DATA MINING TECHNIQUES

Abstract

When referring to a group of methodologies, algorithms, and tools used to draw significant patterns, connections, and insights from huge datasets, we are referring to data mining techniques. These methods are essential for finding concealed information and for developing sensible conclusions across a variety of fields. Structured, semi-structured, and unstructured data are analyzed through data mining to produce useful information that can shape company strategy, advance scientific research, might aid to arrive at decisions -making. The preprocessing of data, which entails cleaning, converting, and combining data from many sources, is one of the fundamental data mining's elements. This ensures that the information is prepared for analysis. Several data mining techniques are used on the data after preprocessing to find patterns and relationships. Classification, grouping, and association rules are some popular data mining techniques. anomaly time series analysis, text mining, and detection. While clustering brings together related data points, classification aids in grouping data into preset classes. Regression theory examines the link connecting variables that are dependent and independent, whereas the mining of association rules uncovers intriguing associations between variables. Unusual patterns are found using anomaly detection, and text analysis uncovers essential details from textual data. The main goal of time series analysis is to examine and predict time-dependent data. The use of data mining techniques is common in many industries. They are employed in healthcare for disease prognosis. The study describes algorithmic discussion of the dataset for the disease acquired from UCI, on line repository of large datasets. The Best results are achieved by using Tanagra tool. Tanagra

Authors

Dr.I.Parvin Begum
Assistant Professor
Department of Computer Applications
B.S.Abdur Rahman Crescent Institute of Science
& Technology
Chennai,Tamil Nadu, India.
parvin@crescent.education

D.Nasreen Banu
Assistant Professor
Department of Computer Applications
B.S.Abdur Rahman Crescent Institute of Science
& Technology
Chennai,Tamil Nadu, India.
nasreen@crescent.education

is data mining matching set. The accuracy is calculate based on addition of true positive and true negative followed by the division of all possibilities. The algorithms are very necessary for intend an automatic classification tools. With help of automatic design tools to reduce a wait in line at the experts..

In conclusion, techniques for data mining provide strong tools to extract insightful knowledge from data, enabling making choices and fostering innovation across a range of industries. Companies may achieve a competitive advantage, boost operational effectiveness, and unleash the hidden value in their data assets by properly utilizing such techniques.

Keywords: Support Vector Machine clustering,k-Nearest Neighbours, classification, and logistic regression.

I. INTRODUCTION

The term "data mining approaches" refers to a group of procedures, algorithms, and tools that are used to extract significant knowledge and structures using huge databases. It involves analyzing data from several sources, such as databases, websites, social media, and sensors, in order to identify unnoticed trends, correlations, and insights that might be helpful for decision-making and predictive modeling. Data mining techniques are widely used in academic studies and industrialization across a variety of areas, including finance and business, healthcare, and medicine. Utilizing computers (IT) is a perk of modern science. Each business in the world currently uses it to offer soft, fast, and optimum working conditions. The mining of data is one of the technology business's fastest-growing industries. A formerly insignificant area in computingIt quickly grew into a pasture for only one person. Another of data mining's biggest advantages can be seen in the variety of techniques and approaches it provides, all of which may be applied to handle a variety of complexity levels. One of the main goals of countries when taking into account the fact that data mining is a routine task to be carried out on a huge amount of sets is to create a complete information transportation, data-mart, and decision-support concentrate of the general population, bounded by professionals from such industries as access to the internet, medical care, insurance, and public transportation.[1][2][3][4].

1. **Objective:** Computation and interpretation are probably going to be the data mining's most essential tasks. In order to forecast unnamed demands or standards of prior variables of attention, calculation is done using an assortment of elements or sections in the data set. The previous hand's explanation placed emphasis on decision patterns related to the evidence with the aim of being understandable by people. As a result, it is possible to find data analysis effectiveness focused on just one of two categories:Analytical data mining, which generates an outline of the structure as it is described by the specified data collection, or
   - Information mining, essentially is based on the available data set, provides particular, nontrivial in sequence results.
   - On the analytical end of the range, data mining's goal is to produce a representation that can be utilized to perform categorization, calculation, evaluation, and other antecedently comparable duties, expressed as an executable sign. The goal is to broaden the responsiveness of the analyzed method with exposure patterns and connection to substantial data sets at the prior, expressive end of the range.
   - On the relative relevance of forecast and tale may show a significant discrepancy for demanding information mining applications.To accomplish the forecast and report objectives, data-mining techniques are used.

2. **Which Data Types Can Be Mind**
   - Data warehouses
   - Data Warehouses
   - Transactional Data
   - Other Kinds of Data

