IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data Mining: Techniques, Tools and its Challenges

[1]Dr. Malla Reddy Jogannagari, [2] Mrs. Maheshwari Manchala
[1]Professor & Head, Department of Computer Science & Engineering, Indur Institute of Engineering and Technology,
[1]Indur Institute of Engineering and Technology Siddipet, Telangana State, India
[2]Lecturer in Computer Science & Applications, Department of Computer Science
[2]Government Degree & Post Graduate College, Siddipet, Telangana State, India

Abstract: "Data warehouse" has become a buzzword since few years and followed with emerging interrelated research area of data mining. Data mining is a process of extracting interesting patterns and models (i.e. previously unknown potentially used) from huge data for analytical processing. The various data mining techniques such as description and prediction analysis used to find the hidden new patterns and improve the performance of existing models. It plays a vital role in pattern evaluation which is used for decision making in every aspect of life. The pattern evolution has its own significance in the various application domains This paper provides the overview of data mining system, related data mining techniques, tools and finally concluded with its challenges. It will reshape the subject area and opens the new horizons for researchers to overcome its challenges**.**

Index Terms **– Data Mining, Cleaning, Integration , Selection, Description Analysis, Prediction Analysis.**

I. INTRODUCTION

The organizations transformed their business operations into digitalized, almost every business process is becoming dependent on multitude of information systems which holds the data. Organization employees from top level executives to individuals need to access and analyze the data for finding actionable insights. These derived insights used to take the effective strategic decisions, identify the risks and opportunities for achieving competitive advantage. The data comes from many different sources, but it is easily accessible for analytics, the organizations need to aggregate in common place. Storing the data in uniform system is easier for analysis and reporting.

The evaluation of data collection started with file management systems in 1960. In 1970 onwards database management system concept came into existence with many features. The general idea of data warehousing was invented in 1980 as a response to flow of data with high volume and velocity. The buzzword "data warehousing' first used by IBM researchers and highlighted into popularity. The "Data mining" was introduced in the 1990s, but evolution of data mining in the area of data warehousing with a long history.

In recent years, data mining has focused with a great deal of attention in the software industry and society, due to availability of tremendous data and imminent need to extract useful information and knowledge. It can be used for wide range of applications such as market analysis, customer retention, financial, production control, fraud detection and research exploration. Data mining is viewed as a outcome of the natural evolution of information technology

Data mining is the process of discovering anomalies, patters and correlations within the data sets to predict outcomes. Modern data mining systems involved with data warehousing, statistical analysis, machine learning and artificial intelligence. The organizations use a broad range of data mining techniques to increase revenues, cost cuttings, improve the customer relationships and reduce risks to achieve the competitive advantage.

The data mining has attained tremendous benefits in the late of 1980's. This research paper highlights the significance of data mining system and its various techniques for future generation. The remaining part the paper organized hierarchically is as follows. The Section 2 describes the literature reviews of data mining system. Section 3 focused on the taxonomy of data mining and steps involved in knowledge discovery process. The Section 4 presents the various techniques and tools involved in data mining system. The various challenges discussed in the Section 5. Finally, conclusion and further extension covered in the Section 6 and ended with acknowledgements in Section 7.

II. LITERATURE REVIEW

From the past few years numerous eminent researchers and database professionals have worked on database mining systems and its applicability. Their innovation and ideas can be useful to extend the knowledge discussion in the area of Data mining. They provided valuable research contributions in this area is as following..

Ansha. N et al[1] conducted a Survey on Medical data with using data mining techniques. In her survey Data mining plays vital role and practically utilized in the various such as clinical research, instruction and human services of health care system

Nitesh Kumar Dokania et al [3], analyzed the various data mining techniques with comparative study. They proposed the "perfect" data mining model for dynamic data existed in the real time environment such as markets, financial , spatial and personal.

Priyanka Gutam[5] presented the research paper, that explore the potential impact of big data challenges, open research issues and various tools associated with it, It provides a platform to explore big data at numerous stages and highlighted with challenges.

