Data Mining Methods: A Review

Dimitrios Papakyriakou
PhD Candidate
Department of Electronic Engineering
Hellenic Mediterranean University
Crete, Greece

Ioannis S. Barbounakis
Assistant Professor
Department of Electronic Engineering
Hellenic Mediterranean University
Crete, Greece

ABSTRACT
The Big Data revolution is taking place due to the evolution of technology, where the technology enables firms to gather extremely huge amount of data, disseminating knowledge to their customers, partners, competitors in the marketplace [1]. The deeper we dive into technology, the more we compound the physical with the virtual world having in mind for instance the IoT (Internet of Things) as a network of physical devices connected together and able to exchange data.

There are many Big Data platforms a company can choose like Hadoop and Apache Spark to analyze large sets of data.Moreover, many data mining techniques like Classification, Clustering Analysis, Correlation Analysis, Decision Tree Induction, Regression Analysis can be used to identify patterns for knowledge discovery. In this paper, there is an extent review and summary of Big Data Mining techniqueswith the most common data mining algorithms suitable to be used to handle large datasets. The review depicts the general pros and cons of these algorithms and the correspondingappropriate fields that apply, and in general acts as a guideline to data mining researchers to have an outlook on what algorithms to choose based on their needs and based on the given datasets.

Keywords

Big Data, Big Data Analytics, Data Mining Algorithms, Data Clustering

1. INTRODUCTION
When we refer to Big Data, we mean the combination of structured, semi-structured, and unstructured data collected by Organizations and used in various projectsin combination with predictivemodeling tools and advanced Big Data analytics applications. The classifications of data referred above are very important to understand due to the rapid increase of semi-structured and unstructured data nowadays on the one hand, and the advanced development of tools that make managing and analyzing these classes of data on the other hand.

Structured data. – Structured data can be created by machines and humans having a pre-defined (fixed) data model, format, structure where a database designer can create in a way that entities can be grouped together to form relations. This makes structured data easy to store, analyze and search.A relational database is a representative example of structured data where tables are linked together using unique IDs and query language to interact with the data. Today the estimated amount of structured data accounts for less than 20 percent of all data whereas a much bigger percentage of all the data is unstructured data in our world.

Unstructured data. –The unstructured data has no inherent structure, cannot be contained in a row-column database and doe does not have an associated data model. The unstructured data is usually stored as different types of files for instance text documents, PDFs, photos, videos, audio files, social media content, satellite imagery, websites, and call center transcripts/recordings. Compared to structure datawere stored in spreadsheets or relational databases the unstructured data is usually stored in NoSQL databases, applications, and data warehouses. Plethora of information in unstructured data can be automatically processed with artificial intelligence algorithms today.

Semi-structured data. –The semi-structured data basically is a mix between structure and unstructured data, has some defining or consistent characteristics with some structure but does not conform to a data model. The semi-structured data lacks a fixed or rigid schema, cannot be stored in a form of rows and columns in Databases but contain tags and elements in the form of Metadata which is used to group data and describe how the data is stored. Examples of semi-structured Data sources are the E-mails, XML and other markup languages, binary executables, TCP/IP packets, Zipped files, and Web pages.

2. BIG DATA DIMENSIONS
The concept of Big Data gained momentum in the early 2000s where Gartner analyst Doug Laney articulated the definition of Big Data analyzing the Volume, Velocity and Variety dimensions the so called three (Vs) [2]. According to that, there are three significant dimensions of the Big data“Figure 1”.

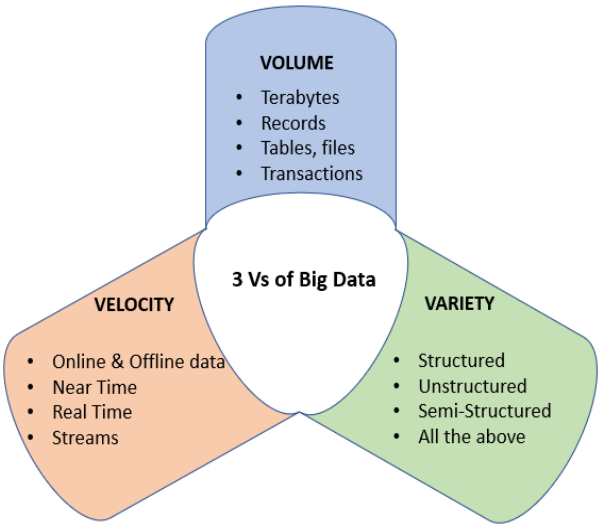

Figure1: The 3 (Vs) of Big Data

Nowadays, we all know that Big Data has penetrated in every industry accepting that is a prevailing driving force for every Organization to succeed across the globe. The “Big Data” as a terminology refers to huge and complex data that it is difficult

to process them by using traditional methods compared to old fashioned data.It is fine whenever business is dealing with data using excel sheets and databases, however when the data cannot be fitted with such tools, then we think about Big Data and Analytics.

Volume. – When we refer to volume, we mean the size of huge amount of data setslying between terabytes and zettabytes, from variety of Sources. Sources such as business transactions, smart Internet of Things (IoT) sensor devices, social media, and other e-commerce platforms where get real-time, structured, and unstructured data.It is estimated that 2.5 quintillion bytes of data is created each day. According to McAffe and Brynjolfsson more data crosses the internet every second than the total amount of data stored online 20 years ago [3]. The Volume of data created, captured, copied, and consumed worldwide is forecast to increase rapidly reaching 59 zettabytes in 2020 and 149 zettabytes in 2024 [4].

Velocity. – Broadly speaking Velocity refers to the speed of generating, processing, and analyzing the data. Nowadays, it is crucial for the Organization to have the information quickly as close to real-time as possible in the sense of paying much more importance to Velocity than to volume giving to Organizations bigger comparative advantage [5], [6], [7]. The appropriate business decisions are strongly dependent to the data availability at the right time since after a couple of hours there may be useless under certain circumstances. For instance, in a machine learning service running in a social media platform with billions of users who post and upload messages or photos and videos, there is a continues transactions of petabytes of data that is being transferred from millions of devices. As we can understand the rate of the volume data that inflows per second is very high defining the velocity of the data. A representative example of data generated with such a high velocity will be Twitter messages and Facebooks posts. Another example of velocity is the sensor data with the Internet of Things (IoT) evolution where the connected sensors are taking off at a dramatic rate with data being transmitted at a near constant rate. Another example of velocity is the packet analysis for cybersecurity, where unfortunately threatening payloads can be hidden in a data flow passing through the firewall. Those data must be investigated and analyzed for patterns of suspect behavior and the situation is getting harder as more data is protected using encryption and the malware payloads are inside the encrypted packets.

Variety. – Variety refers to different data types of formats, namely, the diversity of data types and data sources, from structured numeric data stored in traditional databases to unstructured data types such as text documents, PDFs, photos, videos, audio files, social media content, XML and so on. These kind of heterogeneous data sets possess a big challenge for big data analytics and requires distinct processing capabilities and specialist algorithms [5], [8]. A typical example of high variety of data sets would be the Closed-circuit television (CCTV) audio and video generated in a surveillance area in a city. More than 80 % of the data in the world today is unstructured and at first look does not show any clue of relationships.Moreover, when it comes to Big Data, two additional dimensions are under consideration, the Veracity, and the Value, "Figure 2".

Figure2: The 5 (Vs) of Big Data

Veracity. – Veracity refers to the quality, the accuracy, and the reliability of the collected data since data comes from so many different sources. The first side of the Veracity in Big data it is not just the quality itself but how trustworthy are the data type, the data source considering abnormalities, inconsistencies, duplication as well [5], [6], [9]. The second side of data veracity involves the processing method of the data and the adequate output to objectives based on business needs.

Value. – Value refers to an organization's ability to transform those huge amounts of data into real business since accurate data enables businesses as a steppingstone to get closer to their customer needs and expectations. Namely, Value denotes the *added value* for companies where huge amounts of data (Volume) from highly diverse sources (Variety) with different quality (Validity) are used to quickly make vital business decisions to gain comparative advantage [7], [9].

3. DATA MINING METHODS

When we refer to data mining, we mean the process of finding potentially useful patterns by using huge data sets. During this process, Machine Learning, Statistics, and Artificial Intelligent (AI) is used to extract information about the probability of future events.The diversified aspects of data mining comprise data classification, data integration, data transformation, data discretization, and pattern evaluation and more.Data mining techniques are used to discover hidden and unsuspected relationships amongst the data and used for marketing, sales, fraud detection, scientific discoveries,product development, healthcare, and education. Moreover, data mining techniques are used by the Organizations to solve business problems such as increasing revenues, acquiring new customers, improving cross-selling and up-selling, increasing Return of Investment (ROI) from marketing campaigns. As a result, the Organizations deliver consistent results that keep businesses ahead of the competition.

