# A STUDY ON DATA MINING

Dissertation submitted in partial fulfillment.
the requirements for the

BACHELOR'S DEGREE IN MATHEMATICS

Jointly By

Ann Mary Joy (200021033151)
Arya Shibu (200021033152)
Keerthana T R (200021033171)
Sruthi C K (20002106533182)
Vignesh .K (200021033183)

Under the supervision and guidance of
Dr. Seethu Varghese
Assistant Professor
Department of Mathematics

Bharata Mata College, Thrikkakara
2020-2023

# DECLARATION

We hereby declare that this project report entitled " A STUDY ON DATA MINING" is a bonafied record of work done by us under the supervision of Dr Seethu Varghese and the work has not previously formed the basis for the award of any academic qualification fellowship or another similar title of any other university or board.

Ann Mary Joy :
Arya Shibu :
Keerthana T R :
Sruthi C K :
Vignesh. K :

Place: Thrikkakara
Date:

# ACKNOWLEDGEMENT

We have immense pleasure in presenting this report on " A Study On Data mining". We would like to express our sincere thanks to Dr Seethu Varghese, Head of the Department of Mathematics at Bharata Mata College, Thrikkakara, for her valuable support and guidance that enabled us to make study in this topic and to prepare this dissertation.

We are also thankful to all other teachers in the Department for their cooperation to prepare this report.

We also extend our sincere thanks to the staff and all our friends for their help and support during this study.

Ann Mary Joy

Arya Shibu

Keerthana T R

Sruthi C K

Vignesh K

Place: Thrikkakara

Date: 24.04.2023

# CONTENTS

# ABSTRACT

Chapter 1 includes the preliminaries about data mining.
Chapter 2 discuss about data and its preprocessing.
Chapter 3 includes cluster analysis and its applications in the real world.
Chapter 4 discuss about association analysis in data mining to identify patterns or connections between variables in a data set.
Chapter 5 contains anomaly detection which helps in finding the abnormality given in the data and neglecting it.

# Introduction

Data Mining refers to the process of discovering hidden patterns, relationships, and insights from large datasets. It involves the use of statistical and machine learning techniques to analyse and interpret data. The goal of data mining is to extract meaningful information that can be used to make informed decisions and predictions.

In recent years, the amount of data generated has grown exponentially, making it increasingly challenging to manage and extract useful insights from it. Data mining provides a solution to this problem by allowing organizations to leverage their data to gain a competitive advantage. This has led to an increased demand for professionals who are skilled in data mining techniques.

In this project, we will explore the process of data mining in more detail which includes data, cluster analysis , association analysis and anomaly detection.

Data preparation involves collecting, cleaning, and transforming data so that it can be analysed effectively. This step is crucial as it can greatly impact the accuracy and effectiveness of the data mining process. Once the data is prepared, we can begin applying various techniques to uncover hidden patterns and relationships.

Cluster analysis is a technique used to group similar data points together based on their characteristics. This is useful for identifying patterns in data that may not be immediately apparent. For example, cluster analysis can be used to segment customers based on their behaviour and preferences, allowing businesses to tailor their marketing strategies accordingly.

Association analysis is another popular technique used in data mining, which involves discovering associations or relationships between different variables in a dataset. This can help identify patterns in customer behaviour, identify market trends, and improve product recommendations.

Finally, anomaly detection is a technique used to identify unusual or unexpected data points in a dataset. This is useful for identifying potential fraud or anomalies in financial transactions, identifying abnormal behavior in medical data, or detecting abnormal behavior in computer networks.

CHAPTER 1

# PRELIMINARIES

## 1.1 What is a data mining

Data mining is the process of searching through massive databases for patterns, correlations, and trends using statistical and machine learning approaches. The primary goal of data mining is to extract information that can be utilized to enhance judgment, recognize opportunities, and expedite business processes.

## 1.2 Why data mining

It can be viewed as an advancement of database management systems, transactional data, and information technology. Traditional database management systems are no longer able to handle the complexity and amount of data that organizations are producing at an increasing rate. This sparked the creation of data mining tools, which made it possible for businesses to find undiscovered patterns and connections in their data.

## 1.3 Processes involved in data mining

Data characterization, discrimination, association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis are just a few of the diverse tasks that fall under the umbrella of the continually expanding and dynamic area of data mining. The diversity of applications and the creation of new mining tasks need the development of numerous data mining approaches. Additionally, investigating data in multidimensional space may reveal intriguing relationships between various combinations of dimensions at various levels of abstraction. Data mining is an interdisciplinary endeavor that can be strengthened by incorporating fresh approaches from many fields to increase its flexibility and power.

## 1.4 Future of data mining

Terabytes and petabytes of data are produced daily from a variety of sources, including corporations, scientific and engineering practices, the medical and health industry, social media, and the World Wide Web. This is the era of data. This sparked the development of the field of data mining, which tries to automatically extract useful information from vast amounts of data and organize it into knowledge. In our transition from the data age to the impending information age, the field is vibrant, promising, and will continue to advance significantly.

# CHAPTER 2
# DATA

Data refers to any set of facts, figures, statistics, or other information that can be used for analysis, interpretation, or decision-making.