Dr. Poonam Chaudhary presented a radical review on Data mining system, functionalities and its applications in various areas [6]. Data mining has achieved marvelous success in solving the various problems of different domains. The results gathered from Data mining system is more specific, accurate and useful for decision making.

Ragavi. R et al [7], discussed multidimensional view of data mining system such a data, knowledge, technology and application. They briefly outline methodology and user interaction of data mining system.. The impact of data mining system has strong on decision making in future generations.

Siddardha.K et al [8], discussed the importance of Big data in industry, they highlighted the various challenges, issues, limitations and tools involved in the Big data analytics. It opens the new horizons to researchers to develop the solution to overcome challenges.

Dr. Varun Kumar et al[9], addresses the application of data mining techniques in educational institutions to extract interestingness for the decision making The application of data mining brings a lot of benefits in higher learning institutions

III. TAXONOMY OF DATA MINING

Data Warehouse is a repository of data gathered from heterogeneous sources, stored under unified schema, and that resides at a single site. IT is the core of Business Intelligence system which is build for data analysis and reporting. It is electronic storage of tremendous data which is designed for query and analysis instead of transaction processing. Data warehouse constructed with the process of data cleaning, integration, transformation, loading and periodic refreshing.

The huge amount of data increasing rapidly in the in industry. Traditional statistical and database management systems are inadequate for analyzing this tremendous data. Data Mining is a process of analyzing data in different perspectives and summarizing into meaningful information. Data mining consist of extract, transform and load transaction data onto the data warehouse system. The data mining can extract hidden patterns from tremendous data with algorithms and techniques drawn from the area of Statistical methods, Machine Learning, Artificial Intelligence and Database Management. It uses various tools such as query & reporting tools, analytical processing tools and decision tools.

3.1 Classification of Data Mining System

Data mining is interrelated research area, it is embedded with many disciplines such as databases, machine learning, statistics, visualization, and information science. Moreover, based on the data mining approach used, techniques from other disciplines could also be applied, like neural networks, fuzzy logic, mathematics, knowledge representation, inductive logic programming or high-performance computing. Depending on the kinds of data to be mined , the data mining system may integrate techniques from data analysis, spatial data analysis, pattern recognition, image analysis, signal processing, web technology, computer graphics , business economics or psychology.

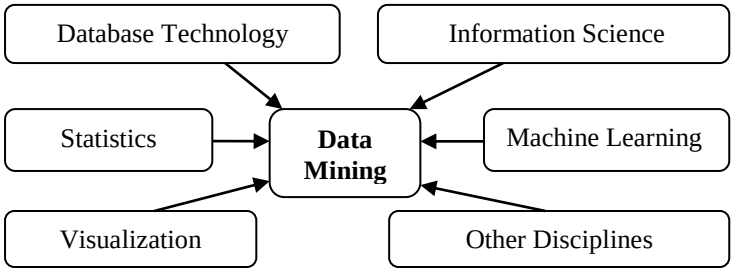


Figure 3.1 Data mining as combination of multiple disciplines

Data mining systems can be classified according to the various criteria as follows.

According to the kinds of databases minded: Data mining system can be classified based on the kind of database and applications s used, each require its own data mining technique. Ex: relational, transactional, and object-rational. The special databases such as spatial, time-series, stream, text, www and multimedia systems have their own mining techniques.

According to the kinds of knowledge mined: Data mining systems can be categorized based on the kinds of knowledge minded and its functionalities These systems classified on descriptive and predictive functionalities. A comprehensive data mining system usually provides multiple data mining functionalities.

According to the kinds of techniques utilized: This kind of data mining systems categorized according the techniques used, it may be degree of user interaction involved or data analysis methods employed. The degree of user interaction systems like autonomous systems, query-driven systems and exploratory systems. Methods of data analysis are other interdisciplinary such as database, data warehouse, machine learning, statistics, visualization, pattern recognition and neural networks etc. Sophisticated data mining system perform effective manner with integrated multiple data mining techniques of various interdisciplinary areas.