3.1 Association Rule Learning

In data science, the association rules technique is used to discover correlations between seemingly independent relational and transactional databases and datasets,and to observe frequently occurring patterns. The constraints on

various measures of significance and interest are used, so that to select the suitable rules among the set of all possible rules.An association rule has always two parts, the antecedent (if) and a consequent (then) where an antecedent is something that is found in data, and a consequent is an item that is found in combination with the antecedent. The two primary patterns that association rules use is support andconfidence which are user defined measures of interestingness [10].

Support. –Support is the measure of how frequent an itemset appears in the dataset where for a given rule, *itemset* is the list of all the items in the antecedent and the consequent.

$$Support\ (\{X\} \rightarrow \{Y\}) = \frac{Transactions\ \ containing\ \ both\ X\ and\ Y}{Total\ \ number\ \ of\ transactions} = \frac{frq\,(X,Y)}{N} \qquad (1)$$

In other words, support denotes the frequency of the rule within transactions. A high value means that the rule involves a great value of database.

Confidence. –Confidence is the measure of the likeliness of occurrence of consequence on the cart given that the cart already has the antecedents.

$$Confidence\ (\{X\} \rightarrow \{Y\}) = \frac{Transactions\ \ containing\ \ both\ X\ and\ Y}{Transactions\ \ containing\ \ X} = \frac{frq\,(X,Y)}{frq\,(X)} \qquad (2)$$

The most common Algorithms used in Association Rule techniques are the "*Apriori Algorithm*", the "*Eclat Algorithm*" and the "*Frequent Pattern (FP) Growth Algorithm*". The association rule mining is suitable for non-numeric, categorical data, in the Market Basket Analysis, Medical diagnosis, Protein sequences, Census data analysis, and Customer Relationship Management (CRM) of credit card business.

3.2 Classification

Classification is named the problem of predicting a discrete random variable Y from another random variable X and sometimes is called discrimination, or pattern classification or pattern recognition. Classification is a method which categorizes data into a definite number of classes and in turn label areassigned to each class. The main idea of the Classification algorithms is to predict the target class by analyzing the training dataset namely, to categorize the data into a given number of classes. We use the training set of data to get better boundary conditions and to assign the new data to preset categories or classes[11], [12]. Classification techniques are used to predict the group membership or class - therefornamed classification techniques) of individuals (data), for predefined group membershipsand also to describe which characteristics of individuals can predict their group membership.

The types of Classification Algorithms can be broadly classified as following:

3.2.1 Linear Classifiers:

Linear classifiers use classification on a linear function of inputs, that is to say, linear models for classification separate input vectors into classes using linear decision boundaries.

Logistic Regression. –Logistic regression is a classification method that models the probability of an observation to one of two classes and like all regression analysis, the logistic regression is a predictive analysis. Logistic regression is used to describe data explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. The heart of the matter of the logistic regression analysis is the task to estimate the log odds of an event.

A statistical model typically used to model a binary dependent variable with the help of logistic function or using another name for logistic function as sigmoid function and given by equation(3).

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} \qquad (3)$$

This function helps the logistic regression to squeeze the values from (-k, k) to (0, 1). From mathematical point of view, logistic regression starts from a linear equation, where this equation is constituted of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between $[0, 1]$. As a result, we can choose a decision boundary and use this probability to conduct classification task. In logistic regression, the odds of an event occurring can be given by the formula:

$$\text{logit} = \log \text{odds where } odds = \frac{\text{P(event)}}{1-\text{P}(event\ )} = e^{w_0+w_1x_1+\cdots w_n n} \qquad (4)$$

The log odds of an event are given by taking (log) of equation (1)

$$logit(p) = \log\left(\frac{p(x)}{1-p(x)}\right) = \log(e^{w_0+w_1x_1+\cdots w_n x_n}) = w_0 + w_1x_1 + \cdots w_n x_n \qquad (5)$$

The odds ratio is log transformed to remove the restricted range as probabilities are in the range $p(x) \in [0, 1]$, $x \in R$. Log transformation changes this to values from negative infinity to positive infinity and moreover the log values are easier to interpret.

Naïve Bayes Classifier. –A classifier is a function (f) that maps input feature vectors $[x \in X]$ to output class labels $[y \in \{1, \ldots, C\}]$ where $[X]$ is the feature space. We typically assume $[X \in R^D \ or\ X = \{0, 1\}^D]$, i.e that the feature vector is a vector of (D) real numbers or (D) binary bits, but broadly speaking, we may mix discrete and continues feature. We assume the class labels are unordered (categorical) and mutually exclusive. This classifier is based on the Bayes' Theorem and the maximum posteriori hypothesis.Bayes theorem provides a way of calculating posterior probability $[P(A|B)]$ from $P(A)$, $P(x)$ and $P(A|B)$, (6).

$$P(c|B) = \frac{P(c \cap B)}{P(B)} = \frac{P(c) \cdot P(B|c)}{P(B)} \qquad (6)$$

Where:

$P(c|B)$: Posterior Probability of class (c, target) given predictor (x, attributes), meaning how often (c) happens given that (B) happens.

$P(c)$: is the prior probability of class.

$P(B|c)$: is the likelihood, which is the probability of predictor given class, meaning how often (B) happens given that (A) happens.

$P(B)$: is the prior probability of predictor, meaning how likely (B) is on its own.

In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes algorithms are used in *real time prediction*, where it is easy and fast to predict class of test data set, in *multi class prediction*, which performs well. Moreover, Naive Bayes algorithms are used in *Text classification*, in*Spam Filtering* and*Sentiment Analysis*, for instancein social media analysis, to identify positive and negative customer sentiments and lastly, to build *Recommendation Systems*that uses machine learning and data mining techniques to filter unseen information.

Fisher's Linear Discriminant. – Linear Discriminant Analysis (LDA) and sometimes also called Fisher's Linear Discriminant is a linear classifier that projects a p-dimensional feature vector onto a hyperplane that divides the space into two half-spaces where each half-space represents a class [+1 or −1].This methodology relies on projecting points into a line and the outputs of this methodology are precisely the decision surfaces or the decision regions for a given set of classes [13].The decision boundary (7) is characterized by the hyperplane's normal vector (w) and the threshold (*w*0).

$$[w_1 \ldots., w_p]^T [w_1 \ldots., w_p] + w_0 = w^T x + w_0 = 0 (7)$$

Given a new input vector ($X \in X^p$), classification is achieved by computing (8) and assigning the resulting class label ($y = -1 \; or \; y = +1$)to the input (x).

$$y = sign\,(w^T \cdot x + w_0) \qquad (8)$$

To compute (w), (LDA) assumes that the class-conditional distributions $P(x|c = 1 \; and \; P(x|c = 2)$ are normal distributions with mean (μ_c) and covariance (Σ_c) $for \; c \in$ {1,2}, [14]. LDA is an extremely popular dimensionality reduction technique which become critical in machine learning,and it is commonly used in the pre-processing step in machine learning and pattern classification projects.

3.2.2 Support Vector Machines

Support Vector Machines (SVMs) are a set of supervised learning methods used for *classification*, *regression*, and *outliers' detections*. This technique is very effective in high dimensional spaces and still remain effective in cases where number of dimensions is greater than the number of samples. Moreover, SVMsworks pretty well when there is a clear margin of separation between classes. On the other hand, SVM algorithm is not suitable for large data sets, do not provide probability estimates, and does not perform well when the target classes are overlapping. In addition, in cases where the number of features for each datapoint exceed the number of training data samples, the SVM technique will underperform [15], [16].

The essence of the SVM is simply involves finding a boundary that separates different classes from each other where in 2-dimentional space, the boundary is named a line, in 3-dimensionnaly space the boundary is named plane and finally in greater dimension than 3 the boundary is called hyperplane. The math behind the SVM is depicted below:

$$\text{MINIMIZE}_{\,a_0 \ldots., a_m} : \sum_{j=1}^{n} MAX\,\{0, 1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0)\, y_j\} + \lambda \sum_{i=1}^{m} (a_i)^2 \qquad (9)$$

where the first part of the formula (9):

$$\sum_{j=1}^{n} MAX\,\{0, 1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0)\, y_j\}$$focuses on minimizing the error, the number of falsely classified points, that the SVM makes, and

the second part of the formula (9):

$$\lambda \sum_{i=1}^{m} (a_i)^2$$

focuses on maximizing the margin, between the two classes.

SVM's are powerful and flexible supervised machine learning algorithms which are used both for classification and regression with astonishing real-life applications such as:

Inverse Geo-sounding Problems where the SVM's helps to determine the layered structure of the planet.

Seismic Liquefaction Potential, with great result accuracy. In this category we use the Standard Penetration Test (SPT) and the Cone Penetration Test (CPT) to check the occurrence and non-occurrence of liquefaction.

Protein Fold and Remote Homology Detection, where different methods to solve the kernel functions that are being used. The kernel functions help to find the similarity between different protein sequences.*Facial Expression Classification* where the SVM's have great use in various life-care systems in normal happy or sad look classification.