The datasets differ in many ways. Attributes used to describe data objects can be of different types – quantitative or qualitative – and data sets have special properties; for example, some data sets contain time series or objects with explicit interrelationships. The type of data determines what tools and techniques can be used to analyse the data.

Data is often not perfect. While most mining techniques can tolerate some level of imperfections in the data, focusing on understanding and improving the quality of the data usually improves the quality of the resulting analysis.

## 2.1 Attributes

Data objects are described by several attributes that capture properties of the object, such as the mass of the physical object or the time when the event occurred.

A useful way to determine the type of an attribute is to identify the properties of numbers that correspond to the underlying properties of the attribute. The following operations of numbers are normally used to describe attributes.

1. Distinctness = and ≠
2. Order <, ≤, >, and ≥
3. Addiction + and –
4. Multiplication x and /

Four types of attributes can be defined: nominal, ordinal, interval, and ratio. Each attribute type has all the properties and operations of the attribute types above it. Therefore, the properties or operations that are valid for nominal, ordinal, and interval attributes are also valid for relational attributes.

- Nominal attribute: -

Normal attribute values differ only in name. In other words, nominal values provide only enough information to distinguish one object from another. (=,≠)

e.g.: - eye colour, gender, etc.

Operations: - contingency correlation, $\chi^2$ test, mode, entropy

- Ordinal attribute: - ordinal attribute values provide enough information to arrange objects. (<, >)

E.g.: - grades, hardness of minerals, street numbers, etc.

Operations: - median, rank correlation, run tests, sign tests, percentiles

- Interval attribute: -

For interval traits, the difference in values is significant. In other words, there are units of measurement. (+, -)

e.g.: -calendar dates, temperature in Fahrenheit or Celsius

Operation: - standard deviation, mean, Pearson's correlation, t and F tests

- Ratio: -

For ratio variables, both difference and ratio are significant. (x, /)

e.g.: - mass, length, age, counts, electrical current

Operations: - harmonic mean, percent variation, geometric mean

Nominal and ordinal attributes are together referred to as categorical (quantitative) attributes. Interval and ratio attributes are together referred to as quantitative attributes.

## 2.2.Describing attributes by the number of values

### 2.2.1 Discrete: -

A discrete property has a finite or countably infinite set of values. Such attributes can be categorical, such as zip codes or ID numbers, or numerical, such as counts. Discrete properties are often represented using integer variables. Binary features are a special case of discrete features and take only two values such as true/false, yes/no, male/female, or 0/1. Binary attributes are often represented as Boolean variables or integer variables that only take the values 0 or 1.

### 2.2.2 Continuous: -

Continuous attributes are attributes whose values are real numbers. For example, we can refer to features such as temperature, height, and weight.

Normally, nominal and ordinal attributes are binary or discrete, and distance and ratio attributes are continuous. However, the distinct count attribute is also a ratio attribute.

## 2.3 DATA PREPROCESSING

Data pre-processing is a broad field and includes several different strategies and techniques that are intricately interrelated. We will present some of the most important ideas and approaches and try to show the interrelationships between them.

In particular, we will discuss the following points:

- Aggregation
- Sampling
- Dimensionality reduction
- Feature subset selection
- Discretization and binarization

### 2.3.1Aggregation: -

Aggregation is an important concept in data mining that involves combining or summarizing data from multiple sources or across multiple variables to generate new insights or patterns. Aggregation is commonly used in data mining to reduce the complexity of large data sets and to help analysts identify trends and patterns that may be difficult to discern at the individual data point level.

There are many motivations for consolidation. First, data reduction results in smaller data sets requiring less memory and processing time, and therefore aggregation often enables the use of more expensive data mining algorithms. Second, aggregation can act as a change of scope or scale by providing a high-level view of the data rather than a low-level view. Finally, the behaviour of groups of objects or attributes is often more stable than that of individual objects or attributes. This statement reflects the statistical fact that an aggregate quantity such as a mean or total has less variability than when individual values are aggregated. The actual amount of variation for the total is greater than for the individual items, but the percentage of variation is less than for the individual items. A disadvantage of aggregation is the potential loss of interesting details.

## 2.3.2. Sampling: -

There are different types of sampling techniques. One of the simplest sampling techniques is simple random sampling. This type of sampling has an equal probability of selecting a particular object. There are 2 types of random sampling sampling without replacement and sampling with replacement. The next sampling technique is stratified sampling. stratified sampling starts with prespecified groups of objects. The same number of objects are drawn for each group, even if the groups are of different sizes. In another variation, the number of objects drawn from each group is proportional to the size of that group.

## 2.3.3. Dimensionality Reduction: -

The main advantage of dimensionality reduction is that many data mining algorithms work better if the dimensionality—the number of features in the data—is low. This is partly because dimensionality reduction can remove irrelevant features and reduce noise. Another advantage is that dimensionality reduction can lead to a more understandable model because the model usually contains fewer features. Also, the dimensionality reduction can allow the data to be visualized more easily.

## 2.3.4. Feature Subset Selection: -

Feature subset selection involves finding a small feature set that can perform as well or better than the full feature set while reducing computational complexity and the risk of overfitting.