According to the application adopted: Data mining Systems can also be categorized according to the applications they adapt such as finance, DNA, stock markets, telecommunication, email etc. Therefore, a generic all purpose data mining system may not fit domain specific tasks

3.2 Steps involved in the Knowledge discovery process

Data mining refers to knowledge discovery from huge amount of data. Data mining as confluence of multiple discipline. It operates on the huge amount of data to extract hidden patterns and relationships for decision making [2][3].

Data mining referred as Knowledge discovery database (KDD) comprises of the following steps which converts raw data into meaning full knowledge patterns and shown in the figure 3.2.

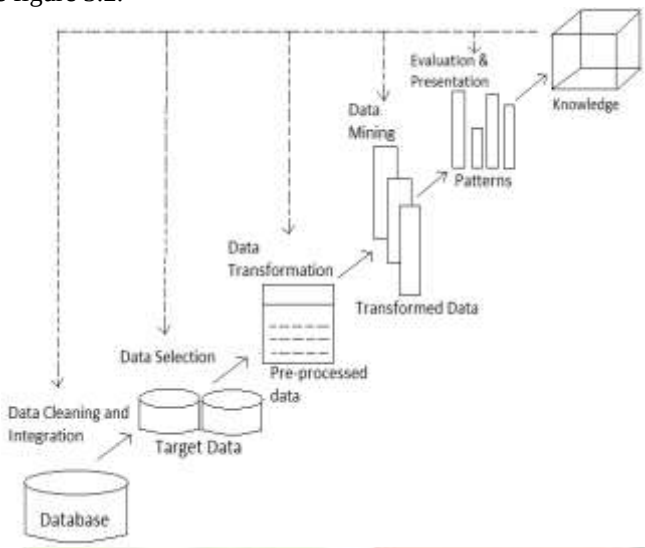


Figure 3.2. Data mining as a core process in KDD

Data cleaning: In real world data may be incomplete, noisy and inconsistent. Data cleaning is process of removing the noisy and inconsistent data from collection. It will attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconstancies in data. Data discrepancy detection and transformation tools can eliminates the anomalies.

.

Data integration: Heterogeneous data collected from multiple sources may be combined into common source. Data can integrated with various tools such as Migration tools, synchronization tools, ETL (Extract Load Transformation) tools.

Data selection**:** At this step, the data relevant to the task analysis is decided and retrieved from the data collection. Data selection can be using various techniques such as neural networks, Decision trees, Naïve Bayes , clustering and Regression etc.

Data transformation: In this step, data is transformed and consolidated into appropriate for data mining with using summary, aggregation and normalization procedures. The data transformation is in two step process such as data mapping and code generation.

Data mining : In this step, intelligent methods are applied in order to extract data patterns potentially needed. Data mining tools predict future trends and helps organizations to take the proactive knowledge driven decisions for getting the competitive advantage. Clustering and association analysis techniques of data mining can transform task relevant into patterns.

Pattern evaluation**:** In this step, strictly interesting patterns representing knowledge are identified based on given measures

Knowledge representation: Final phase of knowledge discovery, the discovered knowledge is visually represented to the user community. This phase uses visualization techniques to help users understand and interpret the data mining results

Decisions/ Use of Discovered Knowledge: The discovered knowledge useful to improve the business operations, reduce the costs, increase the profits and make better decisions with the help of effective data mining technologies. Decisions are made based upon the availability of data as well as the decision situation environment such as certainty, uncertainty and risk.

IV. DATA MINING TECHNIQUES AND TOOLS

Data mining attempts to implement basic processes that extract the structured information and knowledge from the unstructured data. It extracts patterns, associations, changes and anomalies from large data sets. The goal of data mining is to identify authentic, novel, potentially usable and understandable correlations and patterns in existing data.

