumpkins by size

Description automatically generated

**Data suited for a Chart**

**Suitable Uses for Column/Bar Histogram Charts:**

1. **Category Comparison:** Useful for comparing values across different categories, such as comparing sales of various products over time.
   * **Example:** Comparing the sales figures of products A, B, C, and D for the first quarter of the year.
2. **Data Distribution:** Useful for displaying the frequency of data within different ranges or categories.
   * **Example:** Showing the number of users in different age groups (e.g., 10-20, 21-30, 31-40 years).
3. **Sequential Comparison:** Useful for comparing sequential data over time, such as annual revenue changes.
   * **Example:** Comparing the annual revenue of a company over the past five years.

**Additional Examples:**

* **Column Chart:** Comparing monthly sales figures of different stores. For instance, Store A had sales of 100, 120, and 150 units in January, February, and March respectively.
* **Bar Chart:** Showing the frequency of different product types sold in a month, such as the number of products in categories A, B, C, and D sold in April.

These charts help in making data clearer and more understandable for viewers.

**Stacked Column/ Bar Chart**

A Stacked Column/Bar Chart is suitable for displaying data that includes subcategories within main categories. It is useful for visualizing the distribution or proportion of each subcategory within a group and for comparing the total values across different groups.

* It is used to represent composition.
* It shows multiple levels of the same variables.

A graph of sales and a number of individuals

Description automatically generated with medium confidence

**Data suited for a Chart: ( เหมาะกับข้อมูลแบบไหน )**

**Examples of suitable data:**

* **Sales by region:** The chart can show the total sales for each region and break them down by product categories, such as food, beverages, and other goods within the same region.
* **Election results by district:** It can display the total votes for each district, with a breakdown of votes by political parties within each district.
* **Quarterly revenue by type:** The chart can present the total revenue of a company for each quarter, with subcategories for revenue sources like product sales, services, and other income.

In summary, Stacked Column/Bar Charts are ideal for presenting multi-dimensional data with main and subcategories to show the proportion of each within a group.

**Bar Chart**

A Bar Chart is suitable for displaying data that involves comparisons between categories or groups. It is often used for quantitative or qualitative data that can be divided into distinct categories, such as the number of items, sales figures for different months, or the number of users in different segments.

* **It is used when:**
  + The names of categories are long.
  + The number of categories is more than seven and less than fifteen.
  + Representing negative numbers.

A graph with green and white bars

Description automatically generated

**Data suited for a Chart: ( เหมาะกับข้อมูลแบบไหน )**

If you want to compare the sales of four different products in a store, say Product A, B, C, and D, you can use a Bar Chart to represent the sales of each product with vertical or horizontal bars. Each bar will represent the sales figures for one product, such as:

* Product A sold 100 units
* Product B sold 150 units
* Product C sold 75 units
* Product D sold 200 units

The Bar Chart allows us to quickly see that Product D had the highest sales.

**Line Chart**

A Line Chart is suitable for displaying data that changes over time or data that follows a sequence of events, such as showing trends or changes over time. It can also be used to compare more than one dataset.

* It is used to analyse trends.
* It emphasises the continuation of values.

A graph showing the amount of sales

Description automatically generated with medium confidence

**Data suited for a Chart**

Examples:

1. **Showing monthly sales**: In a business, a Line Chart can be used to display sales figures throughout the year, allowing you to see whether sales increase or decrease each month.
2. **Displaying daily temperature changes**: It can be used to show how the temperature changes each day over the course of a week.
3. **Monitoring network bandwidth usage**: A Line Chart can track network bandwidth usage hourly to identify peak usage times.

A Line Chart helps to clearly visualize trends and changes in data over time.

**Bubble Plot**

A Bubble Plot is suitable for visualizing data with three or more dimensions. In a Bubble Plot, the X and Y axes represent the two main variables, and the size of the bubbles indicates a third variable, such as quantity or magnitude.

* It represents three or more dimensions of data.
* In other words, it is a scatter plot with other dimension(s).
  + A scatter plot compares two values, and then adds bubble size as a third variable.
  + If bubbles have similar sizes, you can add labels.

A graph showing the price of product

Description automatically generated

**Data suited for a Chart**

**Example**:  
Suppose you want to display sales data for different months. In this case:

* The **X** axis represents the months.
* The **Y** axis shows the revenue from sales.
* The size of the bubbles represents the number of items sold in each month.

A Bubble Plot helps to clearly visualize which months had the highest sales, revenue, and item quantities, with bubble size indicating the magnitude of the third variable

**Chropleth/Map Chart**

Choropleth maps are ideal for visualizing data that is related to geographical areas such as countries, states, or districts. They display data using color gradients to represent different values within these regions.

* It represents geographically related data.

A map of the world

Description automatically generated

**Data suited for a Chart**

**Types of Data Suitable for Choropleth Maps:**

1. **Population Data**: Showing the number of people in different regions, such as the population of each state in the U.S.
2. **Economic Data**: Representing metrics like income or Gross Domestic Product (GDP), for instance, GDP of different countries or states.
3. **Public Health Data**: Displaying disease prevalence across regions, such as COVID-19 infection rates by country.
4. **Education Data**: Illustrating educational attainment levels across regions, such as the percentage of people with higher education in various provinces or countries.
5. **Environmental Data**: Showing pollution levels or air quality across different areas.

**Examples:**

* **Disease Spread**: A choropleth map could illustrate the rate of influenza infections across U.S. states, with darker colors indicating higher infection rates and lighter colors indicating lower rates.
* **Economic Growth**: A choropleth map might display GDP per capita across different provinces of a country, making it easier to see economic disparities across regions.

Choropleth maps help in understanding complex data by showing how it varies across different geographic areas.

**Data Analytics and Machine Learning**

**Big Data Analytics Steps**

● Discovery - Learn the business domain, identify a problem, and plan required resources.

● Data Collection - Consider types of data and data sources.

● Data Preparation and Storage - Pre-process data and add data into the data file/database.

● Data Processing and Analytics - Process data and analyse data.

● Visualisation - Design visualisation and visualise analysed data as tables, plots, or graphs.

**Data Analytics**

● Raw data itself does not have a meaning until it is processed into information.

● Analytics is a process for creating information.

● Different goals need different technologies, algorithms, or frameworks for data analytics.

● For example,

○ To predict something from the data

○ To find patterns in data

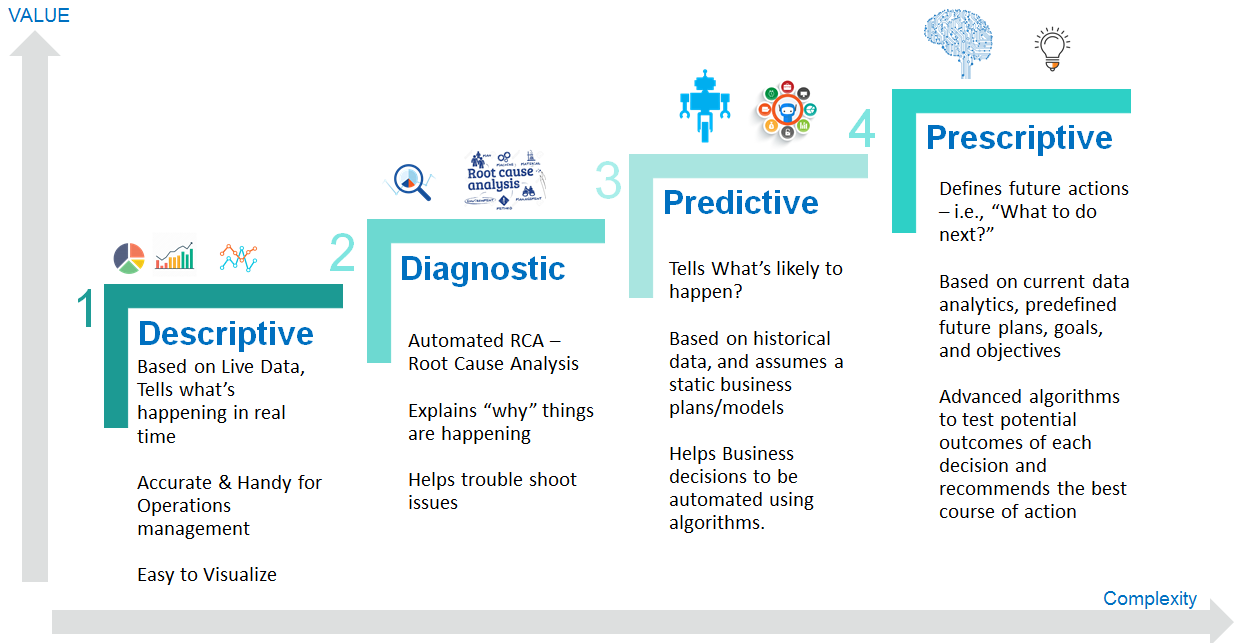
○ To find relationships between data

**Data Analytics** is the process of analyzing raw data to extract meaningful insights that can be used for decision-making or solving problems. Data analytics can take various forms, such as statistical analysis, data mining, predictive analytics, or descriptive analytics, depending on the goal of the analysis.

**Examples of Data Analytics Applications:**

1. **Marketing**: Data analytics is used to analyze customer behavior, predict market demand, and adjust marketing strategies to target the right audience. For example, analyzing online shopping data to identify popular products and using that information to create promotions or advertising campaigns.
2. **Healthcare**: It helps analyze patient data to predict disease trends and improve treatment plans. For example, analyzing health check-up data to forecast risks for diseases like diabetes or heart conditions.
3. **Manufacturing**: Data analytics is used to analyze production data to reduce waste, improve efficiency, and forecast the need for raw materials. For instance, analyzing factory production data to minimize waste and enhance product quality.
4. **Finance**: Banks and financial institutions use data analytics to detect fraudulent activities by analyzing credit card usage patterns or to recommend financial products to customers based on their financial profiles.
5. **Sports**: It helps analyze athlete performance data to plan competitive strategies. For example, collecting data on soccer players' performance to make decisions about substitutions and game plans.