Common methods of subset selection in data mining are as follows:

Filter Method: - This method evaluates the relevance of each attribute independently of the model used. Statistical techniques such as correlation analysis, chi-square test, and mutual information can be used to measure the relationship between each characteristic and the target variable. A subset of features with the highest scores can then be selected for further analysis.

Wrapper Method: - This method uses a model to evaluate the utility of different subsets of performance. The wrapper method generates a subset of features, evaluates them using a specific model, and selects the subset that provides the best model performance. However, wrapper methods are computationally expensive and prone to overfitting.

Embedding Method: - This method involves the selection of features during the model building process. Some machine learning algorithms, such as decision trees and regular regression models, automatically select a subset of features during training. A selected subset of features is available for further analysis.

## 2.3.5.Discretization and Binarization: -

Discretization is the process of dividing continuous numerical data into discrete intervals or bins. This is done to simplify the data and facilitate data analysis. Discretization is done in two ways:

Equal width discretization: This method divides the data range into a fixed number of bins of equal width. For example, if you have a numeric feature with values from 0 to 100 and divide it into 5 bins, each bin has a width of 20 (0-20, 20-40, 40-60, 60-80, 80- 100).

Equal frequency discretization: This method divides the data into bins so that each bin has approximately the same number of data points. For example, if you have a numeric feature with 100 data points and divide it into 5 bins, each bin will contain 20 data points.

Binarization is the process of converting continuous numeric data into binary 0 or 1 data. Binarization can be done in different ways.

Thresholding: In this method, a threshold is selected, 1 is assigned to data points above the threshold and 0 is assigned to data points below the threshold. Sometimes we want to identify objects in an image based on brightness or colour.

Quantization: This method divides the data into a fixed number of levels and assigns a binary value to each data point based on the level it falls on. For example, if you have a numeric feature with values from 0 to 100 and divide it into four levels, data points 0 to 25 are assigned the value 00, and data points 26 to 50 are assigned the value 00. You can be assigned a value like 01.

## 2.4.Measures of similarity and Dissimilarity

There are several measures of similarity and dissimilarity that are commonly used in data mining, including:

Euclidean distance: This is the most used measure of distance in data mining. It measures the distance between two points in a multidimensional space. Euclidean distance is calculated as the square root of the sum of the squared differences between the corresponding attributes of the two objects.

d(x,y) = $\sqrt{\sum_{k=1}^{n}(x_k - y_k)}\ 2$

**Cosine similarity:** This measure is used to determine the similarity between two vectors. It measures the cosine of the angle between the two vectors. A cosine similarity of 1 indicates that the two vectors are identical, while a cosine similarity of 0 indicates that they are orthogonal (completely dissimilar).

Cos(x,y) = $\frac{(x,y)}{\|x\|\|y\|}$

**Jaccard similarity**: This measure is used to determine the similarity between two sets. It measures the ratio of the intersection of the two sets to the union of the two sets. A Jaccard similarity of 1 indicates that the two sets are identical, while a Jaccard similarity of 0 indicates that they have no elements in common.

**Hamming distance:** This measure is used to determine the distance between two binary strings of equal length. It measures the number of positions at which the corresponding symbols are different.

**Manhattan distance:** This measure is also known as city-block distance. It measures the distance between two points by summing the absolute differences between the corresponding attributes of the two objects.

These measures of similarity and dissimilarity are used in different contexts depending on the nature of the data and the problem being solved.

CHAPTER 3

# CLUSTER ANALYSIS

## 3.1.Introduction

Cluster analysis is a popular data analysis technique used to group similar data points together. The goal of the analysis is to identify different.

customer segments that exhibit similar purchasing behaviors.

## 3.2. Applications of Cluster Analysis

Cluster analysis is a powerful tool in data analysis and has several applications in various fields, including marketing, biology, finance, and social sciences. One of the most common applications of cluster analysis is in customer segmentation in marketing. By identifying groups of consumers with indistinguishable characteristics, companies can tailor their marketing strategies and product offerings to each segment, leading to increased customer satisfaction and profitability. Cluster analysis is also used in biology to classify organisms into different species based on shared characteristics, such as morphology, behavior, or DNA sequences.

In the financial sector, cluster analysis has various applications including customer segmentation for targeted marketing campaigns and portfolio diversification for risk management. By identifying similarities and differences between groups of customers or investments, cluster analysis can provide valuable insights for financial decision-making. In social sciences, cluster analysis is used to identify patterns of behavior or attitudes among individuals or groups, such as political ideologies, personality types, or consumer preferences. This can help researchers understand social phenomena and develop interventions to promote positive change. Overall, cluster analysis is a valuable tool for discovering patterns and insights in complex datasets, and its applications are diverse and widespread.

## 3.3. Requirements for Cluster Analysis

Before conducting a cluster analysis, there are several requirements that must be met to ensure accurate and meaningful results.

The first requirement is the presence of homogeneous data. The data must be collected from similar sources and should have similar properties, such as scale, variance, and distribution. If the data is not homogeneous, the clustering algorithm may produce inaccurate results or even fail to cluster the data at all. Therefore, it is important to preprocess the data before conducting cluster analysis to ensure that it meets this requirement.