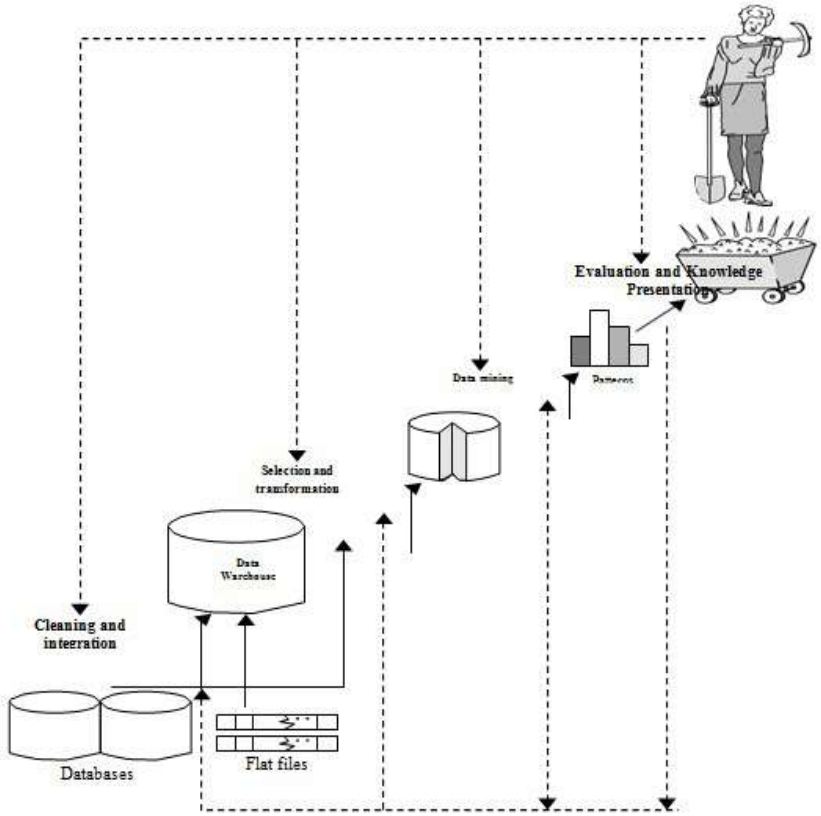


Figure 1: Data mining as a step in the process of knowledge discovery.

Figure 1 depicts the knowledge discovery process as an iterative series of the following steps:

- Cleaning down information (to get rid of noise and incorrect data)
- Data fusion (the combining of data from many sources).
- The choice of data (the process of retrieving data from a database that are pertinent to the mathematical task)
- Data conversion (transformation and consolidation of data into forms suitable for mining through summary or aggregate processes)4
- Data extraction (a crucial procedure that employs clever techniques to extract data structures)
- Evaluation of structures (to find the patterns that indicate information that are actually intriguing based on metrics of interestingness.
- Knowledge appearance (where consumers are presented with knowledge gathered using graphics as well as information presentation approaches)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base**.**

II. TYPES OF DATA IN DATAMINING

1. **Smooth Files:** The majority of data sources used by methods for data mining, especially during the research level, are clean files. Simple files are archives of information in text or binary format that have been organized in a way that the information extraction algorithm recognizes as being helpful. These documents contain data that can be applied to communication, analysis of time frames, systematic capacity, etc.

2. **B. Relational Database:** A relational database, to put it simply, is made up of a collection of tables that also contain standards of individual characteristics, or standards of attributes originating from particular connections. Rows and columns make up a chart, with rows indicating tuples and columns signifying qualities. A hierarchical table's combination is recognized with a set of quality values rather than a single key and corresponds to both an entity and/or a link connecting substance [6] [7].

   Structured Query Language is the most popular type of query used for relational records since it enables both the recovery and use of the data both in the rows and columns and when generating accumulated functions. Like standard, addition, minimum, maximum, and total up. Using relational databases, data mining techniques because they comprehend how to take advantage of the features built within SQL databases, methods for data mining using these databases are more customizable when compared to those specifically written for smooth files. Data mining extends beyond what structured query language might offer, such as identifying, contrasting, detecting deviation, etc., even though it can support SQL for data gathering, manipulation, and aggregation [5].

3. **Data Warehouses**: A depository for data is an information warehouse of data made up of information from several sources, frequently multiple sources, and is intended to be utilized as a comprehensive integrative presentation above and above the same. Data evaluation from multiple places under the same wrapping is possible with the aid of a warehouse of data.

Figure 2: pan and zoom until you locate an active hurricane.

The Atlantic and Northeast Pacific hurricane seasons last from June through November, the South Pacific and Indian Ocean seasons last from November through April, and the Northwest Pacific hurricane season lasts from April through December.

The orange points show the observed track of the tropical storm. The black points and text are the forecasted track of the tropical storm. The gray area indicates the possible margin of error in the forecast.

4. **Transaction Databases**: In organization's dataset is an amalgamation of statements instead of communications, each one featuring a unique identifier, a set of contents, and an occasion trample. Business records may also provide expressive data regarding the material related to them.

5. **Multimedia Databases**: video content, illustrations, sounds, including written text are all included in multimedia databases. They can be stored in a finite amount of object-relational or object-oriented databases, or they can be stored essentially on a folder system. Data mining is made more challenging for multimedia because of its higher dimensionality. Computing graphical representation, graphics programming, picture analysis, and conventional sentence dealing out gets closer may be used while mining data from audiovisual archives.

6. **Common Database**: Standard files are collections of databases that store environmental information in chronological order, such as maps, and global or regional place of residence. The novel challenge posed by such spatial information systems to data mining methods.

7. **Time-Series Database**:Time-series databases contain time-related data like sales and advertising statistics as well as other tracked activity. These databases typically feature a continuous inflow of the latest information, which occasionally necessitates the need for a taxing real-time analysis. In such databases, data mining typically entails learning about correlations and patterns relating the evolutions of odd variables, as well as forecasting trends and timelines for the variables in question.