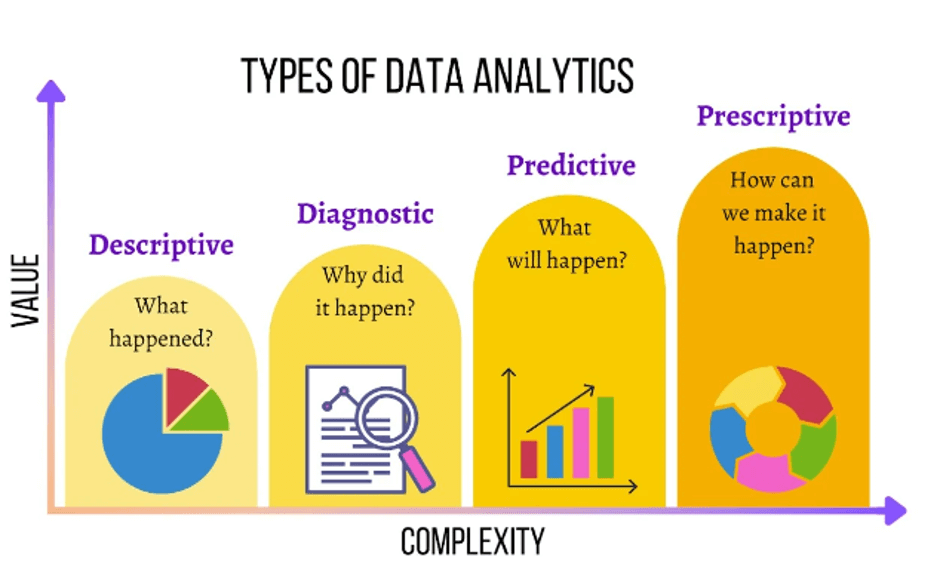
**● Data analytics can be categorised into 4 categories:**

○ Descriptive analytics

○ Diagnostic analytics

○ Predictive analytics

○ Prescriptive analytics



**Descriptive Analytics**

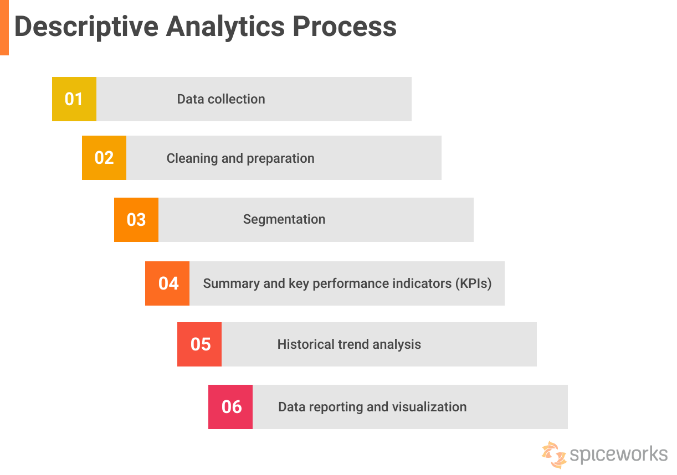
Descriptive analytics is the process of analyzing historical data to summarize and understand past events. It involves using techniques such as summarizing tables, creating charts, and calculating statistical metrics (e.g., averages, medians, minimums, and maximums) to provide insights into patterns, trends, and behaviors in the data. Descriptive analytics helps organizations understand what has happened in the past.

● It analyses the past data to present the analysed data in a summarised form.

● It answers the question “What has happened?”.

● Example: number of reactions of all Lives on Facebook of the year 2023 and which Live gets the highest number of reactions.

● Linear algebra, linear regression, and basic statistics functions such as counts, maximum, minimum, mean, top-N, percentage, etc. can be used for descriptive analytics.



**Examples of Descriptive Analytics Use Cases:**

1. **Retail Business**: A retail company might analyze sales data to see which products sell best during different times of the year. For example, a grocery store could examine monthly sales data to identify which items, like sweets or beverages, are most popular during festive seasons.
2. **Finance**: A bank might analyze customer transaction data to track spending patterns and identify which products customers spend the most on, or what types of transactions are most frequent.
3. **Healthcare**: Hospitals might use descriptive analytics to analyze patient data, such as hospital admission records, diagnoses, and treatments, to observe trends in diseases among different patient groups.
4. **Marketing**: A company could use descriptive analytics to evaluate the performance of past marketing campaigns. For instance, they may analyze customer clicks and responses to advertisements to summarize which strategies were most effective in boosting sales.

Descriptive analytics is crucial for helping organizations understand past data, which can inform future decision-making and strategic planning.

**Advantages**

1. **Understanding Basic Trends**: Helps in understanding the basic trends and patterns in historical data, such as sales figures over a period or customer numbers each month.
2. **Ease of Use**: The tools and techniques used for descriptive analytics are often less complex and can be done with common tools like Excel or Power BI.
3. **Decision Support**: Provides foundational information about past events that can support current or future decision-making.
4. **Time Efficiency**: Since it deals with existing data, results can be generated quickly.

**Disadvantages**

1. **No Future Predictions**: Descriptive analytics does not predict future events or trends. It only describes what has already happened.
2. **Lacks Actionable Insights**: While it provides useful information, it may not suggest the best actions or solutions for current issues.
3. **Complexity in Analysis**: Analyzing large or complex datasets can be challenging and may require sophisticated techniques.
4. **Risk of Misinterpretation**: There is a risk of misinterpreting data if not understood properly, which can lead to incorrect conclusions.

Descriptive analytics is a crucial starting point for understanding data and can serve as a foundation for more advanced analysis, such as predictive or prescriptive analytics.

**Diagnostic Analytics**

**Diagnostic Analytics** is a type of data analysis that focuses on identifying the causes of problems or events. This form of analysis delves into data to answer the question, "Why did this happen?" It often uses statistical techniques such as **correlation analysis**, **pattern recognition**, **causal analysis**, and **drill-down data examination**. Diagnostic analytics is typically performed after **Descriptive Analytics**, which tells what happened.

● It analyses past data to diagnose the reasons why certain events happened.

● Therefore, it answers the question of “Why did it happen?”.

● Example: From all Lives on Facebook, why does early morning or late night reach a higher number of reactions? Does work time affect the social used by users?

● Linear algebra, linear regression, graph analysis, clustering, etc. can be used for diagnostic analytics.

**Examples of Usage:**

1. **Marketing Analysis**
   * An e-commerce company notices a drop in sales over the past month (revealed through Descriptive Analytics). They use Diagnostic Analytics to explore why this decline occurred, such as analyzing customer behavior, comparing the use of promotions, or evaluating competitor changes.
2. **Healthcare Analysis**
   * A hospital finds that certain patients have slower recovery rates than expected. Diagnostic Analytics is used to examine the factors affecting recovery, such as patient medical history, medications, or differences in treatment methods.
3. **Manufacturing Analysis**
   * A factory experiences frequent breakdowns of certain machines. Diagnostic Analytics is used to investigate the causes of the downtime, such as checking the machine's condition, maintenance schedule, or environmental factors in the workspace.

Diagnostic Analytics helps organizations understand the underlying factors behind events or problems, enabling them to plan more effectively for corrections or preventative measures in the future.

**Benefits**

With diagnostic analytics, you can determine why certain events happened and fix the main problems causing them. When you know why things happened, you can make better choices later and adjust your plans.

These helpful insights can improve processes, help solve problems, and mitigate risks. Instead of just dealing with the signs of a problem, diagnostic analytics can help you address the real causes behind it.

You can learn from what happened in the past and use that knowledge to make things better in the future.

**Disadvantages**

Working with large data sets with complex relationships and many variables can be challenging.

It can be tricky to determine why something happened, especially when you’re trying to distinguish it from related events. Just because two events seem connected doesn’t mean one caused the other—this is the causality vs. correlation issue.

Additionally, unaccounted biases or overlooked variables can invalidate conclusions about why something happened, making helpful insights less accurate.

Lastly, diagnostic analysis focuses on historical data, which doesn’t necessarily predict future events. For a more complete picture, you should complement diagnostic analysis with other types of analytics.

**Predictive Analytics**

Predictive analytics is an advanced form of data analytics that attempts to answer the question, “What might happen next?” As a branch of data science for business, the growth of predictive and augmented analytics coincides with that of big data systems, where larger, broader pools of data enable increased data mining activities to provide predictive insights. Advancements in big data machine learning have also helped expand predictive analytics capabilities.

The growth of predictive and augmented analytics coincides with that of big data systems, where broader pools of data enable increased data mining activities to provide predictive insights. Advancements in big data machine learning have also helped expand predictive analytics capabilities.

● It predicts the occurrence of an event or the likely outcome of an event.

● It answers the question of “What is likely to happen?”.

● The accuracy of prediction depends upon the quality of existing data.

● **Example**: What Facebook Live content will attract more views, thus, number of reactions? Will the number of reactions increase if you Live on Facebook in the early morning or late night?

● Linear algebra, linear regression, clustering, graph analysis, Bayesian inference, Markov Chain Monte Carlo, text analysis, Hidden Markov Model, etc. can be used for predictive analytics.

**How does predictive analytics work?**

Data scientists use predictive models to identify correlations between different elements in selected datasets. Once data collection is complete, a statistical model is formulated, trained, and modified to generate predictions.

The workflow for building predictive analytics frameworks follows five basic steps:

1. **Define the problem**: A prediction starts with a good thesis and set of requirements. For instance, can a predictive analytics model detect fraud? Determine optimal inventory levels for the holiday shopping season? Identify potential flood levels from severe weather? A distinct problem to solve will help determine what method of predictive analytics should be used.
2. **Acquire and organize data**: An organization may have decades of data to draw upon, or a continual flood of data from customer interactions. Before predictive analytics models can be developed, data flows must be identified, and then datasets can be organized in a repository such as a [data warehouse](https://cloud.google.com/learn/what-is-a-data-warehouse) like [BigQuery](https://cloud.google.com/bigquery).
3. **Pre-process data**: Raw data is only nominally useful by itself. To prepare the data for the predictive analytics models, it should be cleaned to remove anomalies, missing data points, or extreme outliers, any of which might be the result of input or measurement errors.
4. **Develop predictive models**: Data scientists have a variety of tools and techniques to develop predictive models depending on the problem to be solved and nature of the dataset. Machine learning, regression models, and decision trees are some of the most common types of predictive models.
5. **Validate and deploy results**: Check on the accuracy of the model and adjust accordingly. Once acceptable results have been achieved, make them available to stakeholders via an app, website, or data dashboard.