The second requirement is the selection of appropriate distance or similarity measures.  Different clustering algorithms use different distance or similarity measures . Commonly used distance measures include Euclidean distance, Manhattan distance, and cosine distance.

The third requirement is to determine the number of clusters. The number of clusters in a dataset depends on various factors such as the nature of the data and the clustering algorithm used. Determining the optimal number of clusters is a challenging task in unsupervised learning. There are several methods available, including elbow method, silhouette analysis, and dendrogram, to help estimate the number of clusters. The fourth requirement is the selection of an appropriate clustering algorithm. There are several clustering algorithms, including hierarchical clustering, K-means clustering, and density-based clustering. The choice of the clustering algorithm depends on the type of data being clustered and the objectives of the analysis. Therefore, it is important to choose the appropriate clustering algorithm to ensure that the results are accurate and meaningful.

## 3.4. Basic clustering methods

There are several basic clustering methods, including:

1.K-Means Clustering: This is a popular clustering method that partitions the data into K clusters, where K is a predetermined number. K-means clustering works by randomly selecting K centroids, then assigning each data point to the nearest centroid, and recalculating the centroid based on the new group. The process continues until the centroids do not move any further, or a maximum number of iterations is reached.

2.Hierarchical Clustering: This clustering method creates a hierarchy of clusters by recursively dividing the data into smaller clusters based on similarity. There are two types of hierarchical clustering: agglomerative and divisive. In agglomerative clustering, each data point starts as a cluster and the algorithm merges the closest clusters until there is only one cluster. In divisive clustering, all data points start in one cluster and the algorithm recursively divides the data into smaller clusters until each data point is in its own cluster.

3.Density-Based Clustering: This method groups together data points that are close together in high-density areas and separates data points that are in low-density areas. Density-based clustering algorithms, such as DBSCAN, define clusters as dense regions of data points that are separated by low-density regions.

4.Fuzzy Clustering: This is a soft clustering method that assigns each data point a membership value between 0 and 1, indicating the degree of membership in each cluster. Unlike hard clustering, where each data point belongs to only one cluster, fuzzy clustering allows data points to belong to multiple clusters at the same time.

In summary, the basic clustering methods are K-means clustering, hierarchical clustering, density-based clustering, and fuzzy clustering. Each method has its strengths and weaknesses, and the choice of clustering method depends on the nature of the data and the objectives of the analysis.

## 3.5. Evaluation of Clustering

Clustering is a fundamental task in unsupervised machine learning that involves grouping similar data points together into clusters. The performance of clustering algorithms is typically evaluated using a variety of metrics, including both internal and external measures.

Internal measures evaluate the quality of clustering based on the data itself, without any reference to external information. One commonly used internal measure is the silhouette coefficient, which measures how well each data point fits into its assigned cluster relative to other clusters. The silhouette coefficient ranges from -1 to 1, with higher values indicating better clustering.

External measures, on the other hand, compare the clustering results to some external ground truth or labels. One commonly used external measure is the adjusted Rand index (ARI), which measures the similarity between the true labels and the predicted clusters. The ARI ranges from -1 to 1, with higher values indicating better clustering.

Another commonly used external measure is the F-measure, which is the harmonic mean of precision and recall. Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives. The F-measure ranges from 0 to 1, with higher values indicating better clustering.

In addition to these measures, there are also visual evaluation techniques that can be used to assess clustering quality. For example, scatter plots can be used to visualize the data and the resulting clusters, allowing for visual inspection of how well the clusters separate the data.

However, it is important to note that the evaluation of clustering is not always straightforward, as different clustering algorithms may perform better or worse depending on the specific data set and the desired outcome. Moreover, the choice of evaluation metric may also depend on the particular application and the goals of the analysis.

Overall, clustering is an important tool for unsupervised machine learning, and a variety of metrics and evaluation techniques can be used to assess the quality of clustering results. However, it is important to carefully consider the specific goals and requirements of each application when choosing a clustering algorithm and evaluation.

## 3.6. Summary

Clustering is a machine learning technique used for grouping data points into clusters based on their similarities. It is an unsupervised learning method that doesn't require any prior knowledge or labels.

There are several clustering algorithms such as K-means, Hierarchical clustering, DBSCAN, and Gaussian Mixture Models.

K-means clustering is a popular algorithm that divides the data into K clusters by minimizing the sum of squared distances between data points and their respective cluster centroids.

Hierarchical clustering is a bottom-up approach that creates a tree-like structure of clusters by merging or splitting clusters based on their distance or similarity.

DBSCAN is a density-based algorithm that groups together data points that are in dense regions and separates them from sparse regions.

Gaussian Mixture Models (GMM) is a probabilistic clustering algorithm that models each cluster as a Gaussian distribution and assigns each data point to the cluster with the highest probability.

Clustering has many applications, including image segmentation, customer segmentation, anomaly detection, and document classification.

Evaluation of clustering algorithms can be done using metrics such as silhouette score, Calinski-Harabasz index, and Davies-Bouldin index. The choice of the appropriate evaluation metric depends on the data and the objectives of the clustering task.

Overall, clustering is a useful technique for identifying patterns and relationships in data and can be a valuable tool in various fields, including data science, machine learning, and artificial intelligence.