3.2.3 Quadratic classifiers

QuadraticDiscriminant Analysis (QDA) is closely related to Linear Discriminant Analysis (LDA) with the assumption that the measurements from each class are normally distributed.Quadratic discriminant analysis (QDA) is a variant of LDA that allows for non-linear separation of data.QDA is particularly useful if there is prior knowledge that individual classes exhibit distinct covariances. On the other hand, a disadvantage of QDA is that it cannot be used as a dimensionality reduction technique.

A *quadratic discriminant function* is a mapping $g : X \rightarrow R$ with

$$g(x) = \tfrac{1}{2} x^T W_x + w^T x + w_0, \qquad (10)$$

for some *matrix*$W \in R^{dxd}$, some *vector*$w \in R^d$, and some scalar $w_0 \in$ R. In quadratic discriminant function, the model parameter is $\theta = \{W, w, w_0\}$ and depending on (W) the geometry of (g)can be convex, concave or neither.

3.2.4 Kernel Estimation

A Kernel Distribution is a non-parametric representation of the Probability Destiny Function (PDF) of a random variable. The kernel distribution can be used when a parametric distribution cannot properly describe the data, or in case it is

wanted to avoid making assumptions about distribution of the data. Since the kernel density estimator is the estimated Probability Destiny Function (PDF), for any real values of (x), the kernel estimator's formula is given below:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \qquad (11)$$

Where $x_1,\ x_2, \dots, x_n$ are random samples from an unknown distribution, (n) is the sample size, $K(\cdot)$ is the kernel smoothing function, and (h) is the bandwidth.

K-Nearest Neighbor. –K-Nearest Neighbor (KNN) algorithm is very simply used to solve classification problems where (K) is the number of neighbors in KNN. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these. Broadly speaking, the distance can be any metric measuresuch as Hamming distance, Manhattan distance, Minkowski distance where the standard Euclidean distance is the most common choice. Despite the fact that KNN is very easy to implement, where the (K) value is needed and the distance function (e.g Euclidean), the KNN requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm. Moreover, there is no training period for it, where it stores the training dataset and learns from it only at the time of making real time predictions. As a result, the KNN algorithm is much faster than other algorithms that require training. On the other side, the KNN does not work well with large dataset where performance degradation appears and does not work well with high dimensional data where it becomes difficult for the algorithm to calculate the distance in each dimension. Moreover, the KNN needs to proceed with standardization and normalization before applying the algorithm to any dataset and needs to manually impute missing values and remove outlierssinceKNN is sensitive to noise in the dataset.

Assuming that we have a dataset where $(X)$is a matrix of features from an observation and (Y) is a class label, the formula that estimates the conditional distribution (Y) given (X)classifying an observation to the class with the highest probability is depicted below:

$$P_r(Y = j|X = x_0)^n = \frac{1}{k}\sum_{\mathrm{i}\in \mathrm{N}0} \mathrm{I}\,(y_i = j) \qquad (12)$$

Given a positive integer (k), the $(k - nearest\ neihbors)$ looks at the (k) observations closest to a test observation (x_0) the formula (12) estimates the conditional probability that it belongs to class (j).

The distance between the input data point and other points in the training data can be calculated as such:

Euclidean distance $d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$ (13)

Manhattan distance $d(x,y) = \sum_{i=1}^{p}|x_i - y_i|^2$ (14)

Minkowski distance $d(x,y) = \left(\sum_{i=1}^{p}(|\dot{x}_i - y_i|)^q\right)^{\frac{1}{q}}$ (15)

Where (x), is a point with coordinates $(x_1,\ x_2, \dots, x_p)$, and (y), is a point with coordinates $(y_1,\ y_2, \dots, y_p)$.It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used.

Hamming distance$d(x,y) = \sum_{i=1}^{p}|x_i - y_i|$ (16)

$$x_i = y_i \Rightarrow d(x,y) = 0$$

$$x_i \neq y_i \Rightarrow d(x,y) = 1$$

In case there is a mixture of numerical and categorical variables in the dataset it is necessary to standardize the training set.

3.2.5 Decision Trees Induction

Decision trees are the most popular representation of logic-based classifiers well presented in the literature [17], [18].There are three well known implementations of decision trees, the Classification and Regression Trees (CART) [17], and the Quinlan's univariate tree growing algorithm, which is known as Iterative Dichotomiser 3 algorithm (ID3) [19]. The third one is the C4.5 algorithm [20], which extends the ID3 algorithm by allowing the classification algorithm to deal with numbersand not just categorical values as the ID3 does.

Random Forests. –Random forest is a supervised learningalgorithm, very flexible, easy and one of the most used machine learning algorithmswhich produces great result most of the time. Random forest is based on bagging algorithm and uses Ensemble Learning technique where creates as many trees as possible on the subset of the data and combines the output of all the trees. As a result, we achieve overfitting problem reduction in the decision trees and also variance reduction, which eventually improve the accuracy. Random forest is used in both classification and regression problems and works well with categorical and continues variables. It uses rule-based approach of distance calculation and as a result no feature scaling (standardization and normalization) is needed.Nonlinear parameters do not affect the performance of a Random Forest unlike curve-based algorithms and is very stable and comparatively less impacted by noise.

On the other hand, the random forest creates a lot of trees, - for instance it creates one hundred trees in *Pythonsklearn library*- and as a result requires much more computational power and resources in contrast with the decision tree which is simple and does not require so much computational resources.It also requires much time for training as it combines a lot of decision trees to determine the class and it suffers interpretability and fails to determine the significance of each variable due to the ensemble of decision trees.

In case of regression problems, when using the random forest algorithm, the Mean Squared Error (MSE) is used to determine how the data branches from each node [21].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \qquad (17)$$

Where (N) is the number of data points, (f_i) is the value returned by the decision tree and $(y_i)$is the value of the data point that are tested at a certain node.

The above formula (17) calculates the distance of each node from the predicted actual value, helping to decide which branch is the better decision for your forest.

In case we perform random forests based on classification data, it is often used the Gini Index, or the formula used to decide how nodes on a decision tree branch.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \quad (18)$$

The formula (18) uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. The (p_i) represents the relative frequency of the class that is being observed in the dataset and (C) represents the number of classes.

The Random Forestsis a great choice in *Banking Sector* where problems such as loan default chance of a customer or for detecting any fraud transaction. Moreover, in healthcare sectors random forest can be used to identify the potential of a certain medicine or the composition of chemicals required for medicines. In addition, it be used in hospitals to identify the diseases suffered by a patient, the risk of cancer in a patient, and many other diseases where early analysis and research play a crucial role.

3.2.6 Neural Networks

Neural networks are a set of algorithms, that try to find relationships in a dataset to recognize patterns by simulating the way of how the human brain works.The NeuralNetworks in fact cluster and classify, group unlabeled data according to similarities and classify data when they have labeled dataset to train on.In other words, NeuralNetworks are software routines that can learn from existing data and solve complex real-world problems with an efficient way.NeuralNetwork algorithms are designed to cluster raw input, recognize patterns, and interpret sensory data and despite their multiple advantages, significant computational resources are required.

There are several methods to teach a NeuralNetworkfocusing on the main three learning paradigms examined below:

Supervised Learning. – Supervised Learning (SL) is the machine learning process which is done under the seen label of observation variables contrary to the Unsupervised Learning where the response variables are not available. In (SL), datasets are trained with the training set to infer a Machine Learning algorithm and then will be used to label new observations from the testing set.Supervised learning can be separated into two types of problems when it comes to data mining:

Classification where an algorithm is used to accurately assign test data into specific categories meaning that it recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined.

Regression where is used to understand the relationship between dependent and independent variables and makes projections for instance in businesses in terms of sales revenues. Very popular regression algorithms are the Linear Regression, the logistical regression, and the polynomial regression.

Supervised Machine Learning models can be used to build and advance a number of business applications such as, *Image and object-recognition*, where location, isolation, and object categorization out of videos or images, making them useful when applied to computer vision techniques and imagery analysis. Other applications that use Supervised Machine Learning (SML) models are in *predictive analytics*helping business leaders justify decisions or pivot for the benefit of the organization and in customer sentiment analysis, gaining a better understanding of customer interactions and can be used to improve brand engagement efforts.Broadly speaking, the Supervised Machine Learning models challenge is that require certain levels of expertise to structure accurately, the training is very time intensive, and the datasets can have a higher likelihood of human error, resulting in algorithms learning incorrectly.Unlike unsupervised learning models, supervised learning cannot cluster or classify data on its own.

Unsupervised Learning. –Unsupervised Learning(UL) is a machine learning techniquewhere there is no need for users to supervise the model and instead the model work on its own to discover patterns and information that was previously undetected.The goal for unsupervised machine learning is to model the underlying structure or distribution in the data in order to learn more about the data, by using machine learning algorithms. Clustering and Association are two main types of Unsupervised learning.Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real Artificial Intelligence (AI).

