IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data Mining and Knowledge Discovery Process

Miss. Sneha Vilas Kotawadekar
Principal (I/C)
Maharshi Karve Stree Shikshan Samstha's
College of Computer Applications for Women, Ratnagiri – 415 629 M.S. (India)
(Affiliated to SNDT University, Mumbai)

ABSTRACT

Data is a very powerful weapon which can be used by almost all the profitable as well as nonprofitable organizations worldwide to cherish in the competition of their existence. For this the generated data has to bestow knowledge for its proper convention. Data mining is the computational process of analyzing data from different perspectives, dimensions, angles and summarizing it into meaningful information. Due to this factor, data mining has become a very important and guaranteed branch of engineering affecting human life in various domains directly or indirectly. Knowledge drawn from the large data set through Knowledge Discovery Process also called as KDD helps the user to conclude with very important discoveries which were hidden under large amount of data otherwise. The purpose of this paper is to understand data mining and Knowledge Discovery Process and its importance for an organization handling large amount of data every day. Also the review the challenges faced by an organization while using data mining for knowledge discovery.

Keywords: *Data mining ,KDD ,database,CRM,IDS*

1. INTRODUCTION

Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets.[1] Data mining is actually integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, etc. This is because this techniques help in gathering more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions. Data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, time-series Databases, World Wide Web.

Data mining as a whole process has three main phases:

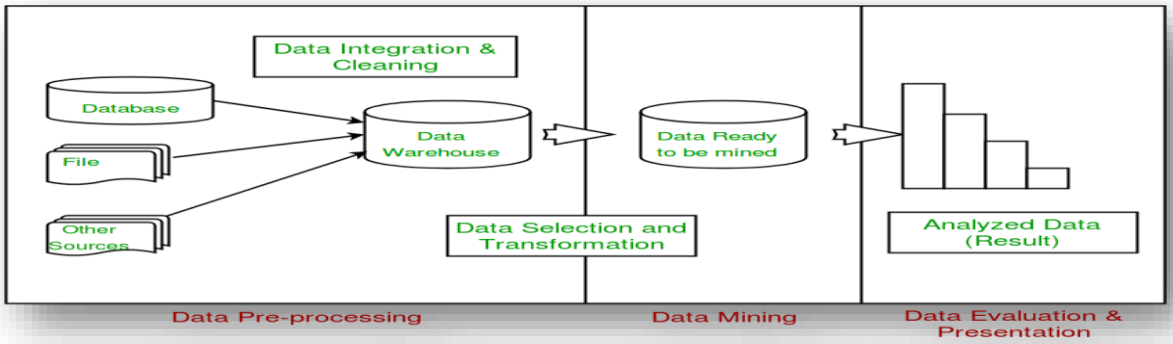

Fig A: Process of Data Mining.

1. Data Pre-processing – Data cleaning, integration, selection, and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results. [2]

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.[3]

The two terms KDD and Data Mining are used interchangeably, but they refer to two related yet slightly different concepts. KDD is the overall process of extracting knowledge from data while Data Mining is a step inside the KDD process; it actually deals with identifying patterns in data. Data Mining is only the application of a specific algorithm based on the overall goal of the KDD process.