4.1. Data Mining Techniques : The data mining functionalities are measured to perceive the type of patterns [3]. It is classified into two categories such as Descriptive Analysis and Predictive Analysis and represented in diagram 4.1.

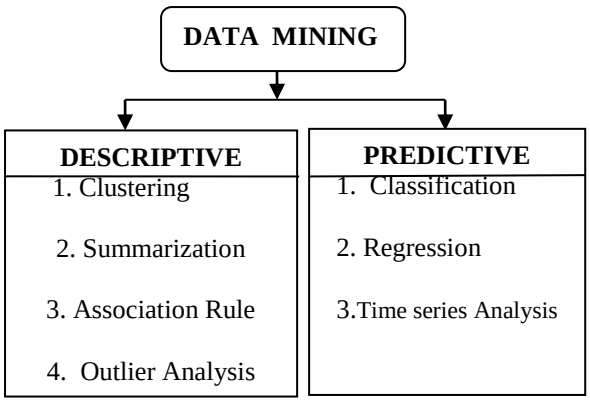


Figure4.1 Classification of Data mining techniques

4.1.1. Descriptive Analysis

Descriptive Analysis is represents the general properties of data stored in database. It focuses finding human-interpretable patterns describing data i.e. cluster, correlation , trends and anomalies etc. Descriptive Analysis classified into various functionalities of the such as a. Clustering b. Summarization c. Association Rule Mining d. Outlier Analysis.

Clustering : Clustering is the process of partitioning objects into groups according to their values, characteristics, similarities and dissimilarities. The objects are clustered on principle of maximizing the interclass similarity and minimizing the interclass similarity. That is, objects in within the cluster are high similarity with each other than the objects in other cluster. Some of the clustering techniques such as K-Means and Y-Means etc. Ex : Clustering of students based on their skills and intelligence represented in 3 clusters such as strong, medium and weak.

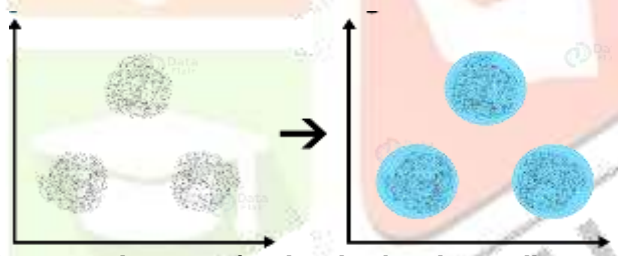


Figure4.2. Clustering of student data based on intelligence

Clustering process used in many areas such as machine learning, bio informatics, pattern recognition, image processing and information retrieval.

Summarization: Summarization technique intended to finding a compact description of data set. Simple summarization methods such as tabulating the mean and standard deviations are often applied in data analytics, data visualization and automated report generations. Clustering is another data mining technique that used to summarize large datasets. For example sum of natural numbers can be written as $\sum_{k=1}^{n}(a_k)$.

Association Analysis: It is the knowledge discovery rule using attribute-value conditions that occur frequently together in a given set of data. Association rule learning is a popular for discovering interesting relations between variables in large databases. It is used to identify strong rules discovered in the databases using different measures of interestingness. Association analysis is mostly used in the market basket analysis and cross marketing.

Association rule is in the form of A=>B, i.e "A1…Am and B1….Bn, where Ai(for 1 to m) and Bj(jto n) are attribute-value pairs.

"If a customer buys bread, he is 70% likely of buying milk." buys (X, "bread") =>buys(X, "milk") [confidence=70%]

The association rule A=>B is interpreted as database tuples that satisfy the conditions in A are also likely to satisfy the conditions in B

Outlier Analysis: Outer analysis is an object in the database which is significantly different behavior from the existing data. " An outlier is an observation which deviates so much from the other observations as to arouse suspicions what it was generated by different mechanism"[], Abnormalities, Deviants, Disorder and Anomalies represented as outliers in the taxonomy of data mining. The Outlier Analysis can be used in fraud detection and prediction of abnormal values.

