

Reporte de unidad 2 & 3

Minería de datos



**Santillan Torres Jesus Javier**

Maestra: Alondra Ramos

9° A

# INTRODUCCIÓN:

Data mining is the process of extracting knowledge from large datasets. It is a multidisciplinary field that combines elements of statistics, machine learning, database systems, and visualization. Data mining techniques are used to identify patterns, trends, and anomalies in data. This information can then be used to make data-driven decisions, solve business problems, and uncover hidden insights.

Here are some of the key concepts in data mining:

* Data: The data that is used for data mining can be structured, semi-structured, or unstructured. Structured data is organized in a regular format, such as a table in a database. Semi-structured data is not as organized as structured data, but it still has some structure, such as XML or JSON documents. Unstructured data is not organized at all, such as text documents, images, or videos.
* Algorithms: Data mining algorithms are used to extract patterns and trends from data. There are many different data mining algorithms, each with its own strengths and weaknesses. Some popular data mining algorithms include:
  1. Classification: Classification algorithms are used to predict the class of an instance. For example, a classification algorithm could be used to predict whether a customer will churn or not.
  2. Clustering: Clustering algorithms are used to group similar instances together. For example, a clustering algorithm could be used to group customers together based on their purchase history.
  3. Association rule mining: Association rule mining algorithms are used to find relationships between items. For example, an association rule mining algorithm could be used to find that customers who buy milk are also likely to buy bread.
  4. Anomaly detection: Anomaly detection algorithms are used to identify data points that are unusual or unexpected. For example, an anomaly detection algorithm could be used to identify credit card transactions that are likely to be fraudulent.
* Data mining process: The data mining process typically involves the following steps:
  1. Data preparation: The data is cleaned and preprocessed to make it ready for analysis.
  2. Data mining: The data is analyzed using data mining algorithms to extract patterns and trends.
  3. Interpretation: The results of the data mining are interpreted to make sense of the patterns and trends that have been found.
  4. Deployment: The results of the data mining are deployed to make data-driven decisions or solve business problems.

Data mining is a powerful tool that can be used to gain insights from large datasets. However, it is important to note that data mining is not a magic bullet. It can only be used to find patterns and trends that are already present in the data. It cannot be used to create new information or to predict the future.

Here are some of the benefits of data mining:

* Improved decision-making: Data mining can be used to improve decision-making by providing insights into customer behavior, market trends, and operational efficiency.
* Increased profitability: Data mining can be used to identify new opportunities for revenue growth, reduce costs, and improve customer satisfaction.
* Enhanced customer experience: Data mining can be used to personalize the customer experience by providing relevant recommendations and offers.
* Improved compliance: Data mining can be used to identify risks and compliance issues, such as fraud and money laundering.

# 

# 

# What is linear regression?

In statistics, linear regression is a linear approach for modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

In simple linear regression, the relationship between the independent variable (x) and the dependent variable (y) is modeled by a line. The equation of the line is:

y = a + bx

where:

* a is the y-intercept
* b is the slope of the line
* x is the independent variable
* y is the dependent variable

The goal of linear regression is to find the values of a and b that minimize the error between the predicted values of y and the actual values of y. This is done using a process called least squares estimation.

Linear regression is a powerful tool that can be used to predict the value of a dependent variable based on the value of one or more independent variables. It is used in a wide variety of fields, including:

* Business: Linear regression can be used to predict sales, customer churn, and other business metrics.
* Finance: Linear regression can be used to predict stock prices, interest rates, and other financial markets.
* Healthcare: Linear regression can be used to predict the risk of disease, the effectiveness of treatments, and other healthcare outcomes.
* Education: Linear regression can be used to predict student performance, the effectiveness of teaching methods, and other educational outcomes.

Linear regression is a relatively simple statistical technique, but it can be very effective at predicting the value of a dependent variable based on the value of one or more independent variables. It is a valuable tool for businesses, organizations, and individuals who want to make better decisions based on data.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# What is logical regression?

Logical regression, also known as logistic regression, is a statistical model that is used to predict the probability of a binary outcome, such as yes or no, based on a set of independent variables. The independent variables can be continuous or categorical.

Logistic regression is a type of regression analysis, but it is different from linear regression in that it does not produce a continuous value for the dependent variable. Instead, it produces a probability value between 0 and 1. This probability value can then be used to make a binary decision, such as whether to approve a loan or not.

The logistic regression model is defined by the following equation:

P(Y = 1 | X) = 1 / (1 + e^(-(b0 + b1X)))

where:

* P(Y = 1 | X) is the probability of the outcome being 1, given the values of the independent variables (X)
* b0 is the y-intercept
* b1 is the slope coefficient
* X is the vector of independent variables
* e is the exponential function

The goal of logistic regression is to find the values of b0 and b1 that minimize the error between the predicted values of P(Y = 1 | X) and the actual values of P(Y = 1 | X). This is done using a process called maximum likelihood estimation.