CHAPTER : 4

# ASSOCIATION ANALYSIS

Association analysis is a method used in data mining to identify patterns or connections between variables in a dataset. It's generally used in request market basket analysis, which involves chancing connections between products that are constantly bought together. Association analysis can also be used in other disciplines, similar as healthcare, where it may be used to identify connections between different medical conditions or treatments. The introductory idea behind association analysis is to look for patterns in a dataset that indicate that certain variables tend to do together more constantly than would be anticipated by chance. One popular algorithm for association analysis is the Apriori algorithm, which works by gradationally erecting up sets of variables that do together more constantly than a specified threshold. Once these patterns have been linked, they can be used to make prognostications or recommendations. For illustration, in request handbasket analysis, the results of association analysis might be used to suggest products that a client is likely to buy grounded on their once buying history.

## Section 4.1. Support and confidence

Support and confidence are two important measures used in data mining to dissect and interpret patterns and connections within datasets. It is generally used in association rule mining, which involves chancing connections between particulars or attributes in a dataset. High support and high confidence values indicate strong connections between particulars, while low support and low confidence values indicate weak connections or no association at all.

Support refers to the frequency of circumstance of an itemset or rule in a given dataset. It's defined as the proportion of deals that contain the itemset or satisfy the rule.

**Support s (x →y) =σ(x∪y)/N** x and y disjoint item set

N: Total no of transition

Confidence measures the strength of association between two particulars in an itemset or rule. It’s defined as the tentative probability of the consequent item given the precedent item.

**Confidence c(x→y) =σ(x∪y)/σ(x)** x and y disjoint item set

## Section 4.2. Apriori Algorithm

Apriori is a popular algorithm used in data mining for association rule literacy. It’s used to find frequent itemset in a given dataset and induce association rules from them. The algorithm works by using a bottom- up approach, starting with individual particulars and gradationally erecting up to larger itemset, grounded on the frequency of circumstance of each item. The frequency of an item is measured by counting the number of deals in which it appears. Apriori uses two crucial measures to identify frequent itemset support and confidence. Support is the proportion of deals that contain a particular itemset, while confidence is the proportion of deals containing an itemset that also contain a particular item. Once frequent itemset have been linked, Apriori generates association rules from them.

## Section 4.3. Association Rule

Association rules are a important method used in data mining to discover patterns or connections between particulars in a dataset. They’re generally used in request handbasket analysis to identify sets of products that are constantly bought together by guests. The introductory idea behind association rules is to identify frequent itemset, which are sets of particulars that constantly do together in the dataset. Once these itemset are linked, the association rule mining algorithm generates rules that describe the connections between these particulars.

## Section 4.4. Market basket Analysis

Market basket analysis is a  method in data mining that analyzes the connections between particulars that are constantly bought together by guests. It's generally used by retailers and marketers to identify patterns in client buying geste, and to induce perceptivity that can be used to ameliorate product immolations, elevations, and pricing strategies. The analysis involves looking at large quantities of transactional data, generally gathered through point- of- trade systems, and using algorithms to identify itemset- groups of particulars that are constantly bought together. This can be done through colorful styles, similar as association rule mining or clustering. The affair of request market basket analysis is a set of rules or recommendations, which can be used to inform product placement, cross-selling, and upselling strategies. For illustration, if a retailer discovers that guests who buy diapers also tend to buy baby wipes and formula, they might consider grouping those particulars together in- store or running targeted elevations on those particulars.

## Section 4.5. Frequent Item set

Frequent item sets are an abecedarian conception in data mining, used to identify patterns and connections in large datasets. An itemset is simply a collection of particulars, and a frequent itemset is a set of particulars that appears constantly in a dataset. The process of changing frequent itemset is called association rule mining or request market basket analysis. This fashion involves assaying large datasets to identify patterns of-occurrence between particulars. For illustration, a supermarket might use association rule mining to determine that guests who buy bread Are likely to also buy milk. To identify frequent itemset, data mining algorithms use a measure called support. Support measures the proportion of deals in the dataset that contain a particular itemset. An itemset with a high support value indicates that it appears constantly in the dataset. Once frequent item sets have been linked, they can be used to induce association rules. Association rules describe the connections between particulars in a dataset, similar as ” if a client buys bread and milk, they’re likely to also buy eggs.” Frequent itemset and association rules are generally used in operations similar to request market basket analysis, product recommendation systems, and fraud discovery. They give precious perceptivity into the connections between particulars in large datasets, helping associations to make data- driven opinions.

## Section 4.6. Collaborative filtering

Collaborative filtering is a popular method used in data mining and machine literacy to make individualized recommendations. The introductory idea behind cooperative filtering is to use the geste and preferences of a group of users to make recommendations for an individual user. In collaborative filtering, the system gathers data on ser preferences, generally through conditions or reviews of particulars like pictures, books, or products. also, using algorithms and statistical models, the system identifies patterns in the data, looking for arallels between users and particulars. The system can lso use these patterns to make recommendations for a particular stoner grounded on the preferences of other users with nalogous tastes. For llustration, if stoner A has rated several ctures argely, and stoner B has alogous preferences, also the system might recommend pictures that user B has not yet seen, but that user A has enjoyed.