III. DATA MINING ARCHITECTURE

Three distinct components make up the information mining architecture. The level of information that has been gathered is closelyfollowed by the level of records and metadata, the application layer that performs data management and algorithms, and the interfacelevel for treatment, input configuration of parameters, and results visualization.

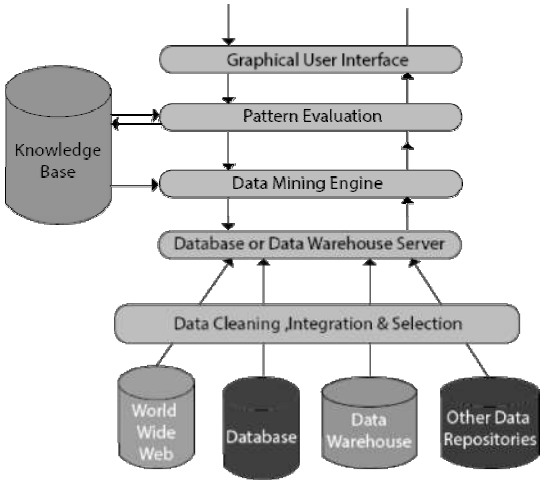


Figure 3: Three-level architecture for Data Mining.

Three distinct components make up the information mining architecture. The level of information that has been gathered is closely followed by the level of records and metadata, the application layer that performs data management and algorithms, and the interface level for treatment, input configuration of parameters, and results visualization.

Data Source: The Internet's (WWW), information databases, data warehouses, file types, and additional records serve as the actual sources of data. For the use of data mining to be efficient, a significant amount of previous data is essential. Data is generally kept by businesses in databases or data warehouses. One or more databases, text files, spreadsheets, or other data sources may be included in a warehouse of data. Even spreadsheets and simple text files may at times involve information. The internet, sometimes known as the World Wide Web, is a further significant repository of data.

Different Processes: The data must be purified, integrated, and approved before it is transmitted to the system or information warehouse server. The data is unable to be utilized immediately for the data mining process since it originates from numerous sources and in different forms, and the data may not be accurate and full. Therefore, the initial data needs to be consolidated and cleaned. Several sources of data will gather more data than is necessary; just the relevant data must be selected and sent to the server. These processes are not as simple as we may assume. As an aspect of choosing, insertion, and cleansing, the data may be analyzed using a variety of approaches.

In other words, the basis of our data mining architecture is the mining of data. It includes programs and instruments for extracting information and insights through data gathered from diverse data sources and kept in the data warehousing. The initial thing phase is the document level, where information is added to and kept alongside data about data. The next level is known as Data Mining Application, and it is used everywhere that knowledge is processed step by step and results are compiled into records. The Front-End threshold, which is represented by the third level in Figure 1, provides a simple to place restrictions on where Data Mining Applications can operate and to showcase the findings in an understandable manner.

IV. DATA MINING TECHNIQUES

Two sorts of information extraction methods

1. **Supervised Data Mining Techniques:** It includes the use for information that has been labeled and in which the goal variable or result is known. These methods use the offered labeled examples to learn how to anticipate or categorize brand-new, unforeseen data situations. A set of characteristics of the input and their identifying target values are used to train the algorithms.

2. **Unsupervised Data Mining Techniques:**Once the target variable or outcome is uncertain, unsupervised data mining techniques are used on unlabeled data. Despite any prior understanding or direction, these approaches seek to identify connections, trends, or classifications in the data.

Table 1: Supervised Vs Unsupervised Data Mining Techniques

Supervised	Unsupervised
Decision Trees	Clustering
Naive Bayes	Association Rule Mining
Support Vector Machines (SVM)	Dimensionality Reduction
Random Forest	Anomaly Detection
Gradient Boosting	Self-Organizing Maps (SOM)

V. SUPERVISED DATA MINING TECHNIQUES

1. **Decision Tree:**It is visualizations of how decisions are made and potential outcomes. To create predictions or categorise the data, they are created by recursively splitting the data based on several qualities. Both category and quantitative information can be handled by decision forests, and they are simple to use.

2. **Naïve Bayes Classifier:** Each of the characteristics are presumed to be conditionally independent of one another in this classification process, which is based on Bayes' theorem. Given the attribute values, it estimates the probability of each class and awards the one with its greatest possibility.

3. **Support Vector Machine(SVM):** It is a sorting by mathematical method that establishes the best hyper plane for classifying the information in question. It seeks to increase the space between the classes in order to strengthen generalization to new data. Using kernel functions, SVM can solve difficulties with classification that are both linear and non-linear.

   SVM handles classification tasks by minimizing classification errors while making the greatest use of the boundaries straightening out together module. The two main characteristics of SVM are kernel purpose, which starts non-linearity in the suggestion gap despite specifically requiring a non-linear procedure, and simplification speculation, which results in a systematic approach to choose a hypothesis. Health assessment SVM proved quite useful in helping human specialists comprehend how the algorithm decides.