2. **KNOWLEDGE DISCOVERY (KDD) PROCESS:**

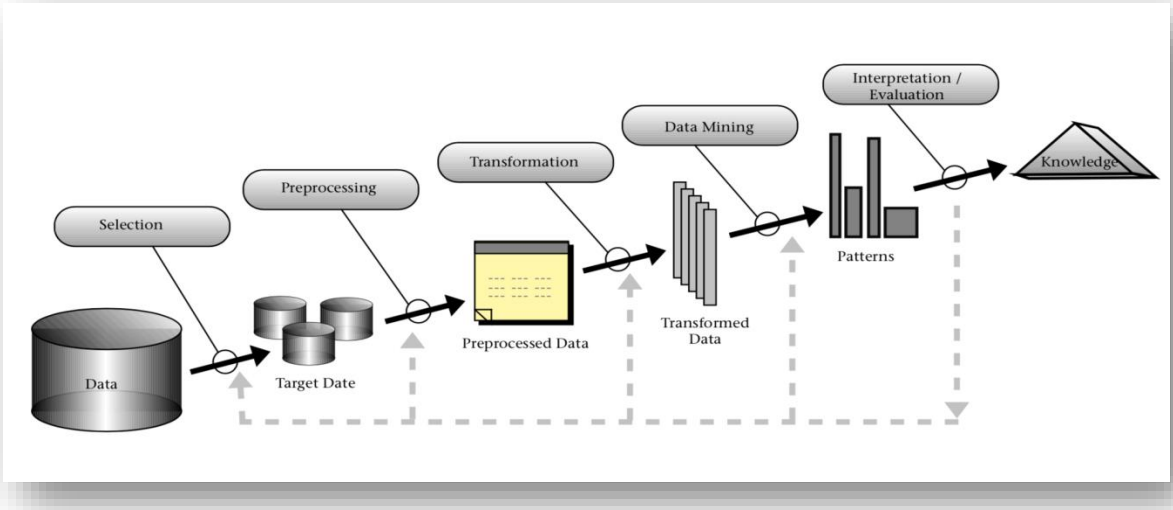

Fig B: Knowledge discovery from data (KDD) Process

Different steps of Knowledge Discovery in Databases are:

Understanding: The first step understands the requirements. We need to have a clear understanding about the application domain and your objectives, whether it is to improve your sales, predict stock market etc. It should also know whether you are going to describe your data or predict information.

Selection of data set: Data mining is done on your current or past records. Thus, you should select a data set or subset of data, in other words data samples, on which you need to perform data analysis and get useful knowledge. We should have enough quantity of data to perform data mining.

A. Data cleaning: Data cleaning is the step where noise and irrelevant data are removed from the large data set. This is a very important preprocessing step because your outcome would be dependent on the quality of selected data. As part of data cleaning, you might have to remove duplicate records, enter logically correct values for missing records, remove unnecessary data fields, standardize data format, and update data in a timely manner and so on.

B. Data transformation: With the help of dimensionality reduction or transformation methods, the number of effective variables is reduced and only useful features are selected to depict data more efficiently based on the goal of the task. In short, data is transformed into appropriate form making it ready for data mining step.

C. Selection of data mining task: Based on the objective of data mining, appropriate task is selected. Some common data mining tasks are classification, clustering, association rule discovery, sequential pattern discovery, and regression and deviation detection. We can choose any of these tasks based on whether we need to predict information or describe information.

D. Selection of data mining algorithm: Appropriate method(s) is to be selected for looking for patterns from the data. You need to decide the model and parameters that might be appropriate for the method popular data mining methods are decision trees and rules, relational learning models, example based methods etc.

E. Data mining: Data mining is the actual search for patterns from the data available using the selected data mining method.

F. Pattern evaluation: This is a post processing step in KDD which interprets mined patterns and relationships. If the pattern evaluated is not useful, then the process might again start from any of the previous steps, thus making KDD an iterative process. G. Knowledge consolidation: This is the final step in Knowledge Discovery in Databases (KDD). The knowledge discovered is consolidated and represented to the user in a simple and easy to understand format. Mostly, visualization techniques are being used to make users understand and interpret information. [4]

3. How does Data Mining work?

Data mining involves exploring and analyzing large blocks of information to pleat meaningful patterns and trends. It is used in a variety of ways, such as database marketing, credit risk management, fraud detection, spam Email filtering, or even to determine the sentiment or opinion of users. Data mining software analyses relationships and patterns in the stored transaction data. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials. Data classes are groups that share easily identifiable characteristics. This explains why they are also referred to as predetermined groups. In the context of a retail business, customers who have purchased a particular product constitute a data class. For example, Amazon.com customers who have purchased business books in the past constitute a class. Knowing the characteristics of the data class takes the guesswork out of "likelihood to buy" factor in sales promotion. The online retailer can use this grouping to develop marketing campaigns for business books and target customers in the group (and underlying subgroups). Depending grouping can significantly improve the efficiency of mass marketing.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, a sports shop that analyzed their data know that there is an 85% chance that a person buying new mountain bike will also buy a helmet, gloves and a water bottle. However, customers who come in requesting a helmet will probably not buy a bike, but they most likely will also buy gloves. This knowledge can assist the manager in ordering the correct stock and assist the sale personnel in suggesting add-on purchasing. Data clusters are similar to classes, but include additional attributes such as logical relationships. In the context of business applications, consumer preferences are often the most useful attributes. Consumer preferences can be used to understand market segments and customer loyalty. Accurate clustering can support cross selling. Again, using Amazon.com as an example, data clusters allow the retailer to identify what other products are purchased by customers who buy business books. Armed with this information, the retailer can

develop "product recommendations" as part of its customer relations management (CRM) programs. The ability to nurture leads efficiently is critical to sales.