**The benefits of predictive analytics are as follows:**

1. **Informed Decision-Making**: Predictive analytics helps customers make decisions based on data. By leveraging predictive models, businesses can make more informed and accurate decisions.
2. **Real-Time Answers**: Predictive analytics can provide real-time answers. Trained predictive models can analyze data in real time and deliver immediate insights.
3. **Understanding Complex Problems**: Predictive analytics can help customers understand complex issues. It can reveal hidden patterns in the data more quickly and accurately.
4. **Competitive Advantage**: Predictive analytics can help companies gain a competitive edge. Companies using predictive analytics have a competitive advantage over those that do not, as they can forecast future events more accurately.

**Prescriptive Analytics**

● It uses multiple prediction models to predict various outcomes and the best action for each outcome.

● It answers the question of “What can we do to make it happen?”.

● It predicts the outcomes based on the current actions.

● **Example**: Should you live on TikTok or YouTube to get more number of views? Does a photo or video attract more views in the early morning? What content type should you post in the early morning to get more views from users with the age above 20?

● Clustering, graph analysis, optimisation, Bayesian inference, Markov Chain Monte Carlo, text analysis, Hidden Markov Model, etc. can be used for prescriptive analytics.

**How do prescriptive analytics work?**

Prescriptive analytics involves using algorithm-based models to generate specific recommendations or decisions based on a defined problem. For instance, if a HR manager needs to upskill a team but some members lack necessary skills, prescriptive analytics can recommend that those members first acquire the required skills through another course before proceeding. The effectiveness of these recommendations depends on the accuracy of the data and the model used, and each recommendation is tailored to the specific situation.

Prescriptive analytics provides actionable recommendations based on data analysis, offering several key **benefits:**

1. **Revenue Generation**: Helps businesses understand customer preferences and optimize sales strategies for cross-selling and up-selling.
2. **Gross Margin Management**: Identifies the optimal product mix to focus on, improving productivity and profitability.
3. **Expense Reduction**: Enhances inventory management, reduces long-term storage costs, and minimizes manual processes, leading to better expense control.

**Applications**:

* **Travel and Transportation**: Optimizes pricing, sales pitches, and route planning using customer data.
* **Fracking and Oil Production**: Improves safety and efficiency in extraction processes.
* **Healthcare**: Combines with predictive analytics to tailor interventions, improve performance, and manage diseases effectively.

**Challenges**:

* **Defining Fitness Functions**: Requires a deep understanding of the business to create effective solutions.
* **Human Bias**: Models can be biased based on human input; machine learning could mitigate this.
* **Complex Constraints**: Solutions must navigate various constraints and rules, which can complicate model accuracy.

The future of prescriptive analytics lies in integrating real-time event processing, distributed computing, and advanced algorithms to deliver timely, actionable insights and support decision-making.

**Differences between descriptive, predictive, and prescriptive analytics**

* Descriptive analytics offers business intelligence insights into what has occurred.
* Predictive analytics forecasts possible outcomes.
* Prescriptive analytics works to find the best possible solution from a variety of options

**Machine Learning**

Big Data Analytics and Machine Learning

* As Big Data works with a vast amount of data (volume) where the accuracy (variety and veracity) and speed (velocity) are of concern for creating information (value), machine learning can be applied to Big Data Analytics.
* Machine Learning (ML) is a trained program to learn from data and then does some actions.

ML is used when:

* Complex problems need to be solved
* Non-stable system where new data can always be ingested into the system
* Improving performance for solving problems that require various rules to find the solution.

**Machine Learning Terminology**

* Feature (independent variable) represents a property of an observation. For example, a row represents an observation while their columns represent different features (different independent variables).
  + A category feature is a descriptive feature (qualitative), for example, a name.
  + A numerical feature is a numerical feature (quantitative), for example, a number of reactions on FBLiveTH.
* Label (dependent variable) is a variable that an ML system learns to predict.
  + A categorical label represents a category, for example, an ML application learns to predict a category of news articles such as business and technology categories. These categories are categorical labels.
  + A numerical label represents a numerical, for example, an ML application learns to predict the price of a house. The price is a numerical label.
* Model estimates the relationship between the dependent variable (label) and independent variable (feature) in a dataset.
* A model can predict the value of a dependent variable (label) using independent variables (features). For example, using the number of reactions, status published, and status type in FBLiveTH to predict a future number of comments.
* In other words, we can say that a model is a mathematical function that uses features as input and outputs a value (label).
* Training Data, also called “training set”, is the data that is used for training a model.
* Training data can be categorised into:
* Labelled dataset - a dataset has at least one of the columns that contains a label.
* Unlabeled dataset - a dataset does not have a column that contains a label.
* Test Data, also called “test set”, is a dataset that is used for testing a model after the model has been trained for evaluating the performance of the model.

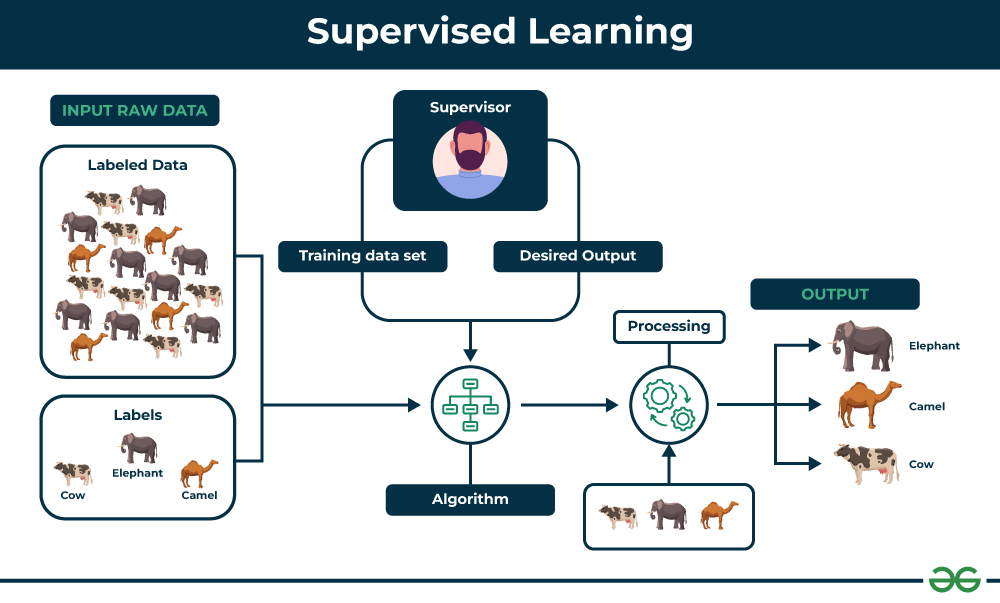
**Supervised and Unsupervised Learning Algorithms**

**Supervised Learning Algorithms**

* A supervised learning algorithm uses a labelled dataset to train a model, where these labels in a training dataset may be generated manually or sourced from other datasets.
* A supervised learning algorithm aims to predict a label for a data.

Supervised learning can be separated into two types of problems when [data mining](https://www.ibm.com/topics/data-mining): classification and regression:

* **Classification** problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox. Linear classifiers, support vector machines, decision trees and [random forest](https://www.ibm.com/topics/random-forest) are all common types of classification algorithms.
* **Regression** is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression, and polynomial regression.



**How it work of the Process**

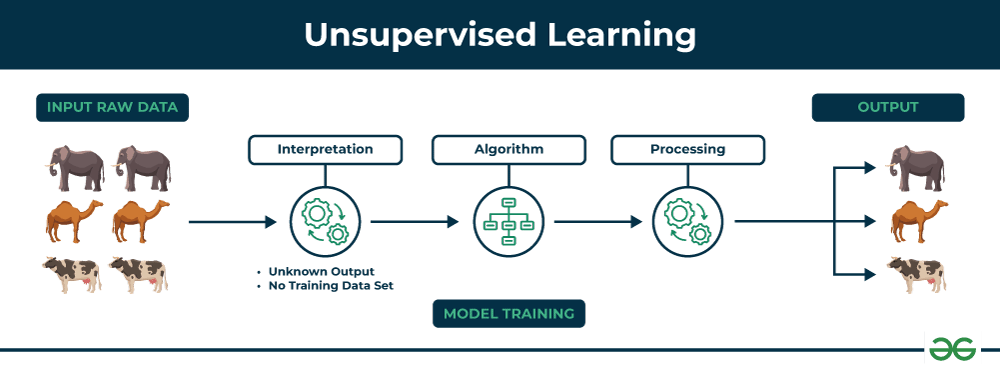
1. **Raw Data (Input Raw Data):** This is data that has not yet been processed, such as images of different animals. For example, this could include images of elephants, camels, and cows.
2. **Labeled Data:** This refers to raw data that has been annotated with labels or tags. For instance, in an image of an elephant, we would label it as "elephant."
3. **Training Data Set:** This is a subset of the labeled data that is used to train the model.
4. **Algorithm:** This is a set of instructions that guides the model to learn from the training data to identify the relationship between input data and output.
5. **Processing:** The model processes the training data repeatedly through multiple iterations to improve its accuracy in predictions.
6. **Output:** After the model has been trained, it can be used to make predictions on new, unseen data. For example, when presented with an image of an animal not included in the training data set, the model can predict what type of animal it is.

**Unsupervised Learning Algorithms**

* An unsupervised learning algorithm uses an unlabeled dataset to train a model.
* An unsupervised learning algorithm aims to discover a structure in data.

Unsupervised learning models are used for three main tasks: clustering, association and dimensionality reduction:

* **Clustering** is a data mining technique for grouping unlabeled data based on their similarities or differences. For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, and so on.
* **Association** is another type of unsupervised learning method that uses different rules to find relationships between variables in a given data set. These methods are frequently used for market basket analysis and recommendation engines, along the lines of “Customers Who Bought This Item Also Bought” recommendations.
* **Dimensionality reduction** is a learning technique that is used when the number of features (or dimensions) in a given data set is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.



**How it works:**

1. **Input Raw Data:** The raw data coming in does not specify the type or category of the data. For example, a collection of animal images without indicating which images are of elephants, horses, or cows.

2. **Interpretation:** The model interprets the raw data by considering various attributes such as shape, color, and size to identify relationships between the data.