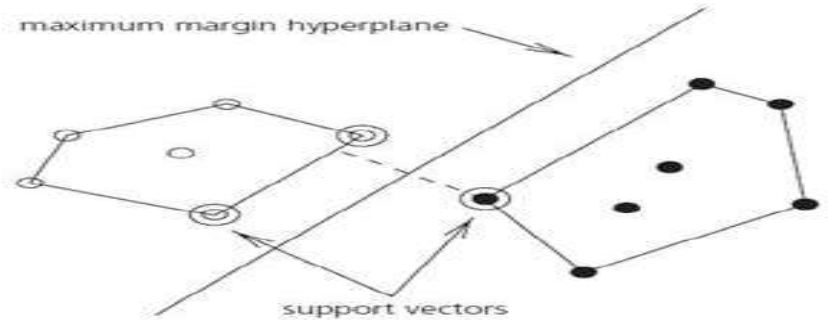


Figure 4: Topology of support vector machine.

Seen in figure 2, SVM does categorization task by making the most of the boundary separate together component while lowering its classification fault values.

4. **Random Forest:** An ensemble learning method called To create forecasts, Random Forest integrates several kinds of decision trees.. A selection that is random of the information and attributes is used to build each tree. The overwhelming vote or the average of the individual trees' predictions serves as the starting point for the final forecast.

5. **Gradient Boosting:** It is an approach to ensemble learning that sequentially combines weak classifiers, usually decision trees. Each subsequent classifier is taught to fix the errors that occurred with the preceding one. Due to their excellent predictive performance, gradient-boosted algorithms like AdaBoost, where the XGBoost and LightGBM as have grown increasingly apparent.

VI. UNSUPERVISED DATA MINING TECHNIQUES

1. **Classification :** Understanding issues, models that over fit the data as a result of the optimization procedures used for barrier range, and geometric actions to choose the best model are all problems with traditional neural network technology approaches. Due to their numerous remarkable characteristics and excellent observed performance, SVMs have been gaining prominence.

2. **Association Rule Mining:** This method is employed to identify intriguing connections or associations between various elements or characteristics in a collection of data. It facilitates the discovery of patterns like "if A, subsequently B" or "the inhabitants who acquire X also buy Y." The algorithms the Apriori and FP-Growth are frequently employed in association rule mining.

3. **Dimensionality Reduction :** Techniques for diminishing the quantity of variables in a dataset while keeping its important details are known as reductions in dimensionality approaches. Popular dimensionality reduction Principal component analysis (PCA) and t-SNE (t-Distributed Statistical Neighbour Embedding) are two techniques.

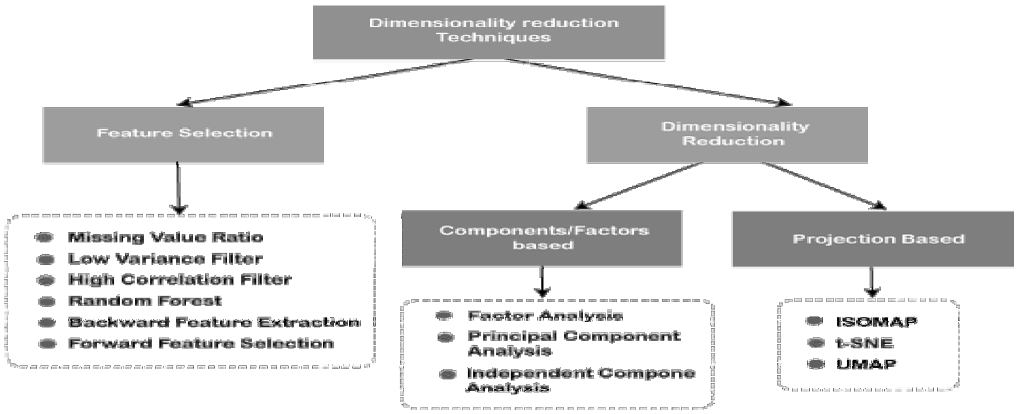


Figure 4: Dimensionality Reduction Techniques

- **Advantages of using dimension reduction**

  - The dimensionality reduction method has the following benefits when applied to the given dataset:
  - By reducing the dimensionality of the features, the amount of storage space required for the collecting of data is reduced.
  - Shorter computation training periods.
  - By taking care of the a degree of multi it eliminates unnecessary characteristics (if any are present).

Approaches of Dimension Reduction

- **The dimension reduction approach can be used in one of the two ways listed below::**
  - Feature Selection
  - Feature Extraction

A certain quantity of crucial qualities must be selected from a group of data in order to build a representation with high accuracy, and any irrelevant details must be left out. The choice of features is the method in question. Take it differently it is a technique for choosing the finest qualities from the information presented. Three methods are used for the feature selection:

- Filters Methods
- Wrappers Methods
- Embedded Methods

Extracting features is the process of reducing the number of dimensions in a multidimensional space. This strategy is helpful especially if we desire to keep all of the information while interpreting it with fewer resources.

- Principal Component Analysis
- Linear Discriminant Analysis
- Kernel PCA
- Quadratic Discriminant Analysis

4. **Anomaly Detection:** Finding outliers or odd patterns in a dataset that significantly vary from expected behavior is the goal of anomaly detection. The above technique proves useful for quality assurance, fraud detection, and network intrusion detection. For the purpose of detecting anomalies, techniques like as statistical modeling, clustering, and support vector algorithms can be used.

5. **AnomSelf-Organizing Maps (SOM):** SOM is a grouping and visualization method for unsupervised neural networks. The topological connections between the data points are preserved while mapping data with high dimensions onto a 2-dimensional grid. Text Mining: The goal of text mining techniques is to glean knowledge and insights from textual material. It entails activities like named entity recognition, topic modeling, document categorization, and sentiment analysis. Text mining frequently uses Natural Language Processing (NLP) algorithms and methods.