There are several types of Collaborative filtering

- **User-based collaborative filtering:** This approach recommends particulars to a stoner grounded on the conditions of nalogous uggies.
- **Item-based collaborative filtering:** This approach recommends particulars that are analogous to articulars that a stoner has liked in history.
- **Model-based collaborative filtering:** This approach uses machine literacy algorithms to identify patterns in toner ste and make recommendations grounded on those patterns.

cooperative filtering has become popular in recent times, particularly e-commerce, social media, and entertainment  diligence, as it allows businesses to give  individualized recommendations to their  guests, thereby  perfecting  users satisfaction.

## Section 4.7. Sequential pattern Mining

Sequential pattern mining is a fashion used in data mining and machine literacy to discover patterns or connections in successional data. This type of data consists of sequences of events, conduct, or particulars that do in a specific order. The thing of  Sequential pattern mining is to identify constant patterns in the data, similar as a sequence of events that frequently leads to a particular outgrowth. For illustration, in retail, Sequential pattern mining can help identify which products are frequently bought together or which products are constantly bought by guests who also buy a specific item.

Sequential pattern mining involves several ways,

- **Data preprocessing :** This step involves cleaning and preparing the data for analysis, similar as removing duplicates, converting the data into a suitable format, and relating the sequence of events.
- **Pattern representation** : This step involves representing the data in a format that can be anatomized, similar as a sequence of deals or a sequence of events.
- **Pattern mining:**  This step involves applying algorithms and statistical ways to identify constant patterns in the data.
- **Pattern evaluation:**  This step involves assessing the linked patterns grounded on certain criteria, similar as support and confidence situations, to determine their significance and utility.

There are several operations of Sequential pattern mining, including client geste Analysis, fraud discovery, and medical opinion. For illustration, successional pattern mining can help identify patterns in medical data that may be reflective of a specific complaint or condition, allowing for early discovery and treatment.

## Section 4.8. Constraint based association mining.

Constraint- based association mining is a method used in data mining to discover intriguing patterns or connections among variables in large datasets. In this approach, constraints are used to guide the mining process and concentrate on specific aspects of the data. The introductory idea of constraint- based association mining is to define certain conditions or rules that must be satisfied by the data in order for a pattern to be considered intriguing. These constraints can be specified in colorful ways, similar as by setting minimal support or confidence thresholds, or by specifying rejection rules that help certain particulars from appearing together in a frequent itemset. One popular algorithm for constraint- based association mining is the Apriori algorithm. This algorithm works by constantly surveying the dataset to identify frequent itemset( sets of particulars that appear together constantly), and also using these itemset to induce seeker rules that meet the specified constraints. Once intriguing patterns have been discovered, they can be used for a variety of purposes, similar as request handbasket analysis, targeted advertising, and client segmentation. Still, it’s important to flash back that correlation doesn’t always indicate occasion, so the perceptivity gained from association mining should always be interpreted with caution and validated through farther analysis.

## Section 4.9. Text Mining

Text mining is a subset of data mining that involves assaying and rooting useful information from unshaped textbook data. It’s also known as textbook analytics or natural language processing( NLP). Text mining uses colorful ways and algorithms to identify patterns and connections in textbook data, similar as sentiment analysis, content modeling, and information extraction .

The process of textbook mining generally involves the following way.

**Data collection :** Gathering large quantities of unshaped textbook data from colorful sources similar as social media, newspapers, client reviews, and other sources.

- **Data preprocessing:** Cleaning and transubstantiating the raw textbook data into a format that can be anatomized, similar as removing stop words, stemming, and tokenization.
- **Text analysis:** Applying colorful algorithms and ways to identify patterns and connections in the textbook data, similar as sentiment analysis, content modeling, and information birth.
- **Evaluation and interpretation:** Assessing the results of the textbook analysis and drawing conclusions grounded on the perceptivity gained.

Text mining has a wide range of operations in colorful fields, similar as marketing, client service, healthcare, and social media analysis. Some exemplifications of how textbook mining can be used include assaying client feedback to ameliorate products and services, relating trends and patterns in social media exchanges, and rooting crucial information from scientific literature.

# ANOMALY DETECTION

Anomaly detection is a technique in data mining that involves identifying patterns or data points that deviate from the expected behavior of a dataset. Anomaly detection can be applied to a wide range of domains, including finance, cybersecurity, healthcare, and industrial quality control. The goal of anomaly detection is to find data points that are significantly different from most of the data. Anomalous objects are frequently known as outliers. These outliers may indicate errors or abnormalities in the data, or they may represent important or interesting phenomena that are worth further investigation. There are various approaches to anomaly detection, including statistical methods, machine learning algorithms, and rule-based systems.

Anomaly detection has numerous applications, such as detecting fraudulent transactions in finance, identifying abnormal behavior in industrial equipment, detecting anomalies in medical imaging, and monitoring network traffic for cybersecurity threats. It can help organizations improve data quality, prevent costly errors, and enhance decision-making by identifying important patterns and outliers in their data.