3. **Algorithm:** The model selects an appropriate algorithm for analyzing the data in an unsupervised manner, such as:

* **Clustering:** Grouping similar data together, for example, clustering animal images by type.
* **Dimensionality Reduction:** Reducing the dimensions of the data to make it easier to visualize, such as transforming data with many dimensions into just two or three dimensions for plotting on a graph.

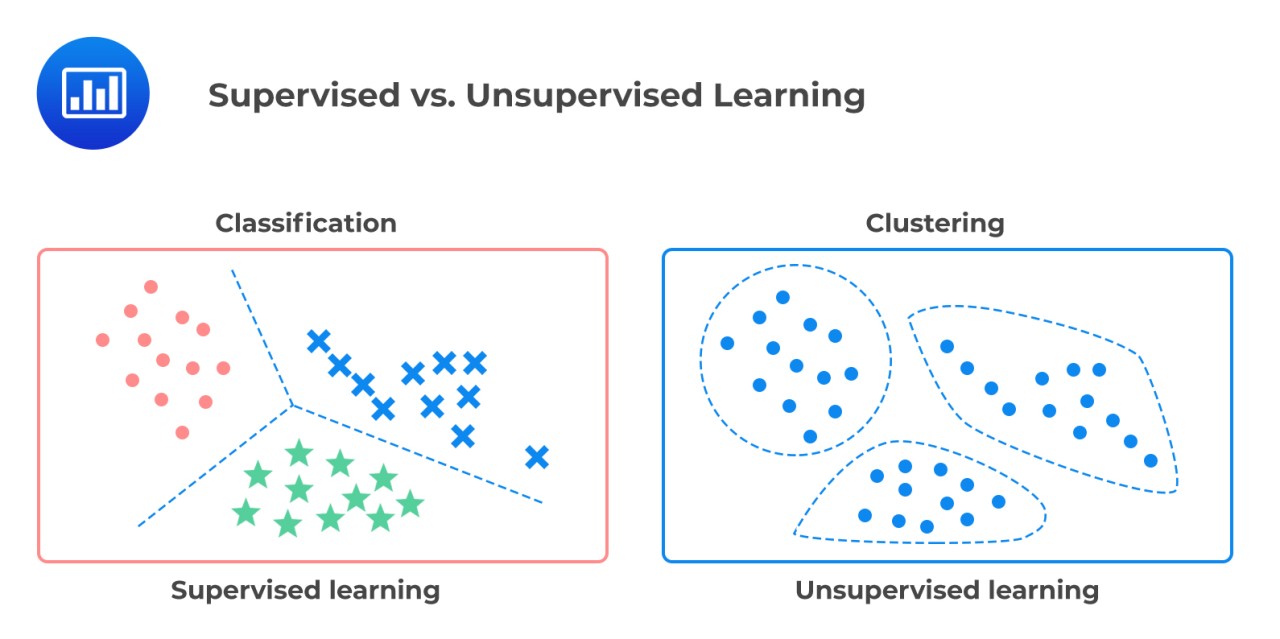
4. **Processing:** The model processes the data according to the chosen algorithm to find hidden structures or patterns.

5. **Output:** The result will be a grouping of data organized by similarity, or data presented in a reduced-dimensional format.

**Different Supervised Learning vs. Unsupervised Learning:**

1. **Supervised Learning vs. Unsupervised Learning**:
   * **Supervised Learning**: Uses data with a target or goal, such as predicting sales using past data as the target and various factors as inputs. Examples include Regression and Classification.
   * **Unsupervised Learning**: Uses data without a target, such as analyzing customer groups based on purchase behavior without predefined groups. Examples include Clustering.
2. **Data Splitting**:
   * **Supervised Learning**: Typically involves splitting data into Training and Testing Sets to prevent Overfitting and to assess the model’s accuracy.
   * **Unsupervised Learning**: Data splitting into Training and Testing Sets is not usually required, as accuracy is not directly measured.
3. **Algorithms**:
   * **Supervised Learning**: Uses algorithms like Random Forests for Regression and Classification.
   * **Unsupervised Learning**: Uses algorithms like Association Rules for finding relationships.
4. **Evaluation**:
   * **Supervised Learning**: Evaluated based on accuracy metrics such as Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE).
   * **Unsupervised Learning**: Evaluated based on practical suitability, such as cluster density for Clustering, with a focus on usability rather than direct accuracy metrics.

In summary, Supervised Learning involves data with targets and often requires data splitting to avoid Overfitting, while Unsupervised Learning deals with data without targets and may not require data splitting.



**SparkML**

* DataFrame from SparkSQL is used as an ML dataset.
* The transformer includes feature transformers and learned models.
  + It inputs a DataFrame, then transforms (transform()) the input to output which is another DataFrame.
  + Transformer examples are such as cutting some features, adding more features, manipulating the current features, or formatting data.
* The estimator is an algorithm or learning algorithm that is used to fit (fit()) data into the machine learning model (transformer).

**Pipeline**

* A pipeline in machine learning is a sequence of algorithms for processing and learning from data.
* In SparkML, a pipeline includes a sequence of stages where each stage can be either a transformer or an estimator.

**Components of a Pipeline**

* **Transformer:** These stages prepare the data for the model. For instance, a Tokenizer breaks text into individual words, and a HashingTF converts these words into numerical features.
* **Estimator:** This is the final stage where a model is fitted to the transformed data. Examples include LogisticRegression, DecisionTree, or RandomForest.

A diagram of a machine learning process

Description automatically generated

**The Pipeline in Action (Referencing the Diagram)**

1. **Raw Text:** This is the unprocessed input data, like a customer review.
2. **Tokenizer:** Splits the text into individual words.
3. **HashingTF:** Converts words into numerical features using a hashing technique.
4. **Logistic Regression:** Trains a model to predict a binary outcome (e.g., positive or negative sentiment) based on the numerical features.

A diagram of a logistic process

Description automatically generated

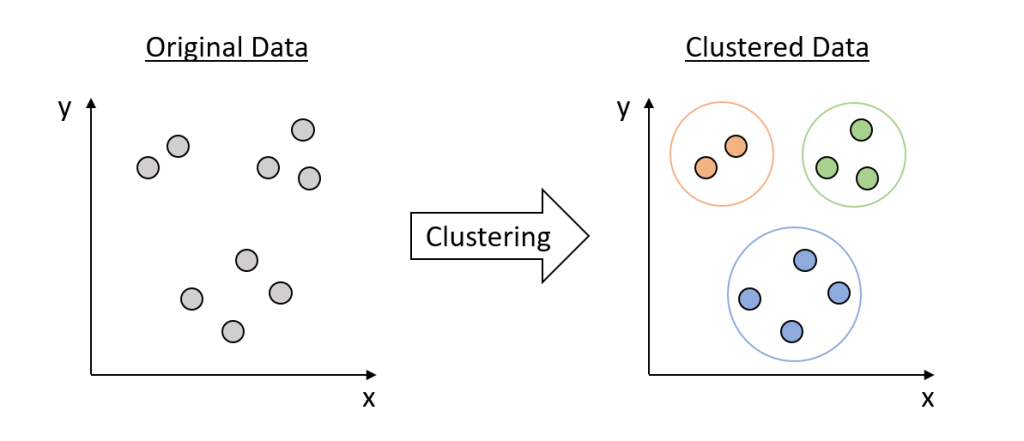
A diagram of a software development process

Description automatically generated

**Clustering**

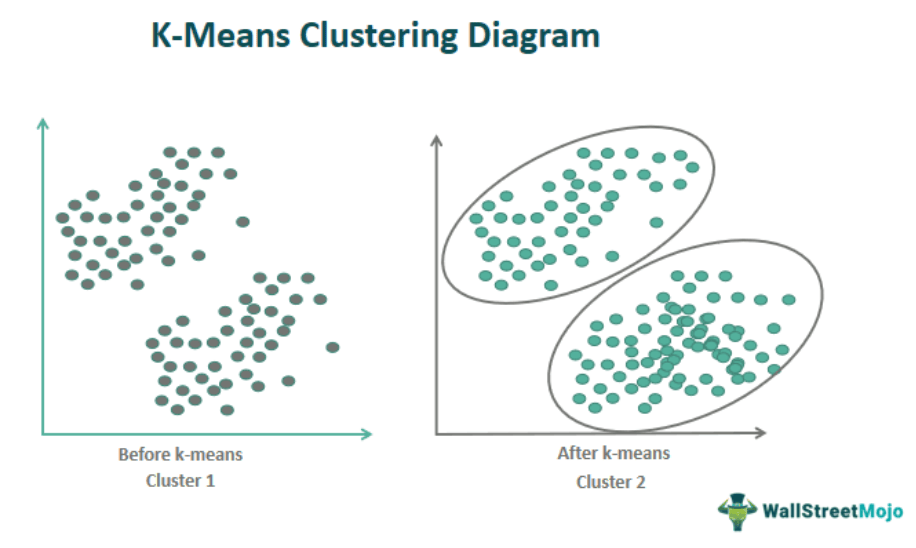
Clustering

* Clustering is one of the unsupervised learning algorithms where the labels will not be determined to apply to the clusters.
* Therefore, clustering is used with unlabeled datasets.
* Clustering finds the similarities between objects based on the object attributes and groups the similar objects into clusters.
* The dataset is split into clusters, where elements in the same cluster are more similar to each other than elements in the other clusters.
* Clustering tasks in unsupervised learning algorithms are such as:
  + K-means
  + Principal Component Analysis (PCA)
  + Singular Value Decomposition (SVD)



K-Mean

* K-Mean is one of the clustering algorithms.
* It is usually used with numerical data.
* It is used for grouping data points into k number of groups.
* The data points in the same group are similar, whereas they are dissimilar to data points in the other groups.
* Steps:
  + Initialise k clusters using random k points (called the centroids)
  + For each new data point, find the distance to the centroid (Euclidean distance), then assign the point to the closest cluster.
  + Update the cluster centroid by calculating mean values.
  + Repeat steps until cluster centroids are not changed or until they reach the maximum number of iterations.