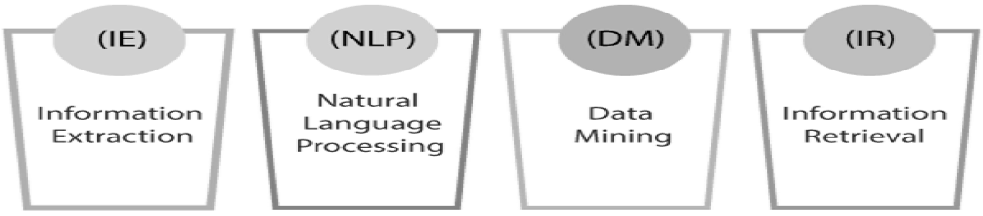


Figure 5: Areas of Text miming.

There are additionally different algorithms and strategies that each have their own advantages and drawbacks. The precise data mining task at hand as well as the features of the dataset will impact the technique to use.

Information extraction is the process of automatically extracting structured data from an unregulated source, such as entities, relationships between entities, and attributes defining entities.

- **Natural Language Processing**: Software for computers can comprehend spoken human language exactly. Artificial intelligence (AI) is basically made up of NLP. Because computers typically demand humans to "Speak" to them using a language used for programming that is precise, unambiguous, and extraordinarily structured, developing the NLP application is challenging. Since human speech tends to be not authentic, it can be influenced by a wide range of intricate factors, such as terminology, context of society, and local accents.

- **Data Mining:** Data mining is the procedure of discovering essential details as well as hidden patterns from huge data collections. Businesses can use data mining techniques to predict future habits and patterns in order to make better data-driven decisions. Many corporate problems that have previously taken too much time to solve can now be solved with the help of data mining tools.

- **Information Retrieval:** Retrieved data is the process of obtaining relevant information from information that is stored in our IT infrastructure. Alternatively, we can use search engines found on domains like e-commerce sites or other websites as an analogy to approach information collection.

VII. DIFFERENT TYPES OF CLASSIFICATION TECHNIQUES

In data mining, data into determined categories or sections based on characteristics using the classification approach and characteristics of the data. Here are some commonly used classification techniques technique used in data extraction called classification divides data into preset classes or categories depending on the properties and characteristics of the data. Here are a couple of instances of commonly utilized classification methods.

1. **Decision Trees*:*** Decision chains are schematic representations of a series of decisions and conceivable results. They divide the data into segments depending on several attributes and then create a classification model that simulates a tree. Decision trees can handle the level and numerical data and are simple for understanding.

2. **k- Nearest Neighbors*:*** k-NN is a latent learning technique that categorizes each data point in the feature space according to the vast majority of grouping of its k nearest neighbors. The degree of smoothing or sensitivity to local differences depends on the choice of k. There are two different closest-by algorithms. The first one was the K-nearest Neighbor algorithm, which diagnoses entities primarily on the nearby training examples in the feature space. Prototype Nearest Neighbors Algorithm was the next. It is one of the simplest artificial intelligence algorithms.

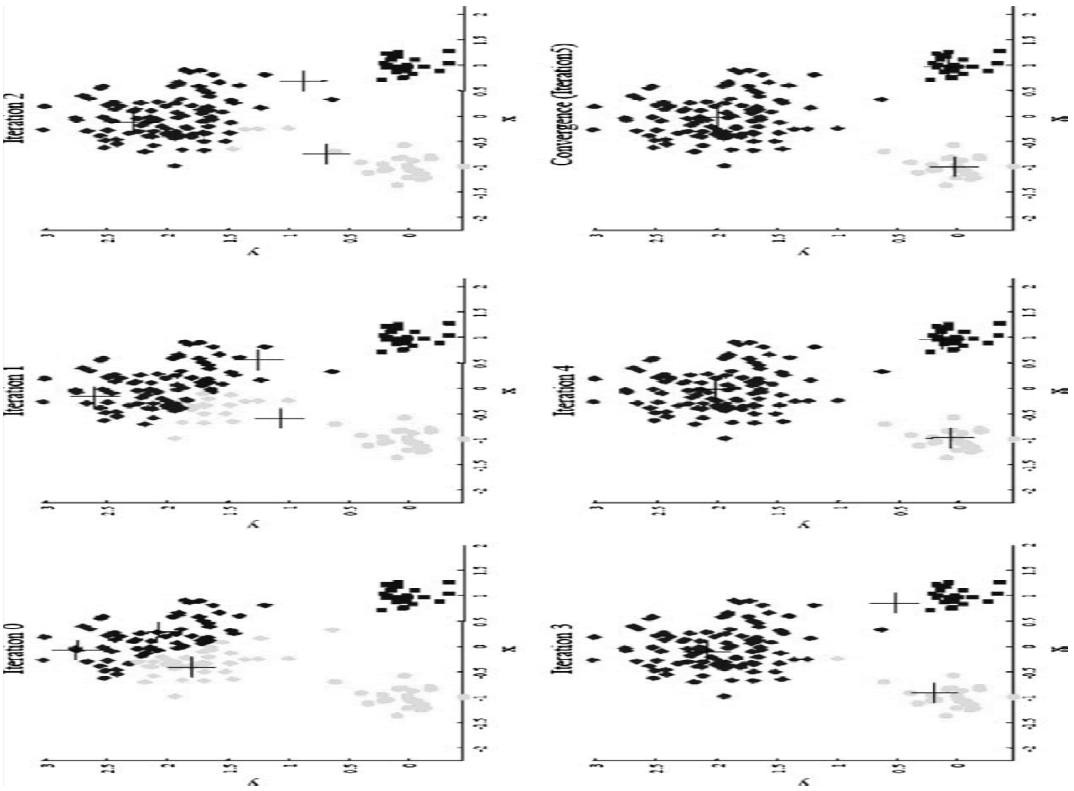


Figure 6: Changes in cluster representative locations.