- Associations: Data can be mined to identify associations. Data associations take clusters further. In the context of business application, associative data mining reveals buying patterns that would otherwise go unnoticed. For example, changes in buying habits induced by shifts in the economy require in-depth analysis for accurate characterization. A clear understanding of the economic shifts can be exploited for marketing purposes
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes. While analyzing past purchases is helpful, some experts believe that the true benefit of data mining is to anticipate customer purchases through predictive analytics. By building on historical data, sequential patterns allow projections to be developed. The projected industry trends are essential for forward-looking business planning and competitive intelligence.[5]

4. Applications Of Data Mining and KDD:

➢ **Scientific Analysis:** Scientific experiments generate largest part of data every day. This includes data collected from nuclear laboratories, space etc. Data mining techniques are capable of the analysis of these data. One can capture and store new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Three typical examples are:

- Sequence analysis in bioinformatics: Genetic data such as the nucleotide sequences in genomic DNA are digital. However, experimental data are inherently noisy, making the search for patterns and the matching of sub-sequences difficult. Machine learning algorithms such as artificial neural nets and hidden Markov chains are a very attractive way to tackle this computationally demanding problem.
- Classification of astronomical objects: The thousands of photographic plates that comprise a large survey of the night sky contain around a billion faint objects. Having measured the attributes of each object, the problem is to classify each object as a particular type of star or galaxy. Given the number of features to consider, as well as the huge number of objects, decision-tree learning algorithms have been found accurate and reliable for this task.
- Medical decision support: Patient records collected for diagnosis and prognosis include symptoms, bodily measurements and laboratory test results. Machine learning methods have been applied to a variety of medical domains to improve decision making. Examples are the induction of rules for early diagnosis of rheumatic diseases and neural nets to recognize the clustered micro-calcifications in digitized mammograms that can lead to cancer. The common technique is the use of data instances or cases to generate an empirical algorithm that makes sense to the scientist and that can be put to practical use for recognition or prediction.[6]

➢ **Intrusion detection:**

Modern network technologies require a high level of security controls to ensure safe and trusted communication of information between the user and a client. An intrusion Detection System is to protect the system after the failure of traditional technologies. Data mining is the extraction of appropriate features from a large amount of data. And, it supports various learning algorithms, i.e. supervised and unsupervised. Intrusion detection is basically a data-centric process so, with the help of data mining algorithms, IDS will also learn from past intrusions, and improve performance from experience along with find unusual activities. It helps in exploring the large increase in the database and gathers only valid information by improving segmentation and help organizations in real-time plan and save time. It has various applications such as detecting anomalous behavior, detecting fraud and abuse, terrorist activities, and investigating crimes through lie detection. Below list of areas in which data mining technology can be carried out for intrusion detection.

- **Using data mining algorithms for developing a new model for IDS:** Data mining algorithm for the IDS model having a higher efficiency rate and lower false alarms. Data mining algorithms can be used for both signature-based and anomaly-based detection. In signature-based detection, training information is classified as either "normal" or "intrusion." A classifier can then be derived to discover acknowledged intrusions. Research on this place has included the software of clarification algorithms, association rule mining, and cost-sensitive modeling. Anomaly-primarily based totally detection builds models of normal behavior and automatically detects massive deviations from it. Methods consist of the software of clustering, outlier analysis, and class algorithms, and statistical approaches. The strategies used have to be efficient and scalable, and able to dealing with community information of excessive volume, dimensional, and heterogeneity.
- **Analysis of Stream data:** Analysis of stream data means is analyzing the data in a continuous manner but data mining is basically used on static data rather than Streaming due to complex calculation and high processing time. Due to the dynamic nature of intrusions and malicious attacks, it is more critical to perform intrusion detection within side the records stream environment. Moreover, an event can be ordinary on its own but taken into consideration malicious if regarded as a part of a series of activities. Thus, its far essential to look at what sequences of activities are regularly encountered together, locate sequential patterns, and pick out outliers.
- **Distributed data mining:** It is used to analyze the random data which is inherently distributed into various databases so; it becomes difficult to integrate processing of the data. Intrusions may be launched from numerous distinctive places and focused on many distinctive destinations. Distributed data mining strategies can be used to investigate community data from numerous network places to detect those distributed attacks.
- **Visualization tools:** These tools are used to represent the data in the form of graphs which helps the user to get a visual understanding of the data. These tools are also used for viewing any anomalous patterns detected. Such tools may encompass capabilities for viewing associations, discriminative patterns, clusters, and outliers. Intrusion detection structures must actually have a graphical user interface that permits safety analysts to pose queries concerning the network data or intrusion detection results.[7]