Unsupervised learning can be separated into two types of problems when it comes to data mining:

*Clustering*which is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.

Association where an association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database.

The most popular unsupervised learning algorithms are the following: *K-means clustering*, *K-nearest neighbors (KNN),Hierarchical clustering*, *Anomaly detection*, *Neural Networks*, *Principal Component Analysis*, *Independent Component Analysis*, *Apriori*, *algorithm* and *Singular Value Decomposition*.

Unsupervised learning is used for more complex tasks compared to supervised learning because, since in unsupervised learning there are no labeled input data. In addition, unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data. As a result, Unsupervised learning is more challenging than other strategies due to the absence of labels. On the other side Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output and moreover, the result of the unsupervised learning algorithm might be less accurate as input data is not labeled since the algorithm do not know in advance the exact output.

Reinforcement Learning. –Reinforcement Learning**(**RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences. In reinforcement machine learning, the machine learns by itself after making many mistakes and correcting themin turn. RL is one of the hottest research topics currently, is very common in robotics and its popularity is growing day by day.Reinforcement Learning method works on interacting with the environment, whereas the supervised learning method works on given sample data or

example.

There are two RL methods which is the "*Positive*" where it is defined as an event, that occurs because of specific behavior, increasing the strength and the frequency of the behavior and impacts positively on the action taken by the agent. The second method is called "*Negative*", and it is defined as strengthening of behavior that occurs because of a negative condition which should have stopped or avoided.The "*positive*" method helps to maximize performance and sustain change for a more extended period, contrary to the "*negative*"method which helps to define the minimum stand of performance. In addition, we have two widely used of Reinforcement Learning models which are the "*Markov Decision Process (MDPs)*" and the "*Q-Learning*". *Markov Decision Process (MDPs)*is mathematical framework to describe an environment in reinforcement learningby which the learner -often called agent- learns to behave in this interactive environment using its own actions and rewards for its actions. The agent discovers which actions give the maximum reward by exploiting and exploring them.*Q-Learning*is an off-policy reinforcement learning algorithm that seeks to find the best action to take, given the current state. It is considered off-policy because the q-learning function learns from actions that are outside the current policy, like taking random actions, and therefore a policy is not needed. "Table 1" and "Table 2" depicts the differences upon some criteria, between Reinforcement Learning, vs Supervised Learning and Unsupervised Learning, respectively.

Table 1. Reinforcement Learning vs Supervised Learning

Criteria	Reinforcement Learning	Supervised Learning
Definition	Learns by Interacting with the environment	Learns by using labelled data
Data type	No predefined data	Labelled data
Decision mode	Helps to take decisions sequentially	A decision considers the input given at the beginning
Dependency on decision	Labels to all dependent decisions are given	Labels are given for every decision
Problem types	Exploitation or Exploration	Regression and classification
Algorithms	Q – Learning, State-Action-Reward-State-Action(SARSA)	Linear and Logistic Regression, SVM, KNN etc.
Applications	Robotics, Machine learning, Aircraft control, AI	Risk Evaluation, Forecast Sales, Object recognition

Table 2. Reinforcement Learning vs Unsupervised Learning

Criteria	Reinforcement Learning	Unsupervised Learning
Definition	Learns by Interacting with the environment	Learns by using unlabeled data without any guidance
Data type	No predefined data	Unlabeled data
Decision mode	Helps to take decisions sequentially	The model work on its own to discover patterns and information
Dependency on decision	Labels to all dependent decisions are given since RL is dependent.	Unsupervised model, provides unlabeled data
Problem types	Exploitation or Exploration	Association and Clustering
Algorithms	Q – Learning, State-Action-Reward-State-Action(SARSA)	K – Means, C – Means, Apriori.
Applications	Robotics, Machinelearning, Aircraft control, robot, AI	Clustering, Anomaly Detection, Visualization, Pattern recognition, find association rules

3.2.7 Learning Vector Quantization

Learning Vector Quantization (LVQ), is a type of artificial neural network algorithm that lets you choose how many training instances to hang onto and learns exactly what those instances should look like, supporting both binary (two-class) and multi-class classification problems. The LVQ is based on prototype supervised learning version of vector quantization where is used when we have labelled input data. This learning technique uses the class information to reposition the Voronoi vectors slightly, so that to improve the quality of the classifier decision regions and is very useful for pattern classification problems [22].

Learning Vector Quantization is a neural net that combines competitive learning with supervision and used for pattern classification. For LVQ we suppose training data $V \subseteq \mathbb{R}^n$ with each $v \in V$has a class label $c(v) \in C = \{1, \dots, C\}$ indicating to which class (v) belongs. Further, we assume (M) prototypes $W = \{w_k \in \mathbb{R}^n, k = 1 \dots M\}$ with labels $c(w_k) \in C$ such that at least one prototype is assigned to each other [23]. Indicative innovative application uses LVQ algorithm is the proposed one in real time adaptive traffic signal control [24].

3.2.8 Boosted Decision Trees Method

When a decision tree is a weak learner, the resulting algorithm is named gradient boosted trees where boosted means that each tree is dependent on prior trees. As a result, boosting in a decision tree is a method of combining many weak learners

(trees) into a strong classifier and tends to improve accuracy with some small risk of less coverage [25], [26]. Each tree attempts to minimize the errors of previous tree.Trees in boosting are weak learners but adding many trees in series, meaning combining a learning algorithm in series, it is achieved a strong learner from many sequentially connected weak learners, making boosting a highly efficient and accurate model.Since trees are added sequentially, boosting algorithms learn slowly. In statistical learning, models that learn slowly perform better. However, the number of trees for instance, in gradient boosting decision trees, is very critical in terms of overfitting whereadding too many trees will cause overfitting so it is very important to stop adding trees at some point.

The theory of a decision tree has the following components: a *root* node which is the first node and the starting point of the tree; *branchesarrows* which connect one node to anothershowing the flow from question to answer. Nodes that have child nodes are called *interior* nodes. *Leaf*or *terminal* nodes are nodes that do not have child nodes and represent a possible value of targetvariable given the variables represented by the path from the *root*. The *branching factor* (b) represents the number of children at each node.

The advantages of Decision Trees (DC) can be summarized as simple to understand and easy to interpret and visualized, where all kinds of data can be handled, making them widely used. DC are considered to be non-parametric meaning that have no assumptions about the data point's space or the classifier's structure. DC are robust since require less effort from users for pre-processing data. They are not influenced by outliers and missing values either.On the other hand, overly complex trees can be developed due to overfitting. Moreover, Decision Trees can be unstable because small variations in the data might result in a completely different tree being generated.In addition, Decision Tree learners create biased trees if some classes are more likely to be predicted or have a higher number of samples to support them. The optimality is one more disadvantage, where the problem of learning an optimal decision tree is known to be NP-complete (nondeterministic polynomial-time complete), since the number of samples or a slight variation in the splitting attribute can change results drastically.

3.3 Regression Analysis

Regression analysis is a well-known statistical learningtechnique used to estimate the relationship between a dependent variable with one or more independent variables, where the independent variable is used as an assumption input that is changed in order to see the impact on a dependent variable. In other words, Regression Analysis is a data mining process that helps to understand the correlation and independence of the variables to determine which factors matter most and which factors can be ignored and eventually, how these factors influence each other.

There are many types of regression analysis techniques, depending on number of factors such as, the type of target variable, the shape of the regression line, and the number of independent variables.Regression Analysis has a wide range of real-life application such as, financial forecasting, Sales and promotions forecasting. The different types of regression are briefly explained below:

Linear Regression. – Linear regression model comprises of a predictor variable and a dependent variable related to each other in a linear fashion. The general linear regression model can be stated by the equation (19):

$$y_i = \beta_0 + \beta_1 X_{1\iota} + \beta_2 X_{2\iota} + \cdots + \beta_k X_{k\iota} + \varepsilon_i \qquad (19)$$

where (β_0) is the intercept, $(\beta_1' s)$ are the slope between (Y) and the (X_ι), and (ε) *pronounced epsilon*is an intercept and (e) the error term that captures errors in measurement of (Y) and the effect on (Y) of any variables missing from the equation that would contribute to explaining variations in (Y). The linear regression should not be used to analyze big size data.

Logistic Regression. –Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete for instance (0 or 1), (true or false).This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

The logistic regression model is based on the logistic function and can be stated by the equation (20):

$$F(x) = \frac{L}{1+e^{-k(x-x_o)}} (20)$$

Where (x_0), is the (x), value of the sigmoid's midpoint; (L) is the curve's maximum value; and (k), the logistic growth rate or steepness of the curve.

Logistic regression works best with large data sets that have an almost equal occurrence of values in target variables. The dataset should not contain a high correlation between independent variables (a phenomenon known as multicollinearity), as this will create a problem when ranking the variables.Logistic regression can suffer from *complete separation*. If there is a feature that would perfectly separate the two classes, the logistic regression model can no longer be trainedbecause the weight for that feature would not converge, due to the fact that the optimal weight would be infinite.