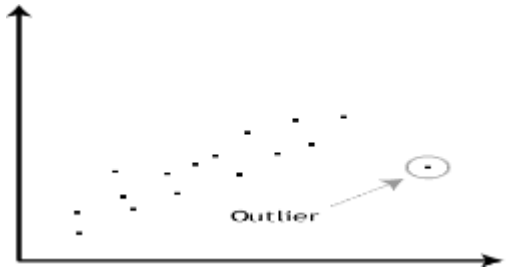


Figure 4.3. Representation of Outlier Analysis

The Outlier can be diagnosed with help of statistical tests that assume probability model for the data.

4.1.2. Predictive Analysis

Predictive analysis model determines the future outcome rather than the existing behavior. It highlights the inference on existing data in order to make the predictions. Prediction involves using some variables in the database to predict unknown or future interestingness. The prediction analysis functionality classified into a. Classification b. Regression Analysis c. Time-series Analysis based on particular perspective.

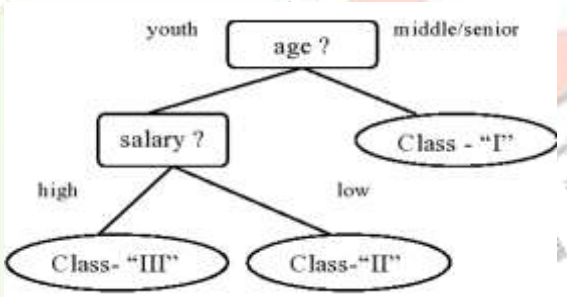
Classification : Classification consists of predicting a certain outcome based on a given input. It is a data analysis method that be used to extract models describing important data classes or to predict future data trends and patterns. In order to process a training data and respective outcome, usually called goal or prediction attribute.

Classification is a data-mining technique that assigns categories to a collection of data to aid in more accurate predictions and analysis. It is used to build models form data with predefined classes as the model is used to classify new instances who classification unknown. The instances used to create the model are known as training data. The derived model may be mentioned with various forms such as classification (IF...THEN) rules, decision trees, mathematical formulae and neural networks.
The Classification of persons into three categories based on the age and salary/his salary. The classification models is as follows.

a. IF .. THEN rule :

*age(X,"young") & salary(X,"high") →Class(X, "I")
*age((X,"young") & salary(X,"low") →Class(X,"II")
*age(X,"middle_aged") → Class(X,"III")
*age(X,"old_aged") → Class(X,"III")

b. Decision trees:



c. Neural Network

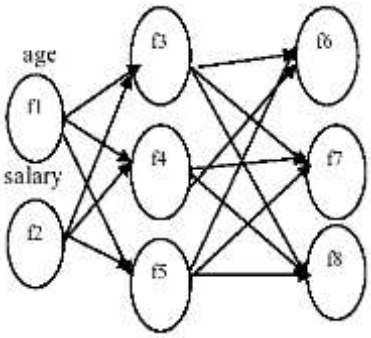


Figure 4.4. A Classification model represented in the form of **(a).** IF..THEN **(b).** Decision Tree (**c).** Neural Network

Regression Analysis : Regression Analysis is a statistical technique used for numeric prediction, although other methods exist as well. It predicts and also encompasses the identification distribution trends based on the available data. The regression analysis predicts profits, sales, temperature and distance. For example predict the asset value based on location, number of rooms and other factors.

This analysis predicts how many dependent variables are effected with independent variable [10]. The simplest form of regression, linear regression formula for equation of a straight line (y=mx+1) and determines the appropriate values for m and b to predict the value of y based on given value of x.

Time Series Analysis: Time series analysis is a statistical technique of trend analysis deals with time series of data that can able to predict the future trends. It means that data is in series at particular time interval. Time series analysis explains why a specific predication was made, analyze the intervals and provide the better understanding of problem.

The sales of various electrical appliances in the year 2019. For example, it would be interesting to predict at which month there is a peak sale of particular appliance consumption. Based on the Time Series Analysis the production of particular appliance can in increased.