➢ Data mining in Business:

The companies with clusters of customer database require a tool to refine that data and highlight only the one that is relevant to them. Data mining process is the best tool to highlight the information that is relevant to one's requirement. There is a possibility of drawing a fine line of classification between closely related categories of information using the tools and softwares with many companies competing for the market share there are sure to be lots of options to choose from it's important to really look closely and consider which service best meets your needs.

- **Marketing Techniques:**

In mining basically a tool is used by businesses to turn mere data into useful information which in turn enables them to improvise their marketing and sales strategies. This makes the dealing processes more efficient and eventually increases productivity as well as profit of the businesses. Data miners have access to clusters of information. They look and analyze it logically by studying the buying patterns and associations and formulate a smaller representation. Further the marketing and the sales teams refine and organize this information in an even simpler format using softwares and prepare reports that are easy to read and handy. This information is of paramount to the companies and is used to beat the other competitors in their market.

- **Customer Relation Strategy**:

Any company's customer database is huge since it has been keeping account of every customer which has made even the smallest deal with the company. At times when information required is a little different than the usual categories; it becomes really tedious to get to the root of such evidences in the given time frame. Using the tool of data mining, information can be traced in the formats of customer fields, the pattern of their deliveries and many more. Also the buying and selling patterns can also be acquired with the help of DM. Customer Relation Management (CRM) is about managing the approach of businesses to the relation building with the current or old customers and improving the quality of the customer reviews. It is possible only by improving the quality of delivery and making it better in terms of communication, quality of products and methods of commuting the deals. The current working folks can have access to apt information about the sales

& exchanges made in the past and use the variant information about the process and utilize it for the current clients.

- **Fraud Detection:**

The next important clause is fraudulent cases and clients that commit such crimes. Surveys conducted by many have shown that millions in money is lost in fraud deals every year by many business organizations. The normal traditional ways to detect fraud are time consuming and complex. By studying the fraud cases, it is observed that most of the fraud instances are similar in pattern and content but not necessarily identical always.

Fraud at Point of Sale (POS) is the most practiced of all the other plots. Data miners used the transaction statements and the CCTV footages to handle cases of fraudulent transactions. Retail shops also have their own processing softwares to classify and differentiate between transactions that are genuine and those that are not. Thus data mining techniques assists businesses in solving various serious issues and validates the acquired results. [8]

- **Market Basket Analysis:**

Market basket analysis is a data mining technique that analyzes patterns of co-occurrence and determines the strength of the link between products purchased together. We also refer to it as frequent itemset mining or association analysis. It leverages these patterns recognized in any retail setting to understand the behavior of the customer by identifying the relationships between the items bought by them.

Market basket analysis comprises the following types.

1. **Descriptive market basket analysis**

This type of market basket analysis offers actionable insights based on historical data. It is a frequently used approach that does not make any predictions but rates the association using statistical techniques between the products. We also refer to it as unsupervised learning based on the way it is modeled.

2. **Predictive market basket analysis**

Predictive market basket analysis considers items purchased in sequence to evaluate cross-sell. For instance, when a consumer purchases a laptop, they are more likely to buy an extended warranty with it. This analysis thus helps in recognizing those considered items in a sequence so they can be sold together.

It finds application in the retail industry mainly to determine the item baskets that are purchased together.

3. **Differential market basket analysis**

Differential market basket analysis is a great tool for the competitive analysis that can help you determine why consumers prefer to purchase the same product from a particular platform even when they are labeled with the same price on both platforms.

This decision of the consumers is often based on several factors, as listed below.

- Delivery time
- User experience
- Purchase history between stores, seasons, time periods, and others.

