Data Mining preparation: Process, Techniques and Major Issues in Data Analysis

Mustafa Abdalrassual Jassim*,1, Sarah N. Abdulwahid1
University of Monastir, Tunis, Al Muthanna University, Muthanna, Iraq, Ministry of Higher Education and Scientific Research, Baghdad, Iraq

*Corresponding author, email: mustafa944@mu.edu.iq

Abstract. Data preparation is an essential stage in data analysis. Many institutions or companies are interested in converting data into pure forms that can be used for scientific and profit purposes. It helps you set goals regarding system capabilities and features or the benefits your company expects from its investment. This purpose creates an immediate need to review and prepare the data to clean the raw data. In this paper, we highlight the importance of data preparation in data analysis and data extraction techniques, in addition to an integrated overview of relevant recent studies dealing with mining methodology, data types diversity, user interaction, and data mining. Finally, we suggest some potential suggestions for future research and development.

1. Introduction
Data mining is a collection of techniques that apply to large and complex databases [1]. To eradicate randomness and discover concealed patterns [2]. Methods of data mining are also computationally-intensive. To discover trends in the data, we use data mining instruments, methodologies, and theories. There are many driving forces present at the moment. [3]. Data mining has therefore become an important area of science. Data mining is one of the powerful new technologies that has emerged. It facilitates both individual and enterprise users to find data from a set or large set of data[4]. The data cached by the structure will be passed through a filter if it is specified and analyzed in the model by the algorithm. The algorithm calculates a set of summary statistics that describe the data, identifies rules and patterns within the data, and then uses those rules and patterns to fill in the form[5][6]. The mining model provides a wealth of knowledge on research data and patterns, including statistics, rules, and regression formulas. Data extraction queries can be generated to recover and use this information for analysis and presentation. There are two types of data mining: descriptive mining and predictive mining with different functions and technologies [6]. Technologies play a significant role in achieving successful models. Techniques of data mining include three major groups: artificial intelligence technologies, machine learning techniques, and statistical techniques.

It should be remembered that each of these technologies has its algorithms to run the models for the best solution [7]. Data production tools enable enterprise users and analysts to address conventional data challenges by simplifying and automating processes for data collection, cleaning, and mixing [8].
In other words, technology enables all users to easily and quickly access, process, enrich, merge and reconcile fragmented data, be it structured or unstructured, from virtually any source. The data can then be prepared for analytical use using tablets and other comprehensive manual processes in a fraction of the time. The data analysis process relies on methods and techniques for taking primary data and mining to obtain insights related to the fundamental objectives. Of exploring this information to transform measures, facts, and figures into improvement initiatives that largely depend on two main areas: methods of quantitative data analysis and methods of data analysis in Qualitative research.

1.1. Context

Data preparation is an essential stage in data analysis. Data preparation processes are the first four processes, namely, data cleaning, data integration, data collection, and data transformation [9]. Data mining, pattern assessment, and information representation were merged to create a single data mining process. [10]. Different data extraction processes can be divided into two groups: preprocessing or data preparation and data extraction [11].

The data mining method cannot be completed in a single stage. In other words, the information required from large volumes of data cannot necessarily be obtained. We assume this is a very complex procedure involving a number of systems. In the order defined, processes such as data cleaning, data integration, data collection, data transformation, data extraction, data pattern analysis, and information representation must be completed. Data preparation is also a tool for the handling and study of raw data. This is an important preprocessing step, often involving data reformatting, data correction, and data fusion to improve the data.

Data processing is often a long-term activity for data professionals or business users. However, data should be put into context as a prerequisite for making observations and removing distortions due to poor data quality. For example, data preparation usually involves data format standardization, enrichment of source data, and outliers' removal. In essence, the preparation of data refers to a collection of procedures that produce data for algorithms. The following are typical steps involved in preparing the data, as shown in figure (1).

![Figure 1. Data preparation structure.](image-url)
Data mining is the central part of the method of knowledge detection. KDP is a process of seeking knowledge in data, using data mining methods (algorithms) to extract challenging knowledge from large quantities of information. The process of information exploration will consist of the following steps:

1- **Data Source:** In the context of computer science and computer applications, the source of the data is where the data used comes from. The source of data may be a database, a static file, live measurements on the physical system, abstract web data, or one of the myriad data services that stream and static over the internet. The primary concern for information accuracy is the data source. In such cases, to help companies and institutions operate more efficiently. Identifying data sources is the first step in any data storage project because you cannot do anything without the data. After setting up the right plan to obtain accurate information (data), the next step is to know how to store it consistently and in the same format so that when you run the reports, you can get the right results for decision making. Ultimately, data sources aim to help users and applications connect with and move data where it needs to be.

2- **Data Cleaning:** Data cleaning is the mechanism by which wrong, corrupt, misformatted, duplicate or incomplete data is corrected or deleted within a dataset.

3- **Data Integration:** Describes how data from multiple sources are integrated into usable and meaningful knowledge using a mix of techniques and business processes.

4- **Data Selection:** The term data selection seeks to choose data that should be preserved or shared/archived when the project is finished during the data collection.

5- **Data Transformation:** The transformation of data is the process of transferring data from one format or structure to another. For activities such as data integration and data management, data transformation is important.

6- **Pattern Evaluation:** The pattern evaluation describes fascinating patterns of information based on various types of interesting steps. A pattern is seen as appealing because it is potentially useful, readily understood by humans, and uses summary and viewing to make data understandable.