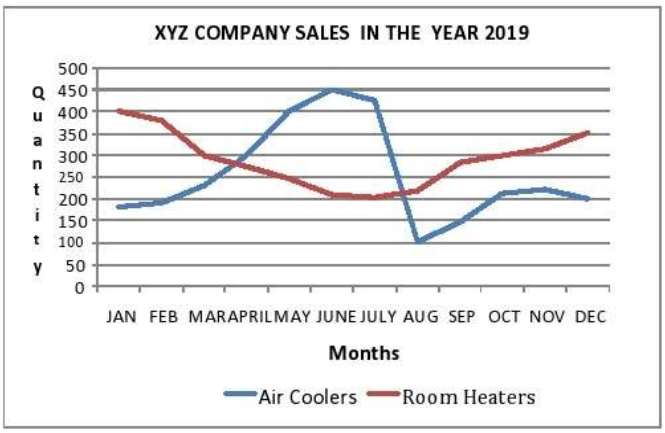


Figure. 4.5 Time Series Analysis of Business data in month wise

Time Series Analysis widely used for non-stationary data such as economic/ financial, weather, stock price, retail sales prediction and forecasting.

4.2. Data Mining Tools : Data mining tools predicts future trends and make business into proactive[4]. The various data mining tools such as Rapid Miner, Mahout, Orange, Weka and Data Melt can tackle the business problems effectively.

Rapid Miner : It is open source software for data and text mining. Rapid analytics is a server product.

Mahout: Mahout offers clustering and based collaborative filtering that run on top of Hadoop

Orange : This project make data mining effective for both novices and experts. It provides a wide variety of visualization plus a tools boxes with widgets.

Weka: Weka provides set of data mining algorithms for knowledge analysis. User can apply data directly or use with java application.

DataMelt : DataMelt provides mathematical computation data mining statistical analysis and data visualization.

V. DATA MINING CHALLENGES

Data Mining is plays vital role in business and many other domains. It is innovative trusted technology, still some challenges have to be resolved. Some of the challenges are highlighted as following.

5.1. User Interface issues: Data Mining tools extract the knowledge of user interestingness. It can be presented with visualization for better understanding.

Mining based on Level of Abstraction: The data warehouse contains the large not of records with more numbers of attributes. High dimensional datasets require large size of search space for efficient model construction to perform mining task. The user can interact with data mining system to view data and discovered patterns using drilling down, rolling up and pivoting at multiple granularities with different angles.

Different kinds of knowledge to be minded : Various users interested in different kinds of knowledge from the same database. Data mining system provide the wide spectrum of tasks for data analysis and knowledge discovery. These tasks perform in different ways and require the development of many data mining techniques.

Incorporation of background knowledge: Knowledge discovery process incorporated with domain knowledge. The user should have the back ground domain knowledge relevant with constraints, rules and other information.. These knowledge is used to find the interesting pattern for evaluation[7].

Presentation of Mining Results: Discovery knowledge easily understandable using in high level languages, visual representations and other expressive formats. The data mining system is to be interactive by adopting expressive knowledge techniques such as trees, tables, graphs and matrices.

Query Languages and adhoc Data Mining : Query languages plays vital role in the relational database for information retrieval. In similar way, Data mining query languages are needed to develop for information retrieval and ad hoc data mining tasks. Such language should be integrated with data warehouse query language for efficient and flexible data mining.

Pattern evaluation and guided mining : The pattern evaluation may depends on user interestingness. A data mining system can uncover thousands of patterns due to lack of interestingness. But, the uncovered also having its own significance

Problems with Noisy and incomplete data: Database sometimes consists with inconsistent and incomplete data due to the problem of noise which leads to erroneous results in data mining. There are various statistical methods to deal with missing inputs and identify noisy attributes values.