Logistic regression is a powerful tool that can be used to solve a variety of problems, including:

* Predicting customer churn: Logistic regression can be used to predict the probability of a customer churning, based on their purchase history, demographics, and other factors.
* Detecting fraud: Logistic regression can be used to detect fraudulent transactions, based on the characteristics of the transactions and the customer.
* Classifying spam: Logistic regression can be used to classify email as spam or not spam, based on the content of the email.
* Predicting risk: Logistic regression can be used to predict the risk of a patient developing a disease, based on their medical history and other factors.

# 

# 

# 

# 

# 

# Web scraping.

Web scraping is the process of extracting data from websites. It is a powerful tool that can be used to collect large amounts of data from the web, which can then be used for a variety of purposes, such as:

* Data analysis: Web scraping can be used to collect data for data analysis, such as market research or customer analytics.
* Competitor analysis: Web scraping can be used to collect data about your competitors, such as their product prices, product reviews, and marketing campaigns.
* Personal finance: Web scraping can be used to collect data about your finances, such as your credit card statements, bank account balances, and investment portfolio.
* Social media: Web scraping can be used to collect data about social media, such as user demographics, sentiment analysis, and brand mentions.
* Research: Web scraping can be used to collect data for research purposes, such as academic research or market research.

Web scraping can be done manually, but it is often done using automated tools. There are many different web scraping tools available, both free and paid.

It is important to note that web scraping is not always legal. Some websites have terms of service that prohibit web scraping. It is important to read the terms of service of any website before you scrape it.

Here are some of the benefits of web scraping:

* Efficiency: Web scraping can be used to collect large amounts of data quickly and easily. This can save you a lot of time and effort.
* Accuracy: Web scraping can be used to collect data that is more accurate than data that is collected manually. This is because web scraping tools can be programmed to follow specific rules and procedures when collecting data.
* Scalability: Web scraping can be scaled to collect data from a large number of websites. This makes it a powerful tool for collecting data for large-scale projects.

Here are some of the risks of web scraping:

* Illegality: As mentioned earlier, some websites have terms of service that prohibit web scraping. If you scrape a website that has such a term of service, you may be violating the law.
* Account suspension: Some websites will suspend your account if they detect that you are scraping their website. This can be a problem if you are using the website for legitimate purposes, such as customer support or research.
* Bot detection: Some websites use bot detection software to identify and block web scrapers. If your web scraper is detected, you may be blocked from the website.
* Data quality: The quality of the data that you collect using web scraping can vary depending on the website that you are scraping. Some websites have high-quality data, while others have low-quality data. It is important to carefully evaluate the data that you collect before you use it for any purpose.

# 

# 

# 

# 

# 

# 

# CONCLUSIÓN:

Data mining, linear regression, logical regression, and web scraping are all powerful tools that can be used to extract knowledge from data. They can be used to solve a wide variety of problems, including:

* Predicting customer churn: Data mining can be used to predict the probability of a customer churning, based on their purchase history, demographics, and other factors.
* Detecting fraud: Linear regression can be used to detect fraudulent transactions, based on the characteristics of the transactions and the customer.
* Classifying spam: Logical regression can be used to classify email as spam or not spam, based on the content of the email.
* Predicting risk: Web scraping can be used to collect data about social media, such as user demographics, sentiment analysis, and brand mentions.

These tools can be used to improve decision-making, increase profitability, enhance customer experience, and improve compliance. However, it is important to note that these tools are not perfect and they can sometimes produce inaccurate results. It is important to carefully evaluate the results of these tools before making any decisions.

Here is a summary of the conclusions about data mining, linear regression, logical regression, and web scraping:

* Data mining: Data mining is the process of extracting knowledge from data. It is a multidisciplinary field that combines elements of statistics, machine learning, database systems, and visualization. Data mining techniques are used to identify patterns, trends, and anomalies in data. This information can then be used to make data-driven decisions, solve business problems, and uncover hidden insights.
* Linear regression: Linear regression is a statistical technique that is used to model the relationship between a dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.
* Logical regression: Logical regression, also known as logistic regression, is a statistical model that is used to predict the probability of a binary outcome, such as yes or no, based on a set of independent variables. The independent variables can be continuous or categorical.
* Web scraping: Web scraping is the process of extracting data from websites. It is a powerful tool that can be used to collect large amounts of data from the web, which can then be used for a variety of purposes, such as data analysis, competitor analysis, personal finance, social media, and research.

# 

# 

# 

# 

# 

# 

# REFERENCIAS:

<https://datascientest.com/es/text-mining-o-mineria-de-textos-definicion-tecnicas-casos-de-uso>

<https://estrategiastrading.com/clustering-jerarquico/>

<https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/>

<https://www.ibm.com/mx-es/topics/data-mining>

<https://www.sas.com/es_mx/insights/analytics/data-mining.html>

<https://www.expertases.com/what-is-clustering-in-data-mining-with-example/>

<https://stats.stackexchange.com/questions/12404/what-exactly-does-it-mean-to-pool-data>

<https://economipedia.com/definiciones/tipos-de-series-temporales.html>

<https://developer.ibm.com/articles/ba-data-mining-techniques>

<https://dataminingweb.wordpress.com/conclusiones/>