IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A STUDY ON DATA MINING-ITS IMPORTANCE

Tamil Elakya.T[1], Dr.K.Manikandan[2]

Ph.D Research Scholar[1], Associate Professor[2]

Department of Computer Science[1 &2]

PSG College of Arts & Science, Coimbatore, India [1 & 2]

Abstract:

Technological advances have resulted in the development of a wide variety of new fields. The advent of Data Mining leads to many positive changes in the technology. This paper discuss about the data mining tasks, techniques it's and tools and applications.

Keywords:

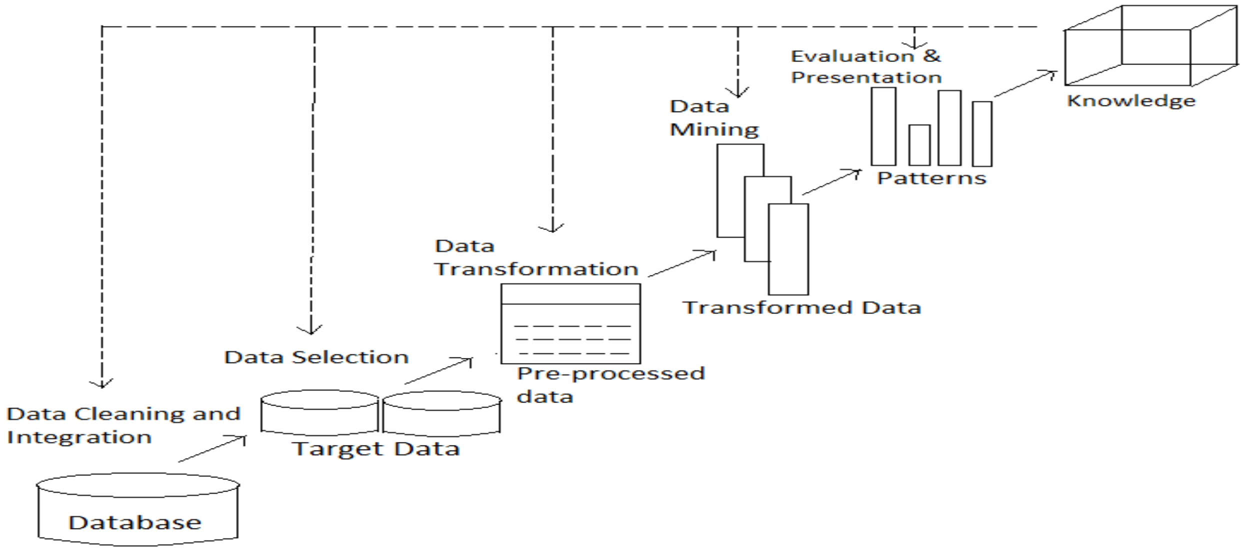
Data Mining, KDD, Data Mining techniques, Data Mining tools.

INTRODUCTION:

The development of information technology has been accompanied by a natural integration of data mining**.** More number of data are generated everyday. Different types of sophisticated tools are required to treat the data that can be converted into knowledge. This led to the birth of data mining. Data mining involves sifting through vast datasets to discover patterns and relationships that can aid in solving business challenges through data analysis. It is also called as Knowledge Discovery in Data. The purpose of data mining is to identify real, novel, potentially usable and understandable connections and patterns in existing data.[10].

STEPS IN DATA MINING:

The process comprises seven essential steps that aid in the discovery of knowledge.



i) Data Cleaning:

The first step in data mining is the concept of data cleaning. The other name of data cleaning is called data scrubbing. It refers to the procedure of rectifying or removing inaccurate, corrupted, inadequately formatted, duplicated, or incomplete data within a dataset.

ii) Data Integration:

It means processing of data from diverse sources, which may have different formats and structures, and merging them seamlessly to maintain a cohesive and unified perspective of the information.

iii) Data Selection:

Data selection is the procedure of identifying and retrieving data that is pertinent to the analysis from the data collection.

iv) Data Transformation:

It is the process that involves transforming, cleaning, and organizing data into a suitable format that can be analyzed to facilitate decision-making.

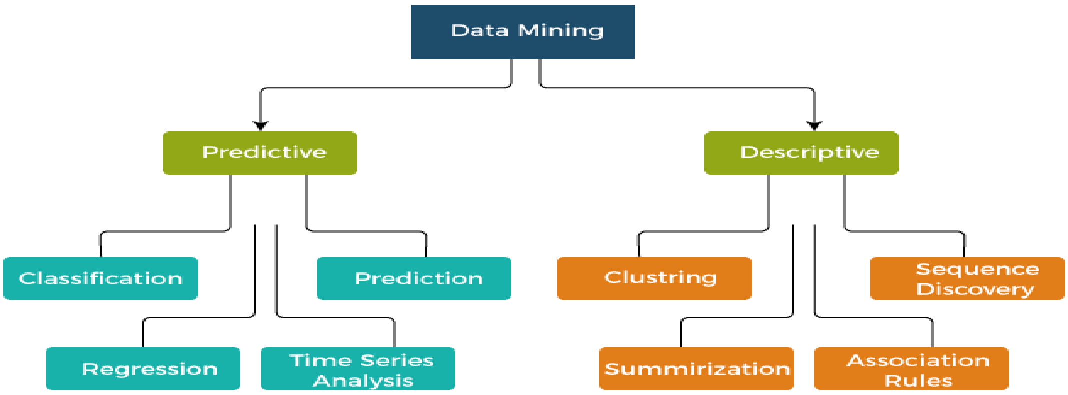
v) Pattern Evaluation & presentation:

Pattern evaluation in data mining is to discover patterns or rules that are assessed and validated to understand their relevance, importance, and potential impact.

Knowledge representation is the process of structuring information in a format that can be efficiently utilized for problem-solving tasks.

DATA MINING TECHNIQUES:

A variety of data mining techniques are applied depending on the type of task being performed. The two major tasks of data mining involves Descriptive and predictive [10] Analyzing past data to identify what occurred in the past is descriptive task in data mining. It primarily aims to summarize and transform data into useful information for monitoring purposes. Example sales tracking. A predictive task involves predicting future outcomes based on historical data. It aims to forecast possible future results. Example medical diagnosis.



Some of the data mining techniques that fall under the predictive include:

Classification:

Classification is a process of building a model that accurately predicts the class label of new instances based on their characteristics. For Example the data can be classified as private data, public data and restricted data in an organization. Data classification is two-step process. [14] one is the training phase and another one is test phase.

Some of the most common classification algorithm includes:

- Decision Tree Algorithm
- Naïve Bayes
- K-nearest neighbour
- Support Vector machine
- Linear Regression
- Random Forest algorithm
- Neural Networks

Prediction:

An objective of this technique is to group unknown data points into predefined categories or classes. It deals with categorical values. A numerical input is found in this technique. The training dataset contains numerical output values and input values as it creates a model.

Regression:

Continuous numerical values can be predicted in regression. The goal of this technique to exhibit a relationship between independent variable (also known as predictors) and a dependent variable (target variable) Example Age and height calculation.

Some of the most common regression algorithm includes:

- Simple Linear Regression
- Lasso Regression
- Multivariate Regression
- Logistic Regression

Time series Analysis:

It carry out the task like analysis, modelling and forecasting of data points collected over time. The main objective of this technique is to understand patterns, trends, and seasonality in order to predict future values.

Descriptive model includes various techniques like

Clustering:

The word cluster refers to the group .In other words, it means that similar data has been grouped together. It is also called a data segmentation. This method helps to considergaps between data and similarities [9]. Most commonly clustering is divided into hard clustering (Objects belonging to the same cluster) and software clustering (Objects belonging to the several cluster).

Some of the most common Clustering algorithm includes

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods Model-based methods

Summarization:

Summarization is a process of identifying and representing patterns in large datasets in a streamlined manner. The process is also known as data generalization or data compression. It includes two kinds like text summarization (compressing large documents into shorter versions) and data summarization (creating concise and informative summaries of large datasets.

Sequence discovery:

It helps to detect or discover similar patterns or trends in transaction data over a given timeframe. [9] It is used to predict sequentialdependencies and sub sequences.[2]

Association Rules:

Association Rule mining technique finds patterns in data and relationships among large data sets and correlations (connection) between them.[2] It is composed of two parts, an antecedent (if) and a consequent (then).Example of Association rule mining is Market Basket Analysis.

Some of the Association rule algorithm includes:

- Apriori Algorithm.
- FP (Frequent Pattern) Growth algorithm.

- CMAR(Classification Based on Multiple Association Rule)
- Maximal Frequent Item set Algorithm.
- Classification Based on Association Rule.

TOOLS IN DATA MINING:

The purpose of data mining tools is to analyze large amounts of data in order to discern meaningful patterns, trends, relationships, and insights within those data sets. Some of the data mining tools include:

1. **IDA (Interactive Data Analysis)**-Oldest Data Mining tool

   Developed year: Early in 1960

   Developed by: Stanford Research Institute

   Purpose of the tool: For exploration and analysis of large datasets.

2. **Orange:**

   Developed year: 1996

   Developed at: Bioinformatics Laboratory at the University of Ljubjana Slovenia.

   Purpose of the tool: It offers a diverse array of visualization options and a collection of toolboxes featuring widgets.

3. **WEKA:** Open Source tool

   Developed year: 1997

   Developed by: H.Witten and Eibi Frank

   Purpose of the tool: Helpful to handle moderately sized data sets.

4. **SAS Enterprise Miner:**

   Developed year: 1999

   Developed by: SAS Institute.

   Purpose of the tool: To analyze large data sets, discover patterns and make decisions.