By considering all these factors that are backing the consumers' decision, organizations can benefit from differential market basket analysis. They can make all the parameters fall in accordance with the consumer excel user experience and increase sales on their platform. [9]

- **Protein Folding:**

It is a technique that carefully studies the biological cells and predicts the protein interactions and functionality within biological cells. Applications of this research include determining causes and possible cures for Alzheimer's, Parkinson's, and cancer caused by Protein misfolding.

- **Fraud Detection:**

Nowadays, in this land of cell phones, we can use data mining to analyze cell phone activities for comparing suspicious phone activity. This can help us to detect calls made on cloned phones. Similarly, with credit cards, comparing purchases with historical purchases can detect activity with stolen cards**.** [2]

There are many other areas like Research, E- Commerce, farming, automation etc. , where data mining and the KDD Process has turned to be the boon for the respective organization. It has aided whole world to overcome many crises and recently from Pandemic Covid-19 which had turned into havoc for entire human community.

5. Challenges with Data Mining and KDD:

These days Data Mining and information disclosure are developing a critical innovation for researchers and businesses in numerous spaces. Data Mining was forming into a setup and confided in control, as yet forthcoming data mining challenges must be tackled.

Some of the Data mining challenges are given as under:

1. Security and Social Challenges

Dynamic techniques are done through data assortment sharing, so it requires impressive security. Private information about people and touchy information is gathered for the client's profiles, client standard of conduct understanding—illicit admittance to information and the secret idea of information turning into a significant issue.

2. Noisy and Incomplete Data

Data Mining is the way toward obtaining information from huge volumes of data. This present reality information is noisy, incomplete, and heterogeneous. Data in huge amounts regularly will be unreliable or inaccurate. These issues could be because of human mistakes blunders or errors in the instruments that measure the data.

3. Distributed Data

True data is normally put away on various stages in distributed processing conditions. It may be on the internet, individual systems, or even on databases. It is essentially hard to carry all the data to a unified data archive principally because of technical and organizational reasons.

4. Complex Data

Data is truly heterogeneous, and may be media data, including natural language text, time series, spatial data, temporal data, complex data, audio or video, images, etc. It is truly hard to deal with these various types of data and concentrate on the necessary information. More often than not, new apparatuses and systems would need to be created to separate important information.

5. Performance

The presentation of the data mining framework basically relies upon the productivity of techniques and algorithms utilized. On the off chance that the techniques and algorithms planned are not sufficient; at that point, it will influence the presentation of the data mining measure unfavorably.

6. Scalability and Efficiency of the Algorithms

The Data Mining algorithm should be scalable and efficient to extricate information from tremendous measures of data in the data set.

7. Incorporation of Background Knowledge

In the event that background knowledge can be consolidated, more accurate and reliable data mining arrangements can be found. Predictive tasks can make more accurate predictions, while descriptive tasks can come up with more useful findings. Be that as it may, gathering and including foundation knowledge is an unpredictable cycle.

8. Data Visualization

Data visualization is a vital cycle in data mining since it is the foremost interaction that shows the output in a respectable way to the client. The information extricated ought to pass on the specific significance of what it really plans to pass on. However, ordinarily, it is truly hard to address the information in a precise and straightforward manner to the end-user. The output information and input data being very effective, successful, and complex data perception methods should be applied to make it fruitful.

9. Data Privacy and Security

Data mining typically prompts significant issues regarding governance, privacy, and data security. For instance, when a retailer investigates the purchase details, it uncovers information about purchasing tendencies and choices of customers without their authorization.

10. User Interface

The knowledge is determined utilizing data mining devices is valuable just in the event that it is fascinating or more all reasonable by the client. From great representation translation of data, mining results can be facilitated, and betters comprehend their prerequisites. To get a great perception, many explorations are done for enormous data sets that manipulate and display mined knowledge.

11. Integration of Background Knowledge

Previous information might be utilized to communicate examples to express discovered patterns and to direct the exploration processes.