Ridge Regression. –Ridge Regression is a model tuning method that is used to analyze any data that suffers from multicollinearity.Ridge Regression performs (L_2) regularizationand is usually used when there is a high correlation between the independent variables. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

The Ridge Regression formula can be stated below:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (21)$$

Where the $\left[\lambda \sum_{j=1}^{p} \beta_j^2\right]$ represents the L2 regularization element. If lambda is zero, then we get Ordinary Least Squares (OLS).However, the high value of lambda will add too much weight which will result in model *under-fitting*, so it is important how we choose the parameter lambda for our model.Overfitting problems may lead to inaccurate and unstable model building so, a technique that helps minimize the overfitting problem in Machine Leaning (ML) models is known as regularization. Ridge regression uses L2 regularization compared to Lasso regression which uses L1 regularization.

Lasso Regression. – Lasso Regression is like linear regression, but it uses a shrinkage technique where the coefficients of determination are shrunk towards zero. Since the Linear regression gives you regression coefficients as observed in the dataset. The Lasso Regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets.Lasso regression penalizes less important features of your dataset

and makes their respective coefficients zero, thereby eliminating them. Hence, it provides with the benefit of feature selection and simple model creation. The Lasso Regression formula can be stated below:

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \qquad (22)$$

Where (λ) denotes the amount of shrinkage.$(\lambda = 0)$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model. $(\lambda = \infty)$ implies no feature is considered. The bias increases with increase in $(\lambda)$and variance increases with decrease in (λ).

Table 3. Differences Between Lasso and RidgeRegression

RidgeRegression	Lasso Regression
It makes use of the L2 regularization technique.	It makes use of the L1 regularization technique
It performs feature weight updates as the loss function has an additional squared term.	It performs the feature weight updates as the loss function has an additional term containing the L1 norm of the weights vector.
It drives down the overall size of the weight values during optimization and reduces overfitting.	It drives down the overall size of the weight values during optimization and reduces overfitting.

Polynomial Regression. –Polynomial regression is a model which transforms data points into polynomial features of a given degree, and models them using a linear model. It works in a similar way to multiple linear regression with a little modification but uses a non-linear curve and it is used when data points are present in a non-linear fashion.Polynomial regression is one of several methods of curve fitting, where curve fitting is a process of constructing the best fit line that passes through all the data points, is not a straight line but a curve line.With polynomial regression, the data is approximated using a polynomial function that takes the form (23).

$$f(x) = c_0 + c_1 x + c_2 x^2 \ldots c_n x^n + residual\ error \quad (23)$$

Where (n) is the degree of the polynomial and (c) is a set of coefficients.

Polynomial Regression provides the best approximation of the relationship between the dependent and independent variable and fits a wide range of curvature. On the other hand, it is very sensitive to the outliers where the presence of one or two outliers in the data can seriously affect the results of the nonlinear analysis. Moreover, there are fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

3.4 Outlier Detection

An outlier is an observation that diverges from the overall pattern on a sample and mainly indicate a variability in a measurement, experimentalerrors, or a novelty. The outliers can be in two categories, the univariatewhen looking for instance, at a distribution of values in a single feature space, and the multivariance in n-dimensional space. In n-dimensional space, there is a need to train a model.Moreover, the outliers can be come out depending on the different type of data: such as point outliers which are single data points that appears far from the rest of the distribution, contextual outliers, that could be noise in data e.g.,background noise signal when doing speech recognition or collective outlierssuch as a signal that may indicate the discovery of new phenomena.

Most common causes of outliers on a data set could bedata entry errors due to human mistake, or measurement errors due to instrument accuracy, experimental errors, data processing and sampling errors.In machine learning and in any quantitative discipline the quality of data is as important as the quality of a prediction or classification model, that is why, detecting outliers is of a major importance for example in Physics, Economy, Finance, Machine Learning and Cybersecurity.

Some of the most popular methods for outlier detection are the *Z-score or Extreme Value Analysis* (parametric), *Probabilistic and Statistical Modeling* (parametric), *Linear Regression Models*such as *Principal Component Analysis* (PCA), and *Least Median of Squares* (LMS)[27],the *Proximity BasedModels* (non-parametric), *Information Theory Models* and last the *High Dimensional Outlier Detection Methods*.

3.5 Predictive Modeling Techniques

When we refer to Predictive Analytics, we mean the use of statistical and machine learning techniques to identify the likelihood of future outcomes based on historical data with the final purpose to streamline decision making producing new insights. Predictive analytics is used to predict behavior and trends, to understand customers and to improve strategic decision making and business performance. Some of the common uses of predictive analytics includes the domain of *fraud detection and security*, *Marketing*, *Operation* and *Risk Identification*. The most used Predictive Analytics models includes the *Classification Model*, which are best to answer Yes or No questions, the *Clustering Model* which sorts data into separate nested smart based on similar attributes. Using the clustering model, it can be quickly separate customers into similar groups based on common characteristics and devise strategies for each group at a larger scale.*Forecast Model* is another predictive technique which can be applied wherever historical numerical data is available such as a call center to predict how many supports calls, they will receive per hour. Moreover, *Outliers* and *Time Series models* are used as predictive techniques, where anomalous data entries within a dataset are identifiedor identify sequence of datapoints using time as an input parameter, respectively.

Broadly speaking, the common predictive algorithms can by separated into two groups: *Machine Learning* and *Deep Learning*. Machine learning involves structural data, comprise both linear and nonlinear varieties, train more quickly, while nonlinear are better optimized for the problems they are likely to face which is more often nonlinear. Deep Learning is a subset of machine learning that is more popular to deal with audio, video, text, and images.With machine learning predictive modeling, there are several different algorithms that can be applied, where the most common are the *Random Forest*, the *Generalized Linear Model (GLM) for two Values*, the *Gradient Boosted Model (GBM)*, the *K-Means*, and the *Prophet* algorithm.

3.6 Sequential Patterns

Similar to association rules mining, by using sequential patterns mining, it can discoverstatistically interesting and useful patterns and rules in a large-scale table that contains sequences of transactions [28]. A sequential pattern is a frequent subsequence existing in a single sequence or a set of

sequences. A sequence $\alpha = \langle \alpha_1 \alpha_2 \ldots \alpha_n \rangle$ is a subsequence of another sequence $\beta = \langle b_1 \, b_2 \ldots b_m \rangle$if there exist integers $1 \leq j_1 < j_2 < \cdots < j_n \leq m$ such that $\alpha_1 \subseteq b_{j_1}$, $\alpha_2 \subseteq b_{j_2}$, … $\alpha_n \subseteq b_{j_n}$ [29].

The algorithms classification suitable and used for sequential pattern mining are the following: *Apriori-like algorithms*, *BFS (Breadth First Search)-based algorithms*, *DFS (Depth First Search)-based algorithms*, *closed sequential pattern-based algorithms*, and *incremental-based algorithms* [30], [32].

The sequential data mining techniques are suitable in *healthcare*where patterns observed in symptoms of a particular disease, and patterns in daily activity and health data, in Education and Web Usage Mining, in Text Mining to discover trends, for text categorization, for document classification and authorship identification. Moreover, the sequential mining techniques are used in Bioinformatics domain for predicting rules for organization of certain elements in genes, for protein function prediction, for gene expression analysis, for protein fold recognition and for motif discovery in DNA sequences.Pattern mining can be used in the field of telecommunications for mining of group patterns from mobile user movement data, for customer behavior prediction, for predicting future location of a mobile user for location-based services and for mining patterns useful for mobile commerce [32].

4. CLUSTERING METHOD

Cluster is a group of objects that belongs to the same class, meaning that similar objects are grouped in one cluster and dissimilar objects are grouped in other clusters based on similarities.

Cluster analysis is a statistical method to group data into subsets with related characteristics to understand the internal structure of the data. Clustering is considered as one of the most important unsupervised learning methods due to the fact that no information is provided about the best answer for any of the object and in fact, it can reveal undetected correlations in a complex data set.Clusters are regions where the density of similar data points is high and in general clusters are seen moreoften in a spherical shape, but it can of any shape. It depends on the type of the algorithm we use which decides how the clusters will be created.

4.1 Clustering Categories

4.1.1 Partitioning Clustering Method

This method is one of the most popular choices for analysts to create clusters where the clusters are partitioned based upon the characteristics and the similarity of the data point. Partitioning-based clustering algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. Since the number of data points in any data set is always finite and, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods. The number of different partitions for (n)observations into (K)groups is a Stirling number of the second kind, which is given by the following form (23):

$$S\binom{(K)}{n} = \frac{1}{K!} \sum_{i=0}^{i=K} (-1)^{K-i} \binom{K}{i} i^n \qquad (23)$$

From the above can be seen that enumeration of all possible partitions is impossible for even relatively small problems and moreover the problem is even more demanding when additionally, the number of clusters is unknown. As a result, the number of different combinations is the sum of the Stirling numbers of the second kind (24).

$$\sum_{i=1}^{i=K_{max}} S\binom{(i)}{n} \qquad (24)$$

Where (K_{max})is the maximum number of cluster and it is obvious that$K_{max} <= n$

Therefore, more practical approach than exhaustive search is the iterative optimization.The advantages and disadvantages of Partitioning Clustering Method are presented below in "Table 4".