he procedure works with the collection The ith observation point is represented by xi kd in the equation D = xi | i = 1... N. The method starts by selecting k spots. from kd to serve as one of the k cluster representatives, or "centroids". Techniques for selecting these preliminary records include picking them at random from the available data and grouping the records separated so as to avoid upsetting the overall standing for the records k times.The algorithm then executes the two cultivate union steps:

- **Step 1:**Transfer of information. Each piece of data is matched to its nearest centroid, with ties being broken at random. This divides all of the records into grades.

- **Step 2:** Moving the word "means" around. The exact center of each and every information point allocated in the direction of each group's representative is moved.

The replacement is the perspective (partisan suggest) of the data partitions if the data points have a probability measure.While restrictions on clusters are initially linear during the known high-dimensional gap, they can grow non-linear following a proposed reversal to the gap, allowing for crucial part k-means to agreement by more challenging clusters.

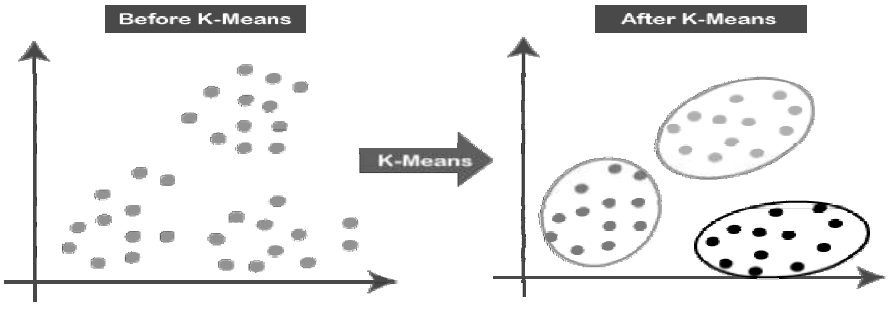
3. **Neural Networks:** Neural networks, also called neural networks, are computational approximations of the brain's organisation and functionality. They are constructed from up of connected neurons grouped in segments. organized in layers. In classification tasks, neural networks learn to assign weights to the connections between neurons to classify data into different classes.

4. **Logistic Regression**: A form of statistics Logistic regression is an approach for issues related to binary classification. The relationship among the factor that is dependent and each of the independent variables is represented employing a logistic function.. One-vs-all or sotmax regression are two strategies that can be used to expand the use of logistic regression analysis when dealing with a variety of class conditions.

VIII. DIFFERENT TYPES OF CLUSTERING TECHNIQUES

According to their innate similarities or patterns, grouping is a technique utilised for putting together similar information elements. These are only a few commonly used clustering techniques:

1. **K-means Clustering:** K-means, an effective clustering methods, breaks down divides data into k clusters, where k is pre-determined. It modifies the center lines on the basis of the average of the allocated points after iteratively assigned data points to the closest cluster centroid. K-means is effective for huge datasets and is computationally efficient. K indicates In the field of data science or machine learning, grouping is a type of unsupervised neural network process that is used for solving clustering problems. determines the optimal number for K centre points or centroids using an iterative method. The most similar k-center is matched with every data point. The information comprises values that are adjacent to a confident k-center gather collectively to form a cluster. Because of this, each cluster is distinctive but comprises information points with some common features. Interests.



The following stages illustrate how the K-Means algorithm functions:

- Step 1: Pick K to get the number of clusters.

- Step 2: Select K centroids or places at random. The supplied dataset might not be the issue.
- Step 3: Assign each information point to its nearest centroid, resulting in the construction of the predetermined K clusters.
- Step 4: Calculate the variance and move the centroid of each cluster.Repeat the third step to reassign each data point to the new average of each group in step 5.
- Step 6: if a transfer occurs, go to step 4; otherwise, go to Complete.
- The finished model is step seven.

2. **Hierarchical Clustering:** An unsupervised learning process known as "hierarchical clustering" establishes novel groups based on existing ones. It operates by clustering data into a tree structure. statistics using hierarchical clustering that regard each data point as a separate cluster. The pointer describes another set of clusters, each of which differs from the others while comprising identical entities.There are two types of hierarchical clustering

   - **T-Clustering:** One of the most prevalent hierarchical clustering techniques for assembling related objects into clusters is agglomerative clustering. Agglomerative nesting, or AGNES, is another name for agglomerative clustering. Agglomerative clustering groups the information in a bottom-up manner at each stage, with each data point acting as its own grouping. Each data object starts out in its cluster. The clusters are joined with other clusters at each iteration to generate one cluster. Analyze how closely individuals resemble other people and every cluster. Locate the relationship matrix.