5. **ELKI (Environment for Developing KDD Applications supported by Index Structures)**

   Developed year: 2009

   Developed by: The Database systems and Information Management DIMA group.

   Purpose of the tool: Efficient for working on KDD and Data mining tasks.

6. **Knime**

   Developed year: 2004

   Developed by: Team of researchers at the University of Konstanz in Germany

   Purpose of the tool: Helps users to visually design data workflows.

7. **Rattle**

   Developed year: 2005

   Developed by: Graham Williams

   Purpose of the tool: Helps in exploring data, preprocess it also helps in evaluating models without the need of complex code.

APPLICATIONS OF DATA MINING:

Data mining finds extensive use in diverse domains, offering a multitude of applications.[2][6][8] Several essential sectors benefit from the utilization of data mining, and some notable examples included

1. Healthcare and Medicine:

i) Diagnosis of a Diseases:

Examining patient data, symptoms, and medical history to assist in precise disease diagnosis and anticipate disease outcomes, ultimately leading to improved treatment planning.

ii) Health Behavior Analysis:

Analyzing patient lifestyle and behavior data, healthcare providers gain insights into the factors that impact health outcomes and can develop effective interventions.

2. Business and Finance:

i) Stock Market analysis:

Examining of past stock market data, data mining can identify patterns and trends that aid in making well-informed investment choices.

ii) Sentiment Analysis:

By analyzing customer feedback, social media posts, and online reviews, businesses can gain insights into customer sentiment and assess public perception of a product or brand.

3. Education:

i) Identifying learning gap:

Educational institutions can pinpoint learning gaps and misconceptions in student comprehension, allowing for targeted remediation and ultimately enhancing learning outcomes.

ii) In Research:

To gain insights into educational trends, polices and factors influencing student performance.

4. Government and public sector:

i) Budget and Resource allocation:

To analyze historical budgetary data and spending patterns. It enables the government to make decisions regarding resource allocation and budget planning.

ii) Traffic Management:

Traffic cameras, sensors, GPS system can be used for enabling the optimization of traffic flow, congestion reduction.

CONCLUSION:

Ultimately, data mining aims to create valuable knowledge from raw data, which allows organizations and individuals to make informed decisions, gain competitive advantage, and gain insights that would otherwise be difficult or impossible to uncover by means of manual analysis.

REFERENCES:

[1] Weini Chen, The application of Machine learning in Data Mining under Big Data Environment in International Conference on Network, Communication, Computer Engineering (NCCE 2018),Volume 147

[2] Koti Neha, M Yogi Reddy," A Study On Applications Of Data Mining" in International Journal of Scientific & Technology Research, Volume: 9, Issue: 2, ISSN2277-8616 (February 2020)

[3] Mrs. Bharati M. Ramageri, " Data Mining Techniques And Applications", in Indian Journal of computer science and engineering Volume: 1 Pages 301-305.

[4] Mr. S. P. Deshpande and Dr. V. M. Thakare, "Data Mining and Applications :A Review", in International Journal of Distributed and Parallel systems Volume:1(September 2010)

[5] Annan Naidu Paidi, "Data Mining: Future Trends and Applications" in International Journal of Modern Engineering Research(IJMER),Volume:2,Issue: 6,ISSN:2249-6645(Nov-Dec 2012)

[6] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The survey of Data mining applications and feature scope", in International Journal of computer Science,Engineering and Information Technology(IJCSEIT),Volume:2 ,No:3(June 2012)

[7] S.A.R,Niha,"Study of Data mining methods and its applications" in International Research Journal of Engineering and Technology(IRJET), Volume:04,Issue:11,e ISSN:2395-0056,p-ISSN 2395-0072,(November 2017)

[8] Mustafa Abdalrassual Jassim and Sarah N. Abdulwahid" Data mining preparation: process, Techniques and Major issues in Data Analysis" in IOP Conference Series: Materials Science and Engineering (2021)

[9] Dr. Malla Reddy Jogannagari, Mrs. Maheshwari Manchala,"Data Mining: Techniques ,Tools and its Challenges" in International Journal of creative Research Thoughts(IJCRT),Volume:0,Issue:07,ISSN:2320-2882(July 2020)

[10] Urvashi Sangwan,"A Survey on use of Data Mining Technique in Different Domain" in International Journal of Latest Trends in Engineering and Technology, Volume:12, Issue:4, e-ISSN:2278-621X

[11] Ashour A N Mostafa , Hamdi Elmburok Abdulghani Mahmoud, "Review of Data Mining concepts and its Techniques" in international Journal of Academic Research in Business& Social Sciences in Volume. 12, No. 6, 2022, Pg. 611 – 619,ISSN: 2222-6990

[12] Anshu,"Review paper on Data mining Techniques and Applications", in International Journal off Innovative Research in Computer Science & Technology(IJIRCST)Volume:7,Issue 2,ISSN:2347-5552(March 2019