12. Mining Methodology Challenges

These difficulties are identified with data mining methods and their limits. Mining methods that cause the issue are the control and handling of noise in data, the dimensionality of the domain, diversity of data available, versatility of the mining method, and so on. [10]

6. Conclusion:

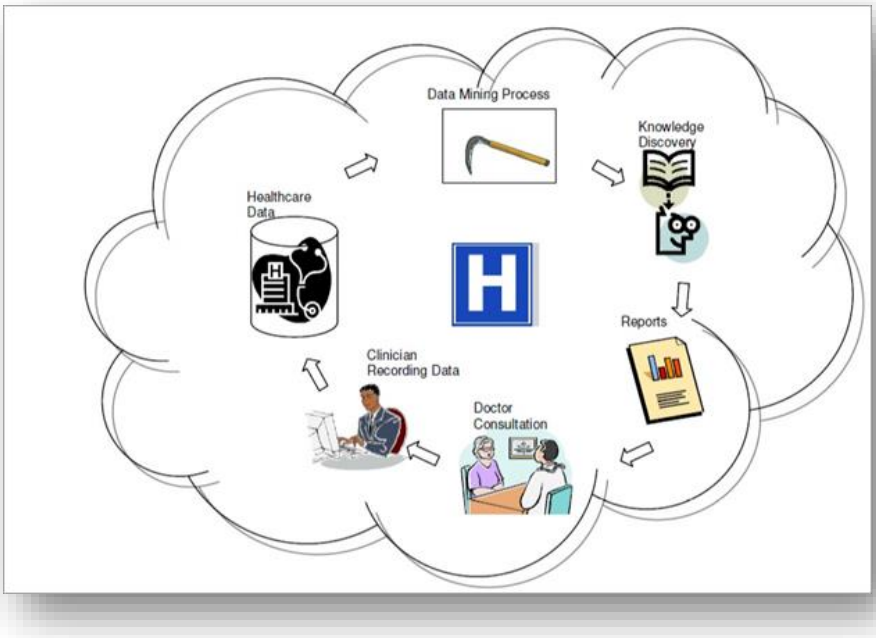

Fig C: Example For Data Mining using KDD for medical treatment.

KDD process helps to derive important patterns and to develop valuable insights for the organization from its own large data set. Data mining is one of the steps in KDD process helping to draw patterns from the Database. The insights/patterns help to make the necessary changes/build new procedures/take advantage of the data in the organization which lead to achieve the business goals in an efficacious way and to grab the competitive advantage. The application of data mining and KDD in real life has abetted in almost every aspect of human life. The advantage gained by an organization can be best analyzed for the organizations with large amount of data. The process may not give good results if the data set is not having amount of data required to derive inferences through it. Data mining and KDD can reveal new possibilities and open new avenues of business opportunities. Even if there are challenges faced by organizations while using data mining and KDD process, it improves business in many domains.

7. REFERENCES

[1] Christopher Clifton,(2022). Data mining. Retrieved from: https://www.britannica.com/technology/data-mining Accessed on 29th October 2022.

[2] Data Mining. Retrieved from: https://www.geeksforgeeks.org/data-mining/ Accessed on 29th October 2022

[3] Knowledge Discovery and Data Mining**.** Retrieved from: https://researcher.watson.ibm.com/researcher/view_group.php?id=144. Accessed on 29th October 2022.

[4] Shivali , Joni Birla , Gurpreet (2015). Knowledge Discovery in Data-Mining. International Journal of Engineering Research & Technology (IJERT), Special Issue – 2015

[5] Sanjuktaranijena, S. Ismail Basha (2015), Data Mining, Knowledge Discovery and its Applications. International Journal of Engineering Research & Technology (IJERT), Special Issue – 2015

[6] H S Venkatesh Prasad , Madhu B K , Lokesha V (2012)A Study on Scientific application of Data Mining . International Journal of Engineering Research & Technology (IJERT),Vol.1(7)

[7] Data Mining For Intrusion Detection and Prevention. Retrieved from: https://www.geeksforgeeks.org/data-mining-for-intrusion-detection-and-prevention. Accessed on 31 st October 2022.

[8] Namita Patil , Data Mining in Business Applications. Retrieved from: https://insightssuccess.com/data-mining-in-business-applications/ Accessed on 31st October 2022.

[9] Market Basket Analysis: Anticipating Customer Behavior. Retrieved from: https://www.turing.com/kb/market-basket-analysis Accessed on 31st October 2022.

[10] Sauvik Acharjee(2021),Data Mining Challenges: A Comprehensive Guide Retrieved from: https://www.jigsawacademy.com/blogs/data-science/data-mining-challenges Accessed on 1st November 2022.