   - **Divisive Clustering:** All of the data points are handled as distinctive clustered in Divisive hierarchical grouping, and in every repetition, the knowledge values that do not appear similar are eliminated from the cluster. The divided data points are handled individually as clusters. N clusters are everything that are left in the end.

3. **Self-Organizing Maps(SOM):** SOM, also referred to as Kohonen maps, is a form of unsupervised training for categorizing data using a neural network. It converts high-dimensional data into a grid of neurons with lesser parameters. SOM may be used to visualize and explore high-dimensional datasets since it maintains the topological structure of the data.

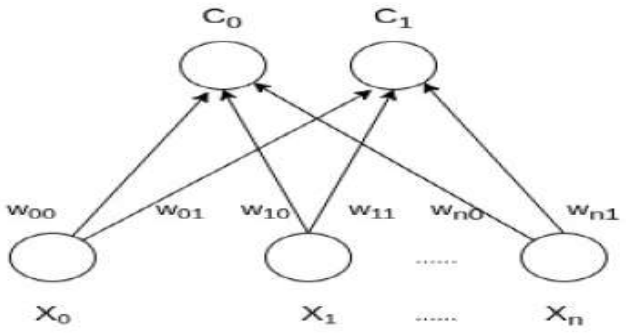


Figure 7: Architecture of the Self Organizing Map with pair clusters and n input features

- **Training**

➢ Step 1: Set the the weights in Step 1 by assuming a random value.
Set the learning rate to zero..
➢ Step 2: Calculate the Euclidean distance squared.
- $D(j) = \Sigma (wij – xi)^2$ where i=1 to n and j=1 to m

➢ Step 3: Determine index J, which will be regarded as the winning index, when D(j) is minimum..
➢ Step 4: Determine the new weight for all i and for each j in a particular area around j..
- $wij(new)=wij(old) + \alpha[xi – wij(old)]$

➢ Step 5: Update the learning rule by using :
- $\alpha(t+1) = 0.5 * t$

➢ Step 6: apply the formula to test the Stopping Condition.

IX. CASE STUDY

Data mining strategies are usually divided into several related sections. The first is a legitimate category, whereas the second is not. Since input and output qualities persist, it is more likely that supervision expertise will be applied if there are a set that includes input data and production data.

1. **Data Mining in the Medicine:** The global optimization strategy to grouping and indicate how gathering can be utilized to tackle the manage data classification problem. The objective functions in this problem have a significant amount of constrained minimizers and are both mutually hard and nonconvex. Global optimization strategies frequently fall short of resolving such issues due to an enormous amount of data and the complexity of the problem. The health care field currently produces a significant amount of complex data about patients, hospital resources, disease detection, sick person reports, pharmaceuticals, etc. Robert Detrano's experiment generated reliable results for classification with a discriminate function produced by logistic regression of about 77%.

2. **Healthcare Resource Management:** Notwithstanding the enormous volume of data in the medical science sector, most of it is disappointingly not regarded as newly discovered or undiscovered. Unknown prototypes can be found during information using sophisticated data mining techniques. Residential reproductions beginning these approaches will be really helpful for wellness assessments practitioners to acquire worthwhile results.Risk-based evaluation of rest home plans using logistic regression models. The character of youngsters who are present in the emergency situation is calculated by network systems. Discovered the possibility of hospital humanity in cancer patients with non-fatal illnesses. Following a predetermined process of knowledge discovery is crucial when using wellness assessments mining of information because of the difficulty requirement close.

They work with genuine information and data to support medical results and to be acknowledged as proof while EBM has been around for millennia.The use of data mining for community health and drug development is a logically immature field. Knowledge Acquisition Health records include a functional way for extracting information to support decisions.

They cause confusion about the laws that can be applied to information retrieval in the area. While some creators define information mining as the invention of data acquisition, others define it as the application of numerical techniques.

3. **History of Diabetes:** Diabetic is certainly not a recently discovered illness; it has existed since the beginning of human civilization as we know it, which was around 1552 B.C. Since this particular outbreak, a number of Greek and French health care providers have worked to raise our awareness of the nature of disease and the organs that cause it. A French doctor had established the connection connecting diabetes and the preparation of a personality diet chart in the 1870s, in addition to the advice to watch what you eat. Both of the name and the brief description of the properties are included.

Te three states and one the Union Territory, or roughly 18.1% of the population, submitted data for the Indian Council of Medical Research-Indian Diabetes (ICMR-INDIAB) study.Extrapolating from these four the facilities, it may be deduced that 62.4 million Indians have diabetes and that 77.2 million of their population has prediabetes. It took into account anthropometric characteristics such body weight, Mass Index, height, and waist circumference. It also assessed each participant's cholesterol and fasting blood sugar levels after a glucose load.Pre-diabetes affects fasting glucose levels and/or sugar level acceptance to a percentage of 8.3%, 12.8%, 8.1 percent, and 14.6%, respectively. India has 62.4 million people with pre-diabetes and another 77.2 million who may develop diabetes in their remaining 19 years.

Table 2: Increasing occurrence of Diabetes: India

Year's	Number Of People effected (In Millions)
1995	124.7
2000	153.9
2025	299.1

4. **Data Source & Experimental Results:** To evaluate these data mining classification use Pima Indian Diabetes Disease information center. The information has nineentities and 768 instances.

Table 3 : Entities of diabetes dataset.