Table 4. Partitioning Clustering Method(pros and cons)

Partitioning Clustering Method	
Advantages	**Disadvantages**
Relatively scalable and simplicity	Poor cluster descriptors and often requires long computation time
Suitable for datasets with compact spherical clusters that are well-separated	High sensitivity to initialization phase, noise, and outliers since it works with squared distances.
Optimal for certain criteria	Needs initial K (objects) and has long computational time

The algorithms that fall into this category are as follows:

K-Means Clustering Algorithm. –K-Means clustering is one of the most widely used algorithms where the value (k) is defined by the user. Basically, K-Means is an iterative process that divides a given data set into (K) disjoint groups based upon the distance metric used for the clustering. In other words, the algorithm adjusts the assignment of objects to the closest current cluster mean until no new assignments of objects to clusters can be madeunder a new iterative process [33]. K-means is perhaps the most widely used clustering principle, and especially, the best-known of the partitioning-based clustering methods that utilize prototypes for cluster presentation.Even that the simplicity is a good advantage, it has some major drawbacks such as: it is very hard to specify number of clusters in advance, and due to the fact that it works with squared distances, it's also sensitive to outliers. K-Means algorithm has linear time complexity, and it can be used with large datasets conveniently. As an unsupervised clustering algorithm K-Means provides many benefits with unlabeled big data. For instance,even if the data has no labels (class values or targets) or even column headers, K-Means will still successfully cluster the data. K-Means is also very easy to use by using default parameters in the Scikit-Learn implementation such as number of clusters where (8)is by default, the maximum iterations where (300) is by default), and like the initial centroid initialization where is 10 by default. All these default parameters can easily be adjusted later on to suit the task goals.Moreover, K-Means returns clusters which can be easily interpreted and even visualized. Just a few examples use cases could be "*customer segmentation*", "*logistic optimization*", "*user suggestions*", "*patient management*", "*trial management*" and "*fraud detection*". On the other hand, K-Means, introduces

drawbacks such as: "Result repeatability" where K-Means algorithm results will differ based due to random centroid initialization. Apart from the fact that K-Means Algorithm needs manual intervention in some parameters (e.g *n_clusters*need to be optimized, adjusted, and reassessed a few times, or *max_iter* and *init*), K-Means algorithm creates spherical clusters that cover the whole dataset without be possible to exclude outliers or certain sample groups.

In summation, the K-Means advantages, and disadvantages can be depicted in the "Table 5" below:

Table 5. K-Means(pros and cons)

K-Means advantages and disadvantages	
Advantages	**Disadvantages**
It's very simple and flexible identify unknown groups of data from complex data sets.	Difficult to predict K-value and does not work well with clusters of different size and density
If variables are huge, then K-Means is most of the times computationally faster than hierarchical clustering, if we keep k smalls. Optimal for certain criteria and suitable in a large dataset	Needs initial K (objects) and has long computational time. When dealing with a large dataset, conducting a dendrogram technique will crash the computer due to a lot of computational load and Ram limits
It's efficient at segmenting the large data set depending on the shape of the clusters. K-means work well in hyper-spherical clusters	K-means doesn't allow development of an optimal set of clusters and for effective results, you should decide on the clusters before
Compared to hierarchical algorithms, k-means produce tighter clusters especially with globular clusters	Lacks consistency where. A random choice of cluster patterns yields different clustering results. K-means algorithm can be performed in numerical data only.
K-means segmentation is linear in the number of data objects thus increasing execution time. Generalize to cluster of different shapes and sizes, for instance elliptical clusters.	It produces cluster with uniform size even when the input data has different sizes and it's very sensitive to scale where rescaling the datasetvia normalization or standardization will change the final results

PAM(K-Medoids)Algorithm. –Partitioning Around Medoids(PAM) Algorithm was introduced by Kaufman and Rousseeuw based on (k) representative objects, named medoids, among the objects of the dataset [34].In k-medoids clustering, each cluster is represented by one of the data points in the cluster.These points are named cluster medoids.The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other members of the cluster are minimal. It corresponds to the most centrally located point in the cluster.Objects are tentatively defined as medoids and are placed into a set (S) of selected objects. If (O) is the set of objects that the set$U = O - S$ is the set of unselected objects.The aim of the algorithm is to minimize the average dissimilarity of objects to their closest selected object, hence, to find themost centrally located objects withinthe clusters.K-Medoids can be considered as a robust alternative to k-means clustering, meaning that the algorithm is less sensitive to noise and outliers, compared to k-means. This is due to the fact that it uses medoids as cluster centers instead of means used in k-means method.The k-medoids algorithm requires the user to specify the (k), and the number of clusters to be generatedwhere the silhouette method is a nice approach to determine the optimal number of clusters. The complexity of k-Medoids is $O(N^2\,KT)$ where (N) is the number of samples, (T) is the number of iterations and (K) is the number of clusters, and this makes it more suitable for smaller datasets compared to k-means which is $\mathrm{O}(\mathrm{N}KT)$.

The advantages and disadvantages of K-Medoids Method are presented below in "Table 6".

Table 6. K-Medoids(pros and cons)

K-Medoids advantages and disadvantages	
Advantages	**Disadvantages**
K-Medoidscan be more robust that k-means in the presence of noise and outliers.	K-Medoids is not suitable for clustering non-spherical (arbitrary shaped) groups of objects.
K-Medoids is efficiently for small datasets while does not scale well for large datasets.	K-Medoidsmay obtain different results for different runs on the same dataset because the first *k* medoids are chosen randomly
K-Medoids is more flexible as it can use any similarity measure.	In k-Medoids, there is a need to specify the value (k) (the number of clusters) in advance

CLARA Algorithm.–Clustering Large Applications (CLARA) Algorithm, is an extension to k-Medoids (PAM) methods dealing with data, comprising a large number ofobjectsin order to reduce computing time and RAM storage problem using the sampling approach.

In CLARA concept, instead of finding medoids for the entire data set, this algorithm considers a small sample of the data with fixed size and applies the PAM algorithm to generate an optimal set of medoids for the sample. The algorithm repeats the sampling and clustering processes a pre-specified number of times in order to minimize the sampling bias. The outcome of this iteration corresponds to the set of medoids with the minimal cost.

4.1.2 Hierarchical Clustering Method

Hierarchical Clustering algorithms can be Agglomerative (bottom-up approach) or divisive (also called as Top-Down Approach) and groups the clusters based on the distance metrics.

In Agglomerative clustering, each data point acts as a cluster initially and then pair of clusters successfully merged one by one until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation

of the objects, named dendrogram.

Divisive is the opposite of Agglomerativeit starts off with all the points into one cluster and divides them to create more clusters.

A key step in a hierarchical clustering is to select a distance measure such as the Manhattan distance, which is equal to the sum of absolute distances for each variable. A more common

measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum [35].

The advantages and disadvantages of Hierarchical Clustering Method are presented below in "Table 6".

Table 6. Hierarchical Clustering Method(pros and cons)

Hierarchical Clustering Method	
Advantages	**Disadvantages**
Fast computation and there is no need to pre-define the number of clusters (k).	Hard to define levels for clusters.Sensitivity to noise and outliers
Embedded flexibility regarding the level of granularity.	Rigid, cannot correct later for erroneous decisions made earlier.
Very well suited in terms of problems involving point linkages.	No ability to make corrections when the splitting/merging decision is taken.
Accepts any valid measure of distance	Lack of interpretability in terms of the cluster descriptors.
Good for data visualization providing hierarchical relation between clusters.	It cannot perform well on a large database

Some typical examples of Hierarchical Clustering algorithms are the following:

CURE Algorithm.– Clustering Using REpresentatives (CURE) is an efficient agglomerative hierarchical clustering algorithm suitable for large datasets, that adopts a balance between centroid based and all point extremes. It starts with a single point cluster, and moves to merge with another cluster, until the desired number of clusters are formed. CURE algorithm, instead of using one point centroid, as in most of data mining algorithms, uses a set of well-defined representative points, so that to efficiently handle the clusters and eliminate the outliers. Compared with K-means clustering, it is more robust to outliers and capable of identifying clusters having non-spherical shapes and size variances. CURE is robust to outliers and can handle large datasets by combining random sampling and partitioning method. Contrary to K-means, CURE supports non-spherical shaped clusters and densities with the disadvantage that cannot handle differing densities. In addition, compared with k-means which deals with data points in spherical datasets, CURE algorithm deals with outliers of non-spherical clusters with random sampling and partitioningto reliably find clusters of arbitrary shape and size.