5.2. Performance issues*:* Data mining algorithms performance issues associated with the following.

Efficiency and scalability: The data mining algorithms must be efficient and scalable in order to extract information from huge amount of data in data warehouse. The execution of data mining algorithms must be predictable time and acceptable.

Network oriented data mining algorithms :The large volume of data distributed over the network environment, computational complexity motivate the development of parallel and distributed data mining algorithms. These algorithms decompose the data into segments and process in parallel then merged together. Moreover, incremental data mining algorithms can update the
database and mine the data again "from scratch".

5.3. Issues related with diversity of databases: Data mining algorithms associated with the following diversity issues.

Complexity of Data types in various databases: Dissimilarity of data types in existed in the various databases such relational, spatial, hypertext, temporal data and multimedia, there is complexity in extracting the insights in data mining algorithms. Specific data mining system should be constructed for specific kinds of database[6].

Mining from Varied Sources: The data is collected from various sources on network environment. The discovery of knowledge in the form of structured, semi structured, unstructured data with semantics poses great challenges to data mining.

5.4. Other Issues :
Security and Social Challenges: Security is most important for every technology. There is no exemption for data mining from third party applications with lack of security. Data Warehouse consists with sensitive data of personal and business.

VI. CONCLUSION

Since the inception data mining has attained amazing success in solving various problems with its techniques. It is practical oriented technology involved in many domains. This paper focused on the Knowledge discovery process and indicates the capabilities of various data mining techniques and its challenges. Many techniques can be implemented for descriptive and predictive analysis. The user should determine data mining techniques, which technique is used and when. Most of the time the techniques to be used are determined by trial and error. Each technique is problem specific, but in real world data may be continuously changing in dynamic way such as market, finance, spatial and video surveillance. It is not possible to build the perfect model with the dynamic nature of data. But sometimes, it is crucial for decision maker before taking a certain decision. Unpredictable nature of data and chaotic ways data mining techniques are not always predicable. There is a need of extensive research in the area of data mining techniques for scientific and business applications.

VII. ACKNOWLEDGMENT

We sincerely acknowledge to the Department of Computer Science & Engineering(R&D), Indur Instittute of Engineering & Technology, Siddipet and Department of Computer Science & Applications, Government Degree &Post Graduate College, Siddipet for their resource, support, guidance, and cooperation during the period of analysis.

REFERENCES

[1]. Ansha. N et al, " A Survey on Medical Data by using Data Mining Techniques", International Journal of Science, Engineering and Technology(IJSETR)", Vol.7(1), Jan,2018.

[2]. Deshpande.S.P. et al, " Data Mining System and Applications : A Review", International Journal of Distributed & Parallel System (IJDPS), Vol. 1(1), Sep, 2010.

[3]. Nilesh Kumar Dokania et al, " Comparative study of various Techniques in Data Mining", International Journal of Engineering Sciences & Research", Vol. 7(5), May.2018

[4]. Nolofar Rehman, " Data Mining Techniques Methods, Algorithms and Tools", International Journal of Computer Science & Mobile Computing, Vol. 6(7),July, 2017.

[5]. Priyanka Gautam, "Impact of Data Mining on Big Data Analytics : Challenges and Opportunities", Vol. 57(1), March, 2018.

[6]. Poonam Chaudhary," Data Mining System, Functionalities and Applications : A Radical Review", International Journal of Innovations in Engineering and Technology (IJIET), Vol 5(2), April, 2015.

[7]. Ragavi. R et al," Data Mining Issues and Challenges : A Review", International Journal of Advanced Research in Computer and Communication Engineering,
Vol. 7(11), Nov, 2018.

[8]. Siddardha. K et al, " Big Data Analytics : Challenges , Tools and Limitations", Internatianl Journal of Engineering and Technical Research, Vol. 3(3), Nov, 2016.

[9]. Varun Kumar, " An Empirical Study of Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2(3), March 2011.

[10]. Vivekananth. P et al, " An Analysis of Big Data Analytics Techniques", International Journal of Engineering and Management Research, Vol. 5(5), Oct, 2015.