No	Entity	Depiction
I	Pregnancy	How many times expectant
II	Plasma	Plasma glucose concentration a 2 hours in an oral glucose test
III	Pressure	Pressure (mm)

No	Entity	Depiction
I	Pregnancy	How many times expectant
IV	Covering	Triceps covering wrinkle width (mm)
V	Insulin	2-Hour insulin
VI	Mass	Body mass index weight in kg
VII	Pedi	Diabetes family background function
VIII	Year	Years(Age)
IX	Set	set changeable (0 or 1)

In the table 3 attributes are exacting, all patients now are females are twenty one years old. If the two hours post load Plasma glucose was as a minimum 200 mg/dl.

Confusion Matrix is a matrix showing the predicted and actual classifications. Suppose m attributes then confusion matrix is of size m x m. In this experiment I had two types of classification.

$$\text{i.e., accuracy} = (TP + TN) / (TP + TN + FP + FN).$$

The recall measure of the predictive performance is the percentage of errors correctly predicted out of all the errors that really occur;

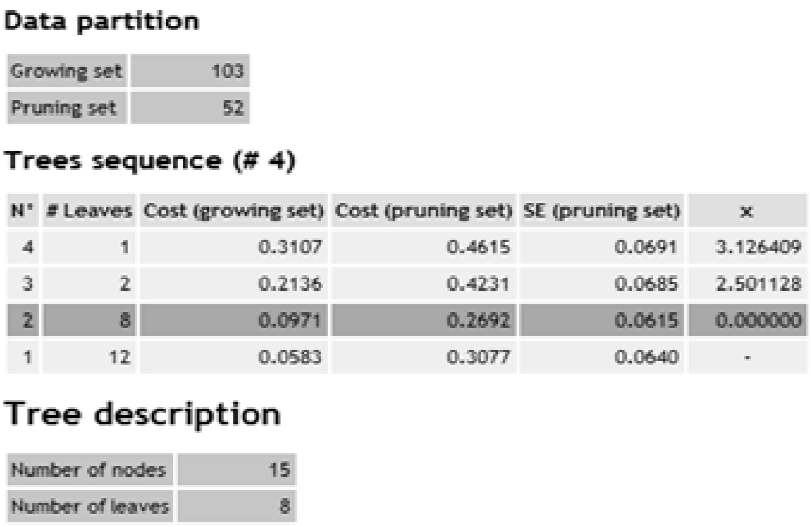
$$\text{i.e., recall} = TP / (TP + FN).$$

Precision is the percentage of the actual errors among all the encounters are divided while fault value via the C4.5.

Precision = TP / (TP + FP).
The F-measure metric is the mean of recollect along with exactness metrics and is computed as:
F-measure = 2 x (recall x precision) / (recall + precision)

Table 4: Confusion Matrix of Cross Validation for CS-CRT.

Computing Time =531 Millisecond			
Confusion Matrix			
	Positive	**Negative**	**Sum**
Positive	**8**	**46**	**54**
Negative	**8**	**88**	**96**
Sum	**16**	**134**	**150**



Data partition

Growing set	103
Pruning set	52

Trees sequence (# 4)

N°	# Leaves	Cost (growing set)	Cost (pruning set)	SE (pruning set)	x
4	1	0.3107	0.4615	0.0691	3.126409
3	2	0.2136	0.4231	0.0685	2.501128
2	8	0.0971	0.2692	0.0615	0.000000
1	12	0.0583	0.3077	0.0640	-

Tree description

Number of nodes	15
Number of leaves	8

Figure 7: Screen shot for Tree Sequence for CS-CRT in HTML format.

Number of leaves, charge (increasing position), charge (reducing position), and SE (reducing location) values are depicted in the figure 5 screen image.The amount of tree leaves influences the highest price rate values for pruning.

There are many more, depending on the details of the problem and the features of the dataset. The technique selected relies on the type of data, the objectives of the research, and the results generated.

X.CONCLUSION AND SUMMARY

Data mining algorithms can be trained from past examples in clinical data and model the frequent times non-linear relationships between the independent and dependent variables. The consequential model represents formal knowledge, which can often make available a good analytic judgment. Classification is the generally used technique in medical data mining. . The accuracy is calculate based on addition of true positive and true negative followed by the division of all possibilities. The algorithms are very necessary for intend an automatic classification tools. With help of automatic design tools to reduce a wait in line at the experts.

REFERENCES

[1] Morgan Kaufmann, “Data Mining Concepts and Techniques”,Third edition, Morgan Kaufmann Publishers an imprint of Elsevier.
[2] YoshimasaTsuruokaTsujii and Sophia Ananiadou “FACTA: a text mining engine for finding associated biomedical concepts” Advanced Access Publication September 4, 2008.
[3] Gordon S. Linoff · Michael J. A. Berry “Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management”,Dec 2018 · Gildan Media · Narrated by Steve Menasche
[4] MaheshT.R. Suresh.M.B, VinayaBabu.,"Text Mining: Advancements, Challenges and Future Directions”, International Journal of Reviews in Computing, IJRIC, 2009-2010.
[5] Honey Mahgoub, “Mining Association Rules from unstructured Documents”, International journal of mathematics and computerscience1:4 2005.
[6] Ramakrishna, Gehrkev, “Database Management Systems”, International Edition, TMH, p-929.
[7] IEEE CS International conference on Data Mining.