ROCK Algorithm.– The RObust Clustering using linKs (ROCK) is a robust agglomerative hierarchical-clustering algorithm based on the notion of links. It is suitable for handling large datasets and most suitable for clustering data that have Boolean and categorical attributes. In this algorithm, cluster similarity is based on the number of points from different clusters that have neighbors in common. The concept of *links* is to measure the similarity/proximity between a pair of data points, where the ROCK algorithm employs links and not distances when merging clusters. ROCK algorithm performs well on real and synthetic categorical dataset, and respectably on time-series data compared to traditional algorithms.

CHAMELEON Algorithm.–CHAMELEONis an agglomerative hierarchical clustering algorithm that uses dynamic modeling in which measures the similarity of two cluster on a dynamic model approach.Adapt to the characteristics of the data set to find the natural clusters where the main property is the relative closeness and relative inter-connectivity of the cluster. Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters.CHAMELEON works by using a two-phase algorithm so that to find the clusters in the datasets. At first phase uses a graph partitioning algorithm to cluster the data into a large number of small sub-clusters where in second phase uses an agglomerative hierarchical clustering algorithm to find the actual clusters by iteratively merges subclusters based on their similarity.The key advantage of the CHAMELEON algorithm is that it determines the pair of most similar sub-clusters by considering both the inter-connectivity as well as the closeness of the clusters.One of the areas of application is spatial datasets.

4.1.3 Density Based Clustering Method

An interesting property of density-based clustering is that these algorithms do not assume clusters to have a particular shape.In this method the clusters are created based on the density of the data points represented in the data space, namely, the density-based clustering algorithm considers cluster as a dense area separated by sparse area in data space.The data points in the sparse regionare considered as noise or outliers. This clustering method creates clusters of arbitrary shapes.Partition-based and hierarchical clustering techniques are highly efficient with normal shaped clusters. Moreover, when it comes to arbitrary shaped clusters or detecting outliers, density-based techniques are more efficient and in particular, are very efficient at finding high-density regions and outliers.

The advantages and disadvantages of Density Based Clustering Method are presented below in "Table 7".

Table 7. Density Based Clustering Method(pros and cons)

Density Based Clustering Method	
Advantages	**Disadvantages**
There is no need to require a-priori specification of number of clusters in advance	Cannot perform well and not suitable with large differences in densities
Ability to identify noise data while clustering and to find arbitrarily shaped	Not suitable for high dimensional data in case of

clusters	DBSCAN and OPTICS
Works well in presence of noisein case of OPTICS but not well in case of DBSCAN	Sensitive to density parameters that should be selected carefully

Some typical examples of Density Based Clustering algorithms are the following:

DBSCAN Algorithm. –The Density-Based Spatial Clustering of Applications with Noise Algorithm(DBSCAN)is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes. The DBSCAN is the most well-known density-based clustering algorithm first introduced in 1996 by Ester et.al [36]. Unlike k-means, DBSCAN does not require the number of clusters as a parameter, where it infers the number of clusters based on the data, and it can discover clusters of arbitrary shape.The DBSCAN algorithm is the fastest of the clustering methods, provided that there is a very clear *Search Distance* to use. The advantages can be summarized as such: DBSCAN does not require a-priori specification of number of clusters, is able to identify noise data while clustering and to find arbitrarily size and arbitrarily shaped clusters. The disadvantages can be summarized as such: DBSCAN fails in case of varying density clusters, and in case of neck type of dataset and moreover, does not work well in case of high dimensional data.

OPTICS Algorithm. – The Ordering Points to Identify Clustering Structure (OPTICS) Algorithm, works as an extension of DBSCAN. The only difference is that it does not assign cluster memberships but stores the order in which the points are processed meaning that for each object stores the *Core Distance* and the *Reachability distance*. The main idea of OPTICS algorithm is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do that, the points of the database are ordered in a way that spatially closest points become neighbors in the ordering. Moreover, for each point a special distance is stored which represents the density that must be accepted for a cluster so as both points belong to the same cluster.

Like DBSCAN, OPTICS requires two parameters: the (ε)which describes the maximum distance to consider and the ($MinPts$) which describes the number of points needed to form a cluster. The key parameter to DBSCAN and OPTICS is the ($MinPts$) parameter which roughly controls the minimum size of a cluster. If this parameter is set too low everything will become clusters where if is set too high at some point there won't be any clusters anymore, but only noise.OPTICS clustering method require more memory to determine the next data point which is closest to the point currently being processed in terms of Reachability Distance. As a result, requires more computational power because the nearest neighbor queries are more complicated compared to radius queries in DBSCAN. Moreover, the OPTICS clustering technique does not need to maintain the (ε)parameter and is relatively insensitive to parameter settings.

4.1.4 Model Based Clustering Method

Model-based clustering is a statistical approach to data clusteringassuming that data points are generated according to a certain probability distribution model, and the clusteringprocess is to adapt all data points to some predefined mathematical models. As a result, the algorithms that falls in this category, can automatically identify the number of clusters and outliers in data points according to the selected mathematical model.However, the noise and outliers are considered while calculating the standard statistics for having robust clustering.In order to form clusters, these clustering methods are classified into two categories: *Statistical* and *Neural Network* approach methods. In the *statistical* approach the model-based algorithms follow probability measures to determine clusters and in *Neural Network* approach, input and output are associated with unit carrying weights.

Representative algorithms that fall into this category are as follows:

GMM Algorithm. – Gaussian mixture model (GMM) algorithm is based on the probability model where the data is decomposed into several models based on the Gaussian probability density function. The GMM algorithm results are expressed in terms of probabilities, which are more visual and can be used to predict in a certain area of interest based on these probabilities. On the other hand, it is necessary to use complete sample information for prediction and lose effectiveness in high-dimensional space and this is considered as a disadvantage.

SOM Algorithm. –Self Organized Maps (SOM) algorithm is based on neural network modelthe input layer receives input signals, and the output layer is arranged by a neuron into a two-dimensional node matrix in a certain way. SOM algorithm has the advantage to map to a two-dimensional plane to achieve visualization and obtain higher-quality clustering results. On the other hand, as a disadvantage is that the calculation complexity is high, and the result depends to a certain extent on the choice of experience.

4.1.5 Grid Based Clustering Method

In grid-based clustering, the dataset is represented into a grid structure which comprises of grids (also named cells) to design a grid-structure.Grid-based methods work in the object space instead of dividing the data into a grid where grid is divided based on data characteristic.After partitioning the datasets into cells, it computes the density of the cells which helps in identifying the clusters.One of the greatest advantages of these algorithms is its reduction in computational complexity. They are more concerned with the value space surrounding the data points rather than the data points themselves. The Grid-based clustering method has fast time of processing than another way and depends on the number of cells in the space of quantized each dimension.Moreover, it applies to any attribute type and provides flexibility related to the level of granularity.

Representative algorithms that fall into this category are as follows:

STING Algorithm. – Statistical Information Grid Approach (STING) Algorithm, the dataset is divided recursively in a hierarchical manner where each cell is further sub-divided into a different number of cells capturing in turn the statistical measures of the cells. STING Algorithm has high efficiency and low time complexity. On the other hand, the fact that the clustering quality is affected by the granularity of the bottom layer of the grid structure, can be considered a disadvantage.

WaveCluster Algorithm. – In this algorithm, the dataspace is represented in form of wavelets where contains a n-dimensional signal helping to identify the clusters. The parts

of the signal with a lower frequency and highamplitude indicate that the data points are concentrated representing clusters that identified by the algorithm. The WaveCluster Algorithm is a fast multi-resolution algorithm where high-resolutioncan obtain detailed information, and low-resolution can obtain contour information. when the processed clusters have no obvious edges the clustering effect is poor, and this can be considered as a disadvantage.

CLIQUE Algorithm. – Clustering in Quest (CLIQUE) Algorithm is a *density-based* and *grid-based* subspace clustering algorithm. *Grid-based* because it discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell. D*ensity-based* since a cluster is a maximal set of connected dense units in a subspace. The CLIQUE algorithm discovers minimal descriptions of the clusters and automatically identifies subspaces of a high dimensional data space that allow better clustering than original space using the Apriori principle.CLIQUE Algorithm is good at handling high-dimensional data and large datasets but has the disadvantage of having low the accuracy of clustering. Another weakness as it happens in all grid-based clustering approaches, the quality of the results crucially depends on the appropriate choice of the number and width of the partitions and grid cells

5. CONCLUSION

The structured data accounts for less than 20 percent of all data whereas a much bigger percentage of all the data is unstructured data in our world. In this paper, it is clear thatnot all the algorithms are suited for all kind ofdatasets. There are different tools, data mining algorithms and methods which are used to analyze the datasets and as result, the choice of the best algorithm to use for a particular analytical task is a big challenge to data mining researchers.