K-Mean Evaluation Example

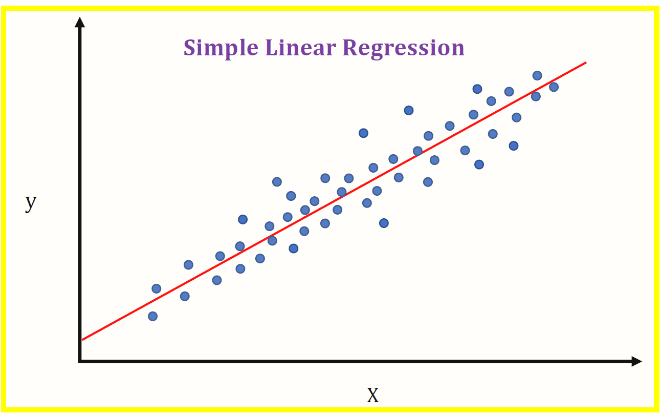
* Silhouette analysis is used to measure the quality of clusters.
* It indicates how far data is in clusters.
* The measure is in the range of [-1, 1]:
  + -1 means the data might be assigned to the wrong cluster.
  + 0 means data in the clusters are very close.
  + 1 means data in the clusters are far away from each other.
* The k number that provides the highest average of the silhouette is the best k for the given data.

**Regression**

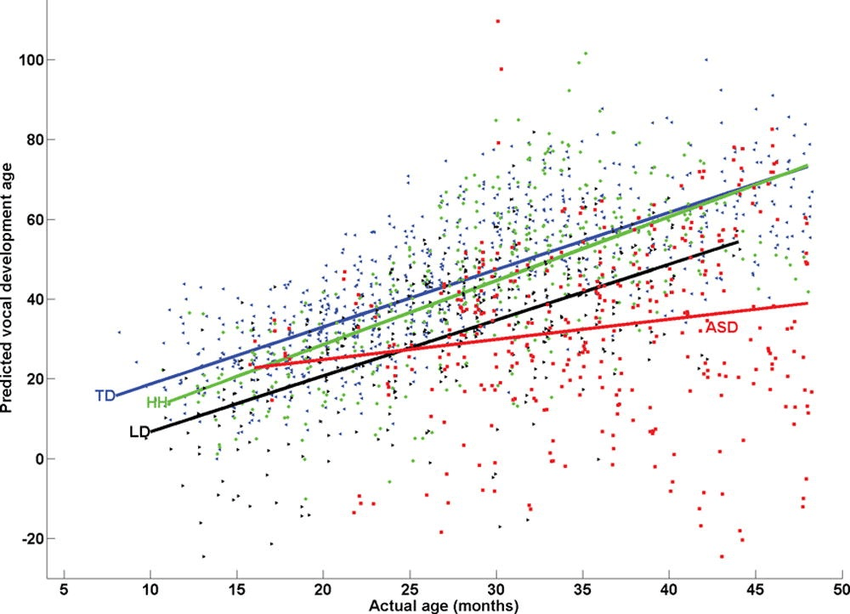
* Regression is one of the supervised learning algorithms.
* Regression is used to predict a continuous variable (real number) from a set of features.
* Thus, it is a training algorithm for predicting a numerical label.
* For example, from a number of reactions, you might want to predict how many people are likely to love the FB live.
* A model trained by a regression algorithm:
  + Simple regression - it involves one label and one feature.
  + Multiple regression - it involves one label and multiple features.
  + Multivariate regression - it involves multiple labels and multiple features.

Regression algorithms are used to model the relationship between dependent and independent variables. Here's a breakdown of the different types of regression you mentioned:

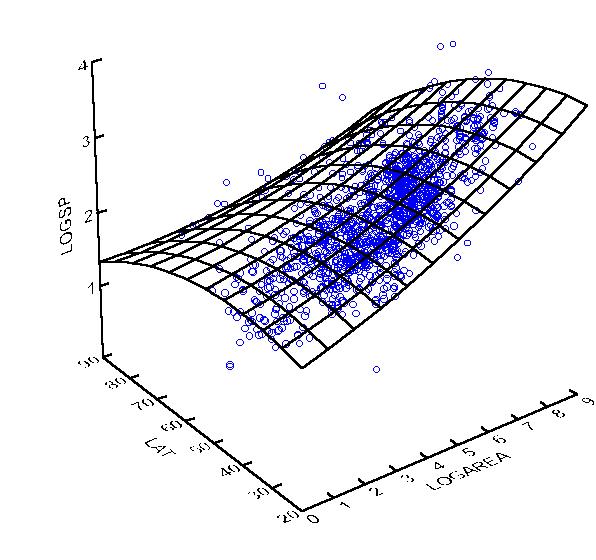
1. **Simple Regression**:
   * **Definition**: This involves **one dependent variable** (label) and **one independent variable** (feature).
   * **Example**: Predicting house prices based on square footage alone.



1. **Multiple Regression**:
   * **Definition**: This involves **one dependent variable** (label) and **multiple independent variables** (features).
   * **Example**: Predicting house prices based on square footage, number of rooms, and location.



1. **Multivariate Regression**:
   * **Definition**: This involves **multiple dependent variables** (labels) and **multiple independent variables** (features).
   * **Example**: Predicting house price and rental price based on various factors like square footage, number of rooms, and location.



In summary:

* **Simple regression**: One label, one feature.
* **Multiple regression**: One label, multiple features.
* **Multivariate regression**: Multiple labels, multiple features.

Here’s a comparison of **Simple Regression**, **Multiple Regression**, and **Multivariate Regression** in English:

**Similarities:**

1. **Main Goal**: All three are types of **regression models**, with the primary goal being to predict the value of a dependent variable (label) based on independent variables (features) that have a linear relationship.
2. **Basic Principle**: They all rely on mathematical and statistical methods to find the relationship between variables using **linear equations**, such as y=a+bxy = a + bxy=a+bx (for Simple Regression) or y=a+b1x1+b2x2+...+bnxny = a + b\_1x\_1 + b\_2x\_2 + ... + b\_nx\_ny=a+b1​x1​+b2​x2​+...+bn​xn​ (for Multiple Regression).
3. **Model Training**: In all three types, the model learns from the data to calculate the coefficients of the equation that best predicts future values.

**Differences:**

1. **Number of Dependent Variables (Labels)**:
   * **Simple Regression**: Involves **one dependent variable** and **one independent variable**.  
     *Example*: Predicting exam scores (dependent variable) based on study hours (independent variable).
   * **Multiple Regression**: Involves **one dependent variable** but **multiple independent variables**.  
     *Example*: Predicting exam scores (dependent variable) based on study hours, sleep quality, and stress level (multiple independent variables).
   * **Multivariate Regression**: Involves **multiple dependent variables** and **multiple independent variables**.  
     *Example*: Predicting both exam scores and health status (multiple dependent variables) based on study hours, sleep quality, and stress level (multiple independent variables).
2. **Model Complexity**:
   * **Simple Regression**: The simplest model, as it only considers one independent variable affecting the prediction.
   * **Multiple Regression**: More complex, as it considers multiple independent variables influencing the outcome.
   * **Multivariate Regression**: The most complex, as it involves multiple dependent variables and multiple independent variables being considered simultaneously.

* Regression tasks in supervised learning
* algorithms are such as:
  + Linear Regression
  + Decision Tree
  + Ensembles

**Linear regression**

* Linear regression uses training data to fit the linear model with a coefficient.
* It is used for predicting the relationship between independent variable (feature) and dependent variable (label)
* For example,
  + A linear model is y = a + bx, where y is the label, x is a feature, a is the intercept of the line, and b is the linear regression coefficient.
  + As we use a training dataset, y, a, b, and x are known, thus, we can train the model to predict the label y.
* Model hyperparameters:
  + family - it can be binary (binary classes) or multinomial (multi-classes)
  + elasticNetParam - it is a floating point in the range of 0 to 1 which is used to specify the mix of L1 and L2 regularisation. L1 regularisation mostly creates the value of 0. L2 regularisation creates a value reaching to 0 but not completely 0. elasticNetParam should be tuned by testing with different values.
  + fitIntercept - it determines whether or not the model should fit an intercept. It can be True or False.
  + regParam - it determines how much weight or coefficient to assign for the regularisation where the value must be greater than or equal to 0.
  + standardization - it can be True or False depending upon whether or not the inputs were standardised before being used in the model.

**Note:** Regularisation is a technique for minimising the loss function (quantifies the difference between predicted and actual values) and preventing overfitting (model learns from noise as trained with too much data) or underfitting (model is unable to learn the training data resulting in low performance).

Training parameters:

* maxInter - it is the total number of iterations, where the default is 100.
* tol - it specifies a threshold and stops the iterations before the number of iterations reaches the maxInter.
* weightCol - it is the name of the weight or coefficient column. You can weigh the labels you know are correct more than the labels you do not know.

Prediction parameters:

* threshold - it is a probability threshold that identifies when to predict a class. It is a Double type within the range of 0 and 1.
* thresholds - it is an array of threshold values used for multiclasses.

**Mean Squared Error (MSE)** is the average of the squared difference between the predicted and

actual values.

* The smaller the MSE, the better because it means the model produces fewer errors, where
* the errors are measured by the dispersion of the data points from its mean).

**R-Squared (R2 = R ยกกำลัง 2)** or Coefficient of Determination is the measure of the strength of the relationship between the model and dependent variable (label).

* A high R2 means the model is good.
* Low R2 means the model is not better than the mean value.
* Negative R2 means the model is worse than the mean value.

**Mathematical Formula**:

A black background with white text

Description automatically generated

A black background with white text

Description automatically generated

**Key Concept**: The algorithm attempts to minimize the **Mean Squared Error (MSE)** between predicted and actual values.