- In finance, anomaly detection can be used to identify fraudulent transactions, such as unauthorized charges, money laundering, or account takeovers. Anomaly detection algorithms can learn the normal spending patterns of a user or a group of users and flag transactions that fall outside of this range as potentially fraudulent.
- In healthcare, anomaly detection can be used to identify fraudulent medical claims, such as overbilling, phantom billing, or billing for unnecessary services. Anomaly detection algorithms can learn the normal billing patterns of healthcare providers and flag claims that deviate from this pattern as potentially fraudulent.

- In insurance, anomaly detection can be used to identify fraudulent insurance claims, such as false claims, exaggerated claims, or claims for pre-existing conditions. Anomaly detection algorithms can learn the normal claim patterns of policyholders and flag claims that deviate from this pattern as potentially fraudulent.

## 5.1 Characteristics of Anomaly detection problems

### 5.1.1 Definition of an Anomaly

The important characteristics of an anomaly are its unusualness, rarity, abnormality, and potentially high importance. Anomalies are significantly different from normal behavior or patterns, occur at a much lower frequency than normal data points. Detecting anomalies is important as they may indicate errors, outliers, or interesting phenomena that require further investigation.

### Definition 5.1.

Anomaly is an observation that doesn't fit the distribution of the data for normal instances, i.e., is unlikely under the distribution of most instances.

### 5.1.2 Nature of data

The nature of data in anomaly detection plays a crucial role in determining the appropriate anomaly detection technique and preprocessing steps required to achieve accurate results.
Univariate or multivariate data: Univariate data refers to a single variable or feature, while multivariate data refers to multiple variables or features that describe the behavior or properties of a system. Univariate anomaly detection techniques can be used to detect anomalies in a single variable, while multivariate techniques can detect anomalies that involve interactions between multiple variables.

Record data or proximity data: anomaly detection can be performed on record data or proximity matrix. Record data refers to data that describes individual instances or records, while proximity matrix refers to data that describes the distance or similarity between instances. Record data can be used for anomaly detection using techniques such density-based approaches, while proximity matrix can be used with clustering-based methods or distance-based approaches .

Availability of labels: Anomaly detection can be performed using labeled or unlabeled data. Labeled data refers to data that has been manually labeled as normal or anomalous, while unlabeled data does not have any labels. Labeled data can be used with supervised anomaly detection techniques such as decision trees, while unsupervised techniques such as clustering or density-based approaches can be used with unlabeled data.

## 5.2 Statistical approaches

Statistical approaches are a common class of techniques used in anomaly detection. These methods use statistical models to identify instances that deviate significantly from the expected behavior. The underlying assumption is that anomalous instances are rare and can be detected based on their statistical properties.

Statistical approaches can be divided into two categories: parametric and non-parametric methods. Parametric methods assume that the data is generated from a known statistical distribution, such as a normal or Gaussian distribution, and use the parameters of the distribution to detect anomalies. Non-parametric methods, on the other hand, do not make any assumptions about the underlying distribution of the data and use non-parametric techniques such as kernel density estimation or nearest neighbor methods to detect anomalies.

### 5.2.1 Using parametric models.

The Gaussian distribution, the Poisson distribution, and the binomial distribution are a few of the typical parametric models that are frequently used to describe many sorts of data sets. They involve parameters that must be identified from the data, such as the mean and variance parameters for a Gaussian distribution.

#### Using the Univariate Gaussian Distribution

The univariate Gaussian distribution, also known as the normal distribution, is commonly used in anomaly detection to model univariate data. The Gaussian distribution is a probability distribution that is characterized by its mean and standard deviation.

The probability density function (PDF) of the Gaussian distribution is given by the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where, x is the value of the variable, μ is the mean of the distribution, and σ is the standard deviation of the distribution.

The PDF of the Gaussian distribution is a bell-shaped curve that is symmetric around the mean. The mean represents the centre of the curve, while the standard deviation represents the spread or width of the curve. The curve is at its highest at the mean and decreases as we move away from the mean in either direction.

In anomaly detection using the univariate Gaussian distribution, the z-score method is commonly used. The z-score measures the number of standard deviations that an instance deviates from the mean of the distribution. The z-score of an instance x is given by the equation: z-score = (x - μ) / σ

Instances with a z-score above a certain threshold are considered anomalous. The threshold can be determined based on the desired level of sensitivity and the trade-off between false positives and false negatives.

### 5.2.2 Using Non-parametric models.

Histogram-based anomaly detection is a simple non-parametric approach for detecting anomalies in univariate data. The basic idea is to divide the range of the data into a fixed number of intervals or "bins" and count the number of instances that fall within each bin. The resulting histogram provides a visual representation of the data distribution and can be used to identify instances that fall outside the expected range or distribution of the data.

To detect anomalies using a histogram, one common approach is to use the interquartile range (IQR) to identify outliers. The IQR is calculated as the difference between the 75th and 25th percentiles of the data, and any instances that fall outside of 1.5 times the IQR above or below the 75th and 25th percentiles, respectively, are considered anomalous.

### 5.2.3 Strengths and Weaknesses

Strengths of statistical approaches for anomaly detection include their ability to model data distributions and identify instances that deviate significantly from expected patterns. They can be particularly effective for detecting anomalies in univariate data or when the data follows a known distribution.

However, statistical approaches also have some weaknesses. They may not perform well when the data distribution is complex and can be sensitive to outliers or data pre-processing techniques. Parametric methods can also be limited by assumptions made about the underlying data distribution, and if these assumptions are violated, they may not accurately detect anomalies.