The supervised learning algorithms are those for which the class attribute values for the dataset are known before runningthe algorithm. These kinds of datasets are named labelled data or training data. Classification for example, is a popular data mining technique referred to as a supervised learning technique because an example dataset is used to learn the structure of the groups. Examples of supervised learning algorithms commonly used in data mining are the Classification category (*Decision tree Learning*, *Naive Bayes Classifiers*, *K-Nearest Neighbor*, *Support Vector Machine* etc. algorithms), Regression category (*Linear and Logistic regression*, etc. algorithms).

In the unsupervised learning algorithms, there is no need for users to supervise the model and instead the model work on its own to discover patterns and information that was previously undetected. Association Rule Learning for instance, is one of the unsupervised data mining techniques in which an item set is defined as a collection of one or more itemsthat is used to discover relationships between variables in datasets. Normally when discussing the unsupervisedlearning, most researchers focus on clustering.In clustering, the data is often unlabeled where the label for eachinstance is not known to the clustering algorithm, and this is main difference between supervised and unsupervisedlearningExamples of unsupervised learning algorithms commonly used in data mining are Clustering category (*K-Means*, *Density based*, *Apriori* etc. algorithms).

In bottom line, Data mining techniques such as classification, clustering, prediction, association, etc., it helps to find the patterns, forecasting, discovery of knowledge etc., in different business domain to decide upon the future trends in businesses to grow.

6. FUTURE WORK

The current review acts as a guideline to data mining researchers to have an outlook on what algorithms to choose based on their needs and based on the given datasets. The next step is to design and deploy a High-Performance Computer (HPC) based on Raspberry Pi 4, to benchmark the efficiency of a Beowulf, Hadoop, and Spark Cluster architectures suited to deal with Big Data Analytics needs [37], [38]. Following the successful results referred above, the final goal is to proceed with a comparative study on various clustering Algorithms, Bringing HPC to Big Data Algorithms. For instance, a comparative study between parallel K-Means and K-Medoids using Message Passing Interface (MPI) and MapReduce in a Hadoop architecture would be very interesting topic to see the results when the computes nodes are increased gradually in terms of computing performance [38]. Moreover, a survey of parallel Clustering Algorithms based on Spark architecturewith Raspberry Pi 4 would be another very interesting topic to see the results when the computes nodes are increased gradually in terms of computing performance.

7. ACKNOWLEDGMENTS

My sincere gratitude to assistance Professor Ioannis S. Barbounakisfor the precious suggestions, and knowledge contribution for the successful completion of this paper.

8. REFERENCES

[1] X. Zhu, B. Song, Y. Ni, Y. Ren, R. Li, (2016). Business Trends in the Digital Era:Evolution of Theories and Applications, Springer.

[2] Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6.

[3] McAfee, A. and Brynjolfsson, E. (2012). Big Data. The Management Revolution. Harvard Business Review, 90(10), pp. 60–9.

[4] Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024. Statista 2020. [Online]. Available:https://www.statista.com/statistics/871513/worldwide-data-created/.

[5] Brands, K. (2014). Big Data and Business Intelligence for Management Accountants. Strategic Finance, 96(6), pp. 64–5.

[6] Gandomi, A. and Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analysis. International Journal of Information Management, 35(2), pp. 137–44.

[7] Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. and Khan, S.U. (2015). The rise of "Big Data" on cloud computing: Review and open research issues. Information Systems, 47(1), pp. 98–115.

[8] Bendler, J., Wagner, S., Brandt, T. and Neumann, D. (2014). Taming uncertainty in Big Data: Evidence from social media in Urban Areas. Business & Information Systems Engineering, 6(5), pp. 279–88

[9] Ishwarappa, K. and Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48(1), pp. 319–324.

[10] Trupti, A. Kumbhare, and Santosh, V. Chobe, (2014). An Overview of Association Rule Mining Algorithms, International Journal of Computer Science and Information Technologies, Vol.5(1), pp. 927-930.

[11] Sudhir, M. Gorade, Ankit Deo and Pritesh Purohit, (2017). A Study of Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology. Vol. 4, Issue. 4, pp. 3112-3115.

[12] J. Han, M. Kamber and J. Pei, J (2010). Data Mining Concepts and Techniques (3rd ed.) University ofIllinois.Chapter 8, pp. 99-117.

[13] Duda RO, Hart PE, and Stork DG, (2000). Pattern classification, 2nd ed. New York: John Wiley & Sons.

[14] Rao, R. P. N., & Scherer, R. (2010). Statistical Pattern Recognition and Machine Learning in Brain-Computer Interfaces. In *Statistical Signal Processing for Neuroscience and Neurotechnology* (1 ed., pp. 335-368). Elsevier B.V.

[15] Auria, Laura and Moro, R. A., Support Vector Machines (SVM) as a Technique for Solvency Analysis (August 1, 2008). DIW Berlin Discussion Paper No. 811, Available atSSRN: https://ssrn.com/abstract=1424949.

[16]S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan and M. j. Rajabi, (2014). "Advantage and drawback of support vector machine functionality," *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pp. 63-65, doi: 10.1109/I4CT.2014.6914146.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). Classification and Regression Trees. Chapman & Hall, New York, NY.

[18] S. K. Murthy, S. Kasif, and S. Salzberg, (1994). A system for induction of oblique decision trees. J. Artif. Int. Res., 2(1):1–32.

[19] J. Quinlan, (1986). Induction of decision trees. Machine Learning, 1(1):81–106.

[20] J. Quinlan, (1993). Morgan Kaufmann,C4.5: Programs for Machine Learning.

[21]Mean Squared Error (MSE). [Online]. Available: https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php

[22] Nova, D., Estévez, P.A. (2014). A review of learning vector quantization classifiers. Neural Comput & Applic 25, 511–524,https://doi.org/10.1007/s00521-013-1535-3

[23] D. Nova and P. Estevez, (2013). "A Review of Learning Vector Quantization Classifiers," Neural Computing and Applications, vol. 25, pp. 511–524.

[24] A. Priyono, M. Ridwan, A. J. Alias, R. A. O. Rahmat, A. Hassan, and M. A. M. Ali, (2012). "Application of LVQ neural network in realtime adaptive traffic signal control," Jurnal Teknologi, vol. 42, no. 1, pp. 29–44.

[25] Y. Freund, (1995). "Boosting a weak learning algorithm by majority", Information and computation. 121(2):256–285.

[26] Y. Freund and R.E. Schapire, (1999). "A short introduction to boosting" Journal of Japanese Society for Artificial Intelligence, 14(5):771-780.

[27] Huh, Myung-Hoe, & Lee, Yonggoo. (2006). "LMS and LTS-type Alternatives to Classical Principal Component Analysis". Communications for Statistical Applications and Methods, 13 (2), 233–241. https://doi.org/10.5351/CKSS.2006.13.2.233

[28] R. Agrawal and R. Srikant., (March 1995)."Mining Sequential Patterns". In *Proc. of the 11th Int'l Conference on Data Engineering*, Taipei, Taiwan.

[29] Fournier-Viger, P., Lin, J.C.W., Kiran, R.U., Koh, Y.S., Thomas, R. (2017)."A survey of sequential pattern mining". Data Sci. Pattern Recogn. s1, 54–77.

[30] Thabet Slimani, and Amor Lazzez. (2013). "Sequential Mining: Patterns and Algorithms Analysis", International Journal of Computer and Electronics Research, Volume 2, Issue 5, pp 639-647.

[31] Mooney, C. H. & Roddick, J. F., (Feb 2013) "Sequential Pattern Mining — Approaches and Algorithms", ACM Computing Surveys, vol. 45, no. 2, pp. 1–39, DOI: 10.1145/2431211.2431218.

[32] Kum, H.-C., Chang, J. H., & Wang, W. (2006). "Sequential Pattern Mining in MultiDatabases via Multiple Alignment". Data Min. Knowl. Discov., 12(2-3), 151-180.

[33] S. Anitha Elavaras, (Jan 2011). "A Survey on Partitional Clustering Algorithm", International Journal of Enterprise Computing and Business Systems, Vol. 1 Issue 1.

[34] Kaufman, L., & Rousseeuw, P. J., (1990). "Finding groups in data: an introduction to cluster analysis." New York, Wiley.

[35] T. Soni Madhulatha. (April 2012). "An overview on Clustering Methods". IOSR Journal of Engineering., Vol. 2(4) pp: 719-725.

[36] Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In Proc. KDD.

[37] Dimitrios Papakyriakou, Dimitra Kottou and Ioannis Kostouros. (April 2018). "Benchmarking Raspberry Pi 2 Beowulf Cluster. International Journal of Computer Applications" 179(32):21-27.

[38] Dimitrios Papakyriakou. (August 2019). "Benchmarking Raspberry Pi 2 Hadoop Cluster". International Journal of Computer Applications 178(42):37-47.