**Advantages**:

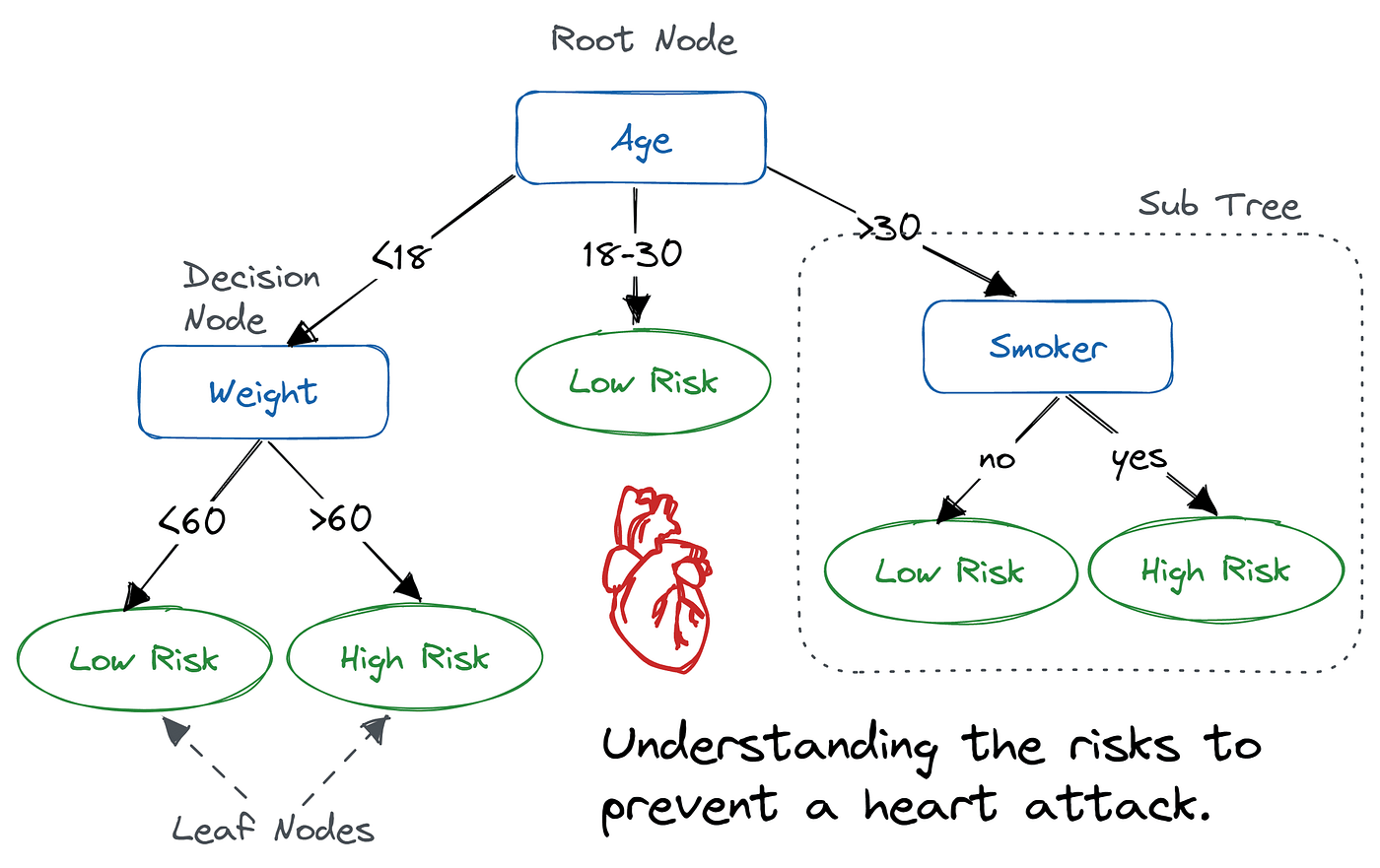
* Easy to implement and interpret.
* Efficient for small datasets.

**Disadvantages**:

* Assumes a linear relationship between the variables, which may not always hold.
* Sensitive to outliers.
* Performs poorly when multicollinearity (high correlation between independent variables) exists.

**Decision Tree Regression**

* **Overview**: A Decision Tree Regression model splits the dataset into subsets based on feature values. The tree is constructed by recursively partitioning the data into smaller regions to minimize prediction error within each partition.
* **Working Mechanism**:
  + The data is split at different decision points (nodes) based on the input features.
  + Each split aims to minimize the difference between actual and predicted values within the partitions, often using a measure like **Mean Squared Error (MSE)** for regression.
  + The process continues until a stopping criterion (e.g., minimum number of samples per leaf node) is met.
* **Key Concept**: Decision trees create a model that makes predictions by learning decision rules from the data.
* **Advantages**:
  + Non-linear relationships can be captured.
  + Easy to visualize and interpret.
  + Requires little data preprocessing, such as feature scaling or normalization.
* **Disadvantages**:
  + Prone to overfitting, especially if the tree grows too large.
  + Sensitive to small variations in data.
  + Poor generalization to unseen data compared to more advanced models like ensemble methods.



Decision Tree

**How the Decision Tree Works**

1. **Root Node:** The starting point, or "root" of the tree, is the **age** of the individual. This is the initial factor considered in the risk assessment.
2. **Decision Nodes:** As you move down the tree, you encounter decision nodes. These nodes represent **split points** based on the values of certain attributes. For example, the age attribute might be split into three ranges: less than 18, between 18 and 30, and over 30.
3. **Branches:** The branches extending from each decision node represent the different possible outcomes of that decision. For instance, if an individual is over 30, the branch would lead to the next node, which might consider another attribute like smoking status.
4. **Leaf Nodes:** The end points of the tree are called leaf nodes. These nodes represent the **final classification** or prediction. In this case, the leaf nodes indicate whether an individual is at "low risk" or "high risk" of a heart attack

**ในสไลด์**

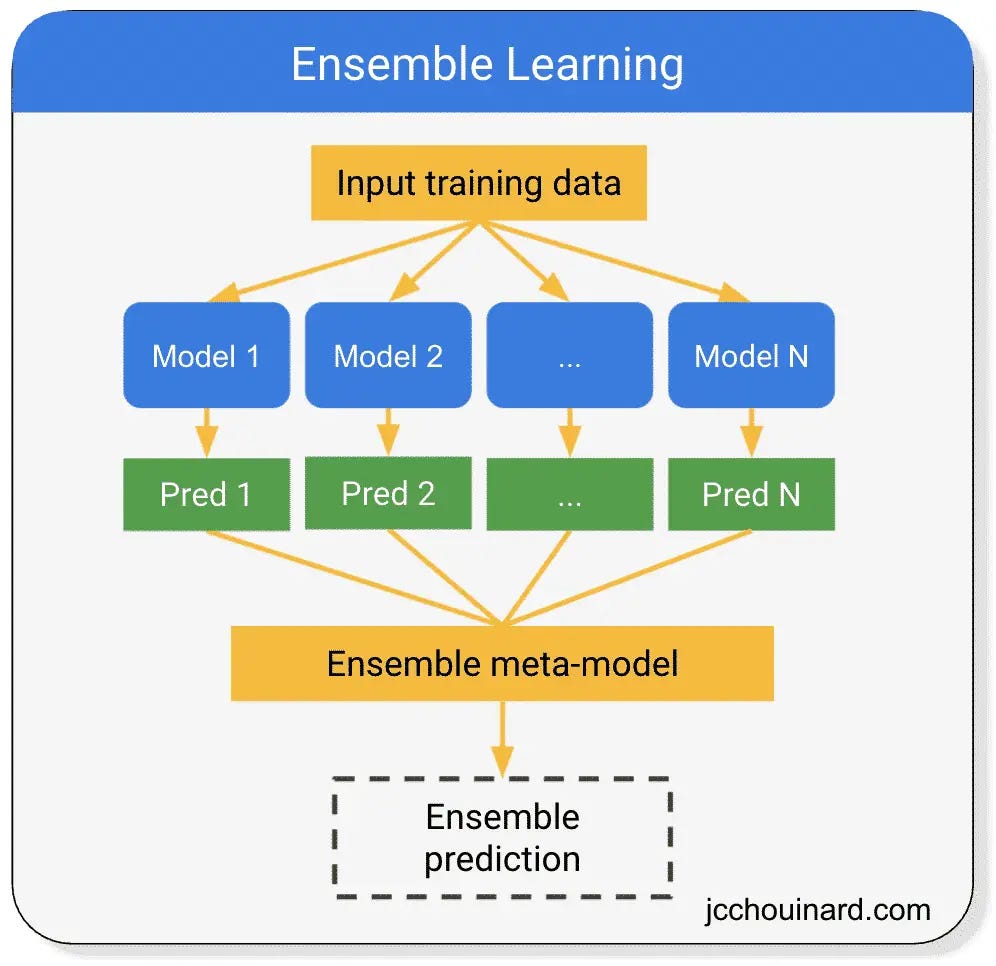
* Decision Trees can be used for both regression and classification.
* Decision trees for regression create output that is a single number per leaf node.
* Decision trees for classification create output that is a label per leaf node.
* It simply creates a big tree of decisions.
* Thus, it is easy for decision making.
* However, it has high cost consumption.

**Ensemble Methods**

* **Overview**: Ensemble methods combine multiple models (often called "weak learners") to improve the overall performance. The idea is that a group of weak models can form a strong model by averaging or combining predictions. In regression tasks, ensemble techniques can reduce variance and increase accuracy.
* **Types of Ensembles**:
  + **Bagging (Bootstrap Aggregating)**:
    - Bagging generates multiple models by training each model on random subsets of the training data (with replacement). The predictions of all models are averaged to produce the final output. The most well-known example of bagging is **Random Forest**.
    - **Advantages**: Reduces overfitting, less variance.
  + **Boosting**:
    - Boosting focuses on training models sequentially, where each subsequent model attempts to correct the errors made by the previous models. Popular boosting algorithms include **AdaBoost** and **Gradient Boosting**.
    - **Advantages**: Good at reducing bias and variance, better accuracy.
  + **Stacking**:
    - In stacking, multiple models are trained independently, and their predictions are combined using another model, often called a "meta-learner," to make the final prediction.
    - **Advantages**: Can improve accuracy by leveraging the strengths of different models.
* **Key Concept**: The strength of ensemble methods lies in combining the predictions of multiple models to improve robustness and accuracy.
* **Advantages**:
  + Reduces the risk of overfitting compared to individual models.
  + Often provides better generalization to new data.
* **Disadvantages**:
  + More computationally expensive.
  + Harder to interpret compared to a single decision tree or linear regression model.

**Ensemble Learning Explanation with Visual Representations of Each Method:**

* **Bagging (Bootstrap Aggregating):** The image will show the creation of multiple models from bootstrapped samples of the original data (with replacement). The results of each model will be combined (e.g., by averaging) to produce the final outcome.
* **Boosting:** The image will emphasize the sequential creation of models, where each new model aims to correct the errors made by the previous models. The models are combined to achieve more accurate results.
* **Stacking:** The image will describe the creation of multiple models, with the outputs of each model passed to a meta-learner model to learn how to combine those outputs together effectively.