## 5.3 Clustering-based approaches

Cluster-based approaches in anomaly detection involve identifying anomalies as data points that do not belong to any cluster or group of similar data points. These methods typically use clustering algorithms, such as k-means or DBSCAN, to group similar data points together based on some distance metric or similarity measure. Anomalies are then identified as data points that do not fit into any of these clusters or have low membership scores in all clusters. Cluster-based approaches can be effective in identifying global anomalies that deviate significantly from the expected patterns of behavior in the data but may struggle with detecting local anomalies or anomalies that are similar to some but not all data points.

### 5.3.1 Finding Anomalous clusters

Finding anomalous clusters in data involves identifying clusters that are significantly different from other clusters in the dataset. This can be achieved by analyzing the characteristics of each cluster, such as the number of data points it contains, its density, its shape, and its distance from other clusters. Anomalous clusters may have one or more of these characteristics that deviate significantly from the expected patterns of behavior in the data. Once an anomalous cluster is identified, further analysis can be done to identify the specific data points that are contributing to the anomalous behavior. This approach can be useful in identifying anomalous behavior that is not readily apparent from analyzing individual data points alone. However, it can also be more computationally intensive and may require more sophisticated algorithms to identify anomalous clusters accurately.

### 5.3.2 Finding Anomalous instances

Finding anomalous instances in data involves identifying data points that deviate significantly from the expected patterns of behavior in the data. This can be achieved using various techniques, such as statistical analysis, machine learning algorithms, and clustering. Anomalous instances may be those that have unusual values for one or more features, or those that are located far away from most of the data points in the dataset. Once an anomalous instance is identified, further analysis can be done to understand why it is anomalous and to determine whether it is a legitimate outlier or a data error. This approach can be useful in a wide range of applications, including fraud detection, intrusion detection, and medical diagnosis. However, it can also be challenging to identify anomalous instances accurately, particularly in large and complex datasets, and may require domain expertise to interpret the results effectively.

### 5.3.3 Strength and weaknesses

Clustering-based approaches have several strengths in anomaly detection, such as their ability to handle large datasets and identify complex anomalies that may not be detectable using other techniques. They can also be used to group similar data points and identify patterns within the data.

However, clustering-based approaches also have some weaknesses, such as their sensitivity to the choice of clustering algorithm and the selection of clustering parameters. Additionally, the interpretation of the results may be challenging without domain knowledge, and the approach may require significant computational resources.

## 5.4 Information Theoretic approaches

Information-theoretic approaches in anomaly detection work by using a compact representation of the data that captures the relevant information content. This can be achieved using techniques such as data compression or feature extraction. By compressing or reducing the data to a smaller, more compact representation, the approach can then measure the deviation from expected levels of entropy or information content in this representation to identify anomalous patterns.

For example, one approach is to use the Minimum Description Length (MDL) principle, which seeks to find the most compact representation of the data that preserves the maximum amount of information. This approach works by comparing the compression length of the data under different models, and selecting the model that yields the shortest description length as the best representation of the data. Anomalies can then be detected by measuring the deviation from expected levels of entropy or information content in this compact representation.

Another information-theoretic approach is based on the concept of Kolmogorov complexity, which measures the shortest possible algorithmic description of a data sequence. Anomalies can be identified by measuring the difference in the Kolmogorov complexity between the observed data and the expected model.

### 5.4.1 Strength and weaknesses

Information-theoretic approaches have several strengths, such as their ability to handle high-dimensional data, their ability to capture complex patterns and relationships, and their unsupervised nature. They also provide a compact representation of the data, which can be useful for visualizing and interpreting the results.

However, there are also some weaknesses of information-theoretic approaches. One of the main limitations is they can be computationally expensive for large datasets and may not perform well with noisy or incomplete data. Finally, they typically do not provide explicit information about the nature or cause of the anomalies, which may limit their usefulness in certain applications.

## 5.5 Evaluation of Anomaly detection

Evaluation of anomaly detection is an important aspect of the anomaly detection process. It involves measuring the performance of an anomaly detection algorithm by assessing its ability to correctly identify anomalies in the data. Evaluation metrics such as precision, recall, accuracy, false positive rate, and false negative rate are used to evaluate the effectiveness of an anomaly detection algorithm. The choice of evaluation metrics depends on the specific problem and the type of data being analyzed.

In anomaly detection, false alarm rate (FAR) refers to the ratio of falsely detected anomalies to the total number of non-anomalous instances. It is also known as the false positive rate (FPR). A high false alarm rate means that the system is detecting a large number of false positives, which can be problematic in real-world scenarios.

# CONCLUSION

Data mining is a powerful tool for extracting useful insights and knowledge from large data sets. In this project , we have explored techniques of data mining such as data preprocessing, cluster, association analysis and anomaly detection. Data mining has broad applications in various domains such as healthcare, finance, marketing and many others. Moreover data mining skills are in high demand in the job market.

# REFERENCE

1. Introduction to data mining, second edition by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar
2. Data mining concepts and techniques by Han Jiawei, Kamber, Micheleine, Pei, Jian.