Ensemble learning is a machine learning technique that combines multiple models to improve predictive performance. **Here's how it works:**

**1. Input Training Data:**

* The process begins with a dataset that will be used to train the ensemble model.

**2. Model Creation:**

* Multiple individual models are created from the training data. These models can be of different types (e.g., decision trees, neural networks, support vector machines) or even the same type.

**3. Predictions:**

* Each individual model makes a prediction on the same input data.

**4. Ensemble Meta-Model:**

* An ensemble meta-model is used to combine the predictions from the individual models into a final prediction. This meta-model can be a simple majority vote, weighted average, or more complex techniques like stacking or boosting.

**5. Ensemble Prediction:**

* The final prediction is made by the ensemble meta-model, which combines the predictions from the individual models.

**Key Benefits of Ensemble Learning:**

* **Improved Accuracy:** By combining multiple models, ensemble learning can often achieve higher accuracy than any individual model.
* **Reduced Overfitting:** Ensembles can help to reduce overfitting by averaging out the errors of individual models.
* **Increased Robustness:** Ensembles are more robust to noise and outliers in the data.
* **Better Generalization:** Ensembles can generalize better to unseen data.

**Common Ensemble Techniques:**

* **Bagging:** Creates multiple models by randomly sampling the training data with replacement.
* **Boosting:** Creates multiple models sequentially, with each model focusing on the errors of the previous models.
* **Stacking:** Combines the predictions of multiple models using a meta-model.

In summary, these regression algorithms vary in complexity and performance:

* **Linear Regression** is simple but limited to linear relationships.
* **Decision Trees** capture more complex relationships but can overfit without proper tuning.
* **Ensemble Methods** like Bagging and Boosting combine the strengths of many models, offering greater accuracy and robustness but at the cost of interpretability and computational power.

**Sure! Here's a comparison of Linear Regression, Decision Trees, and Ensembles**

**Differences:**

1. **Linear Regression**:
   * This method creates a linear model between independent variables and a dependent variable.
   * It is used for predicting continuous dependent variable values.
2. **Decision Trees**:
   * This method splits data into groups (nodes) based on the most important features.
   * It is suitable for both classification and regression tasks.
   * The output is a tree structure that is easy to interpret.
3. **Ensembles**:
   * This technique combines multiple models to improve accuracy.
   * It includes methods like Random Forests and Gradient Boosting.
   * It reduces overfitting and increases the robustness of the model.

**Similarities:**

* All three methods are techniques for modeling that use data to make predictions or outcomes.
* They can be applied to statistical analysis and machine learning problems.
* Performance metrics, such as accuracy and cross-validation, can be used to evaluate the effectiveness of all models.

Sure! Here are examples for each of the three methods: ( ยกตัวอย่างการใช้งานแต่ละอัน )

**1. Linear Regression:**

* **Example**: Predicting house prices based on features like square footage, number of bedrooms, and location. The model will find the best-fitting line that represents the relationship between these features and the price.

**2. Decision Trees:**

* **Example**: Classifying whether a customer will buy a product based on their age, income, and previous purchase behavior. The decision tree will split the data into branches based on the most significant features, leading to a classification (e.g., "Yes" or "No").

**3. Ensembles:**

* **Example**: Using a Random Forest model to predict whether a loan will be defaulted on. The Random Forest combines multiple decision trees to improve prediction accuracy and robustness. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions from all trees.

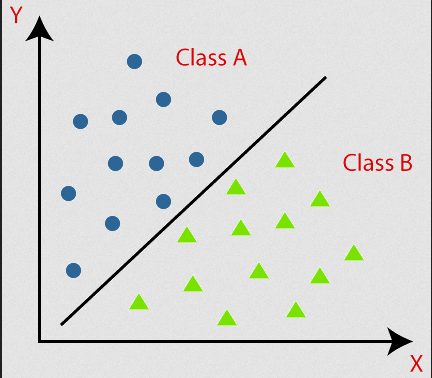
If you have any specific scenarios in mind or want more details about any example, let me know!

**Classification**

* Classification is a common type of supervised learning algorithm.
* It is a training algorithm for predicting a categorical label.
* Classification tasks can be grouped into:
  + Binary classification
  + Multiclass classification
  + Multilabel classification

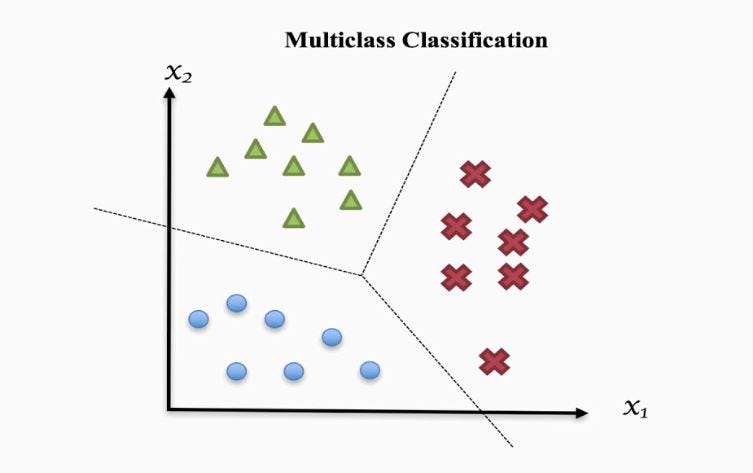
**Binary Classification**

* Binary classification is the most common.
* It classifies an observation into two labels which are positive and negative.
* Example:
  + Classify emails into a spam or not spam emails.



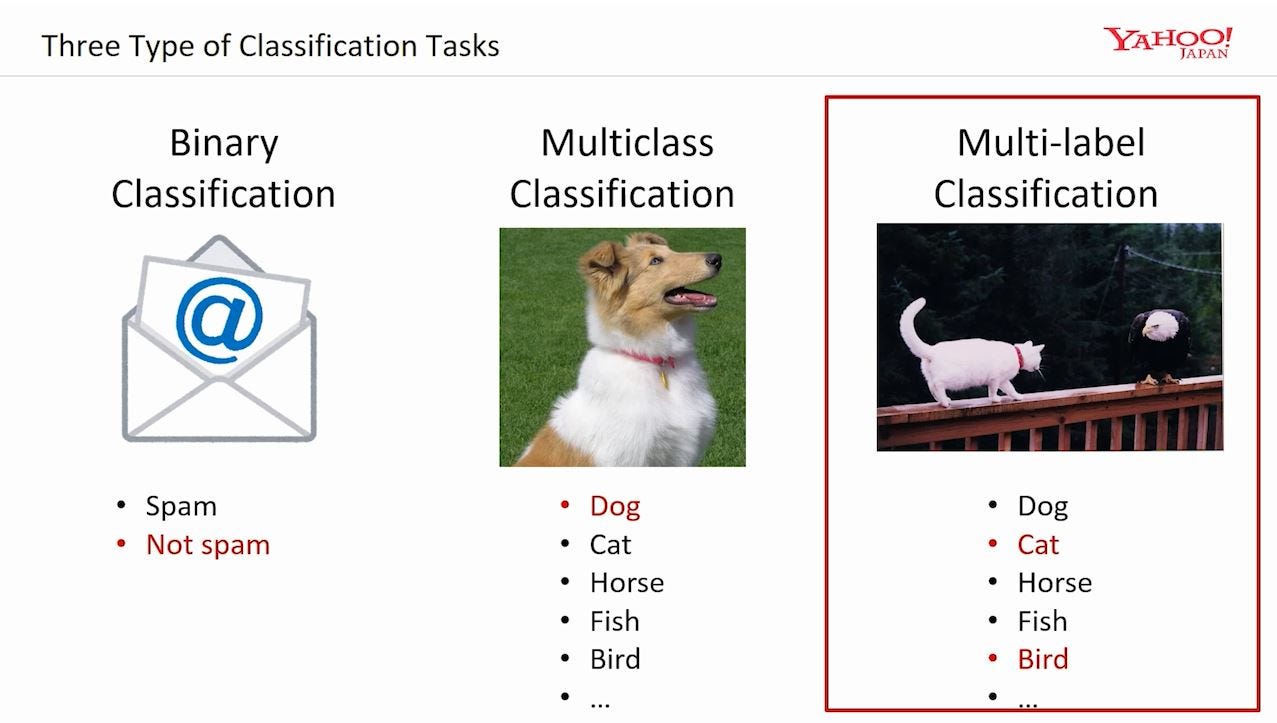
**Multiclass Classification**

* Multiclass classification classifies an observation into one label where the label is chosen from more than two labels.
* Example:
  + Predicting the weather from rainy, sunny, cloudy, etc. labels



**Multilabel Classification**

* Multilabel Classification predicts more than one label for the same observations.
* In other words, an input can produce multiple labels.
* Example:
  + A news article can be labelled as business and IT.



**ความแตกต่าง และ ความเหมือน**

Classification tasks can be divided into three main types:

1. **Binary Classification**:
   * **Description**: There are only two possible outcomes, such as "yes" or "no" (e.g., whether an email is spam or not).
   * **Example**: Predicting whether an email is spam or not (spam/not spam).
2. **Multiclass Classification**:
   * **Description**: There are more than two possible outcomes, but only one can be selected at a time (e.g., categorizing fruits as apple, banana, or orange).
   * **Example**: Classifying animals in an image as dog, cat, or rabbit.
3. **Multilabel Classification**:
   * **Description**: There are multiple possible outcomes that can be selected simultaneously (e.g., one image may contain multiple objects, such as a dog and a cat).
   * **Example**: Predicting emotions in a text that may convey multiple feelings, such as "happy" and "sad" in the same sentence.

**Differences**:

* **Number of Categories**: Binary has 2 categories, Multiclass has more than 2 but can choose only one at a time, Multilabel can select multiple categories simultaneously.
* **Usage**: Binary is used when there are only two options, Multiclass is used when there are multiple non-overlapping options, and Multilabel is used when there are overlapping choices.

**Similarities**:

* All three types fall under the category of classification tasks and utilize similar techniques, such as machine learning models, to predict outcomes.

**Classification Evaluation Example**

* Accuracy refers to the number of correctly classified over the total number of data.
* It is calculated as:

A black text on a white background

Description automatically generated

where TN is true negative, TP is true positive, FN is false negative, and FP is false positive.

* Precision refers to the number of actual positives over the total predicted positives.
* It is calculated as:

A mathematical equation with black text

Description automatically generated

* Recall refers to the number of positives over the total actual positives.
* It is calculated as:

A mathematical equation with black text

Description automatically generated

* F1 measure is the harmonic mean of precision and recall. It shows a balance between precision and recall.
* It is calculated as:

A close-up of a sign

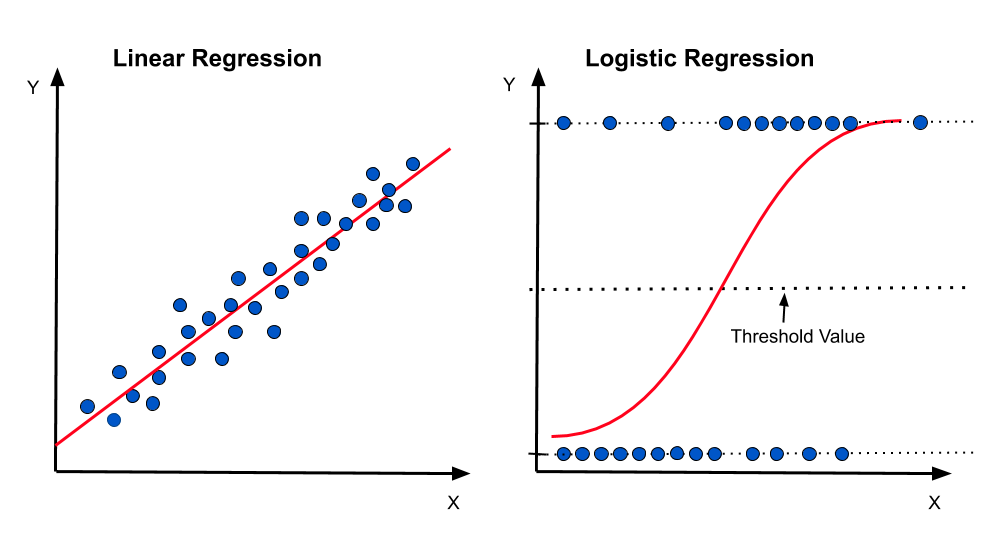
Description automatically generated

Classification tasks in supervised learning algorithms are such as:

* Logistic Regression
* Decision Trees
* Support Vector Machine (SVM)
* Ensembles

**Logistic Regression**

* Logistic Regression combines inputs (features) with the weights (coefficients) to get the probability of the class.
* The weights help to represent the importance of the feature.
  + Large weight means the variations of features significantly affect the outcome.
  + Smaller weight means the feature is less important.
* SparkML for Logistic Regression can use the same set of parameters as the Linear Regression.



**Linear Regression และ Logistic Regression แตกต่างกันตรงไหน มีตรงไหนเหมือนกัน และก็ และ 2 อย่างนี้ใช่ในกรณีไหน**

**Differences:**

1. **Type of Data Used**:
   * **Linear Regression**: Used for predicting continuous values, such as price or scores.
   * **Logistic Regression**: Used for predicting categorical outcomes, such as "yes" or "no" (binary classification) or multiple categories (multiclass classification).
2. **Regression Function**:
   * **Linear Regression**: Uses a linear function y=mx+by = mx + by=mx+b.
   * **Logistic Regression**: Uses a sigmoid function to keep outputs between 0 and 1: p=11+e−zp = \frac{1}{1 + e^{-z}}p=1+e−z1​, where zzz is the linear combination of independent variables.
3. **Evaluation Methods**:
   * **Linear Regression**: Evaluated using RMSE (Root Mean Squared Error) or R².
   * **Logistic Regression**: Evaluated using metrics like Accuracy, Precision, Recall, or AUC-ROC.

**Similarities:**

1. **Regression Basis**: Both methods use regression techniques to estimate outcomes.
2. **Independent Variables**: Both involve independent variables for predictions.

**Use Cases:**

* **Linear Regression**: Used when predicting continuous outcomes, like estimating house prices based on various features.
* **Logistic Regression**: Used when predicting classifications, such as whether a customer will buy a product based on available data.

I hope this clarifies the differences and similarities for you!

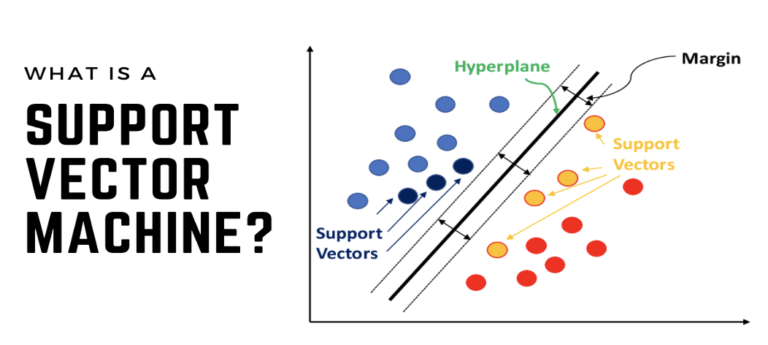
**Decision Trees Classification ใช้ได้ทั้ง** Regression และ Classification

* Decision Trees can be used for both regression and classification.
* Decision trees for regression create output that is a single number per leaf node.
* Decision trees for classification create output that is a label per leaf node.
* It simply creates a big tree of decisions.
* Thus, it is easy for decision making.
* However, it has high cost consumption.

**Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a powerful classification technique that finds the optimal hyperplane to separate different classes in the feature space. Here are some key aspects of SVM:

* **Margin Maximization**: SVM aims to maximize the margin between the closest points of different classes (support vectors). A larger margin typically leads to better generalization on unseen data.
* **Kernel Trick**: SVM can handle non-linearly separable data by transforming the original feature space into a higher-dimensional space using kernel functions (like polynomial or radial basis function). This allows SVM to create complex decision boundaries.
* **Soft Margin**: To accommodate noise and overlapping classes, SVM introduces a soft margin that allows some misclassifications. The trade-off between maximizing the margin and minimizing misclassifications is controlled by a parameter called C.
* **Applications**: SVM is widely used in text classification, image recognition, and bioinformatics due to its effectiveness in high-dimensional spaces.



**Example SVM**

Here are some examples of how Support Vector Machines (SVM) are used in various applications:

**1. Text Classification**

**Example**: Spam Detection

* **Description**: SVM can classify emails as spam or non-spam by analyzing the frequency of words and phrases. The model is trained on a labeled dataset of emails, learning to identify patterns associated with spam.

**2. Image Classification**

**Example**: Face Recognition

* **Description**: SVM is used to classify images based on features extracted from the images, such as edges, colors, and textures. In face recognition, SVM can differentiate between faces by creating decision boundaries in the feature space.

**3. Bioinformatics**

**Example**: Cancer Diagnosis

* **Description**: SVM can classify tumor samples as malignant or benign based on gene expression data. By analyzing the patterns in the gene expression profiles, SVM helps in early diagnosis and treatment planning.

**4. Handwriting Recognition**

**Example**: Digit Recognition

* **Description**: SVM is often used in optical character recognition (OCR) systems to identify handwritten digits. The model is trained on labeled digit images and learns to distinguish between different digits based on pixel intensity patterns.

**5. Intrusion Detection Systems**

**Example**: Network Security

* **Description**: In cybersecurity, SVM can be applied to detect anomalies in network traffic, identifying potential intrusions or malicious activities. By training on historical data of normal and anomalous traffic, SVM can classify new traffic patterns effectively.

**6. Sentiment Analysis**

**Example**: Social Media Sentiment

* **Description**: SVM can classify social media posts or comments as positive, negative, or neutral based on the textual features. This application helps businesses analyze public sentiment about their products or services.

These examples illustrate the versatility of SVM across different domains, making it a powerful tool for classification tasks.

**Ensembles ใช้ได้ทั้ง Regression กัย Classification**

**Recommendation System**

* A recommendation system is used to recommend a product or item to the user.
* The recommendation system learns from user’s behaviours and preferences in the past to recommend a product or item.
* Explicit preference - express preferences through ratings.
* Implicit preference - through observation such as the number of clicks, number of likes, number of loves.
* Recommendation forms:
  + Content based
  + Collaborative filtering
  + Hybrid (combination of content based and collaborative filtering)

**Content based**

* It recommends a product based on its characteristics that match the previous product which the user is interested in.
* It uses explicit preference to determine product similarity for making recommendations.
* For example, Netflix recommends movies based on the genre that a user often watches, such as an action, a romantic, or a comedy movie.

**Collaborative Filtering**

* It recommends a product to a user based on the preferences of users who have similar interests.
* It learns users with similar preferences and similar properties of products using data from rows in a tabular input dataset, where each row contains a user ID, product ID, and rating.
* The products that will be recommended to a user are the products with high rates from other users who have similar preferences.

**Collaborative Filtering with Alternating Least Squares**

* Alternating Least Squared (ALS) is a popular collaborative filtering recommendation system.
* ALS finds the k-dimensional feature vector for each user and product.
* ALS conducts a dot product of each user’s feature vector with each item’s feature vector, thus, it can approximate the user’s rating for that product.
* It requires a tubular input dataset where each row contains a user ID, product ID, and rating.
* Each rating can be an explicit (numerical rating) or an implicit (such as the number of visits to aparticular page).
* It uses input Dataframe to predict user’s ratings for products which have not yet been rated.

**Cold Start Problem**

* It arises when new users or products have no rating history.
* It also occurs when using a random split because users or products in the testing set are not in the training set.
* Spark will assign NaN prediction.
* This can ruin the ability of your model evaluation.
* Assigning NaN can be useful as you can design an overall system to fall back on default recommendations when a new user or new product is added to the system.
* Spark coldStartStrategy parameter is allowed to be used to drop any rows in the DataFrame of predictions that contain NaN values.
* Therefore, the evaluation can be conducted over non-NaN data in the Dataframe.

**Root Mean Square Error (RMSE)**

* Root Mean Square Error (RMSE) is the measure of the differences between predicted and actual
* values.
* The smaller the RMSE is the better of predictions from the model.