7- **Knowledge Presentation:** The representation of knowledge is a show of knowledge to the user in terms of trees, tables, rules, graphs, charts, matrices, etc. Represented one Methods for understanding the ins and outs of preparing data are the following items shown in Table 1.

### Table 1: Representation of preparing data.

| (a) Deliver | 1- Delivery is all about structuring the distilled data into the format that the consuming process or user needs. |
|-----------------|------------------------------------------------------------------------------------------------------------------|
|                 | 2- The data set(s) delivered for ongoing detention, and in the case of detention, should also be evaluated. |
|                 | 3- Supporting metadata should be added to the data catalog. These steps allow data to be discovered by other users. |
|                 | 4- The use of the data provided should be monitored - and the unused data should be deleted - after a specified period of time. |

### 1.2. Motivation and contribution

Data is collected for many purposes. Hence, there is often a need to identify and extract relevant data for the specific purpose of analysis. Each system has specific requirements regarding how the data for
analysis should be submitted and therefore the data should be converted to fulfill these demands. Moreover, the selection of specific data to be analyzed can greatly influence the models learned. It is often the most time-consuming part of any data mining project. Many researchers have recognized high-end data extraction expertise and information as essential. research subject in machine learning and the database system and in many industrial enterprises as an interesting field with the potential of generating substantial revenues. Data, information, or knowledge have an important role in human activities. Data mining has significance in finding patterns, forecasting and discovering knowledge, etc. in various business fields. Data mining is used in the Medical Sciences, the detection of malicious executables, statistical techniques, identifying patterns, sales forecasting, basket analysis, mathematical organizations, and the presentation of information in a way that can be easily interpreted by people. This allows businesses and organizations to concentrate their stored data on the most relevant information. The massive increase in data in recent years has led to this, prompting recording, processing, and analyzing these records.

Research Contributions To automate data preparation, several challenges must be addressed, including: (1) accommodating several different components of the same task, establish alternative approaches; (2) Coordination of the various components of data preparation that rely on different evidence about the problem to be solved; (3) Identify the evidence that enables us to perform steps on the data; (4) Creating several options between candidate and alternative solutions, in light of multiple requirements; (5) Linking multiple components for the purpose of forming transactions.

1.3. Review methodology

Data mining is the process by which useful data, patterns and trends from several data are collected, using techniques like clustering, classification, regression and correlation [12]. Data, information, and knowledge are the exciting roles of human life. Massive data warehouses with the rapid development of file technologies require big data analysis and modelling to predict future information trends [13]. Data mining is called a technique by which the necessary information in databases can be extracted from raw information. Using the data mining prediction analysis methodology, future scenarios can be predicted with regard to current knowledge. Forecast analysis is a combination of classification and aggregation [14] [18]. Data mining is used for data extraction from a great deal of information. Data mining is made up of two predictive and descriptive models. Data management aims to collect data in stratification files with either the ultimate aim of learning new effects or seeing new areas [15]. Data analysis techniques were applied to the higher education institution's educational data. The analytical data included event records which extracted hidden information from the data by performing pattern recognition and forecast modelling tasks [16] [19]. Solve many complex problems, including energy efficiency and energy use, structural analysis, building materials, smart cities, design and optimization, technology forecasts, soil engineering and construction engineering. Accurate forecasting of traffic information, predicting the occurrence of COVID-19, the user must use the information published effectively through prediction models [17]. In this section, the applied methods used in our approaches, such as decision support systems, data mining, and correct data preparation and extraction, have an effective role in showing the right results through their application to data mining techniques.

The results differed between accurate and imprecise due to their reliance on the techniques and algorithms used to show results on prepared data. Therefore, the main features of the conceptual data management platform must be followed to prepare the data to offer more accurate results before implementing it, which are:

1. Taking the initial data and preparing it for analysis.
2. Apply algorithms to raw data to reveal new insights.
3. Create database categories to place raw data.
4. Data collection and classification.
5. Collect data from multiple sources and collect them together.
6. Converting unstructured data into data ready for analysis.
7- Merging different types of data into a unified system.
8- Converting data from one type to data of another kind.

2. Data Mining Implementation Process
Many different sectors are leveraging data mining to enhance their businesses' efficiency, including manufacturing, banking, marketing, aerospace, education, health, etc. Therefore, the need for the traditional data extraction process has effectively improved. Data mining techniques should be reliable and reproducible by company personnel with little or no knowledge of the data mining context. Several steps determine the classic presentation (see Table 3). First, The problem has to be identified in terms of work or academic objectives and converted into concrete data mining and analytical objectives. The second stage is the data step, the discovery, fusion and transformation of primary data sources to be used for the related data mining mission. This is typically the longest step unless the process is completely automated. The third step is step modelling. Algorithms are used to extract real data patterns, to predict or to calculate my metadata. In the fourth stage these patterns and models are evaluated in a quality and content format. The extracted forms will be added to the new data during the final publication stage and the findings are combined with other details for effective action.

Table 3: Data Mining Process techniques

| Process   | Attributes                                      |
|-----------|-------------------------------------------------|
| Planning  | Define problem objectives                       |
|           | Translate into Data Mining objectives           |
|           | Define Analysis Approach                        |
|           | Data Requirements                               |
| Data      | Collecting Data from Multiple Source            |
|           | Combining Data into Single Source               |
|           | Cleaning Data, Selecting, Transforming Data     |
| Modeling  | Predictive Data Mining                          |
| Evaluation| Descriptive Data Mining                         |
|           | Evaluate Quality                                |
|           | (Accuracy, Cost, Variance, Implication, Action ability, ....) Application |
| Deployment| Post Processing                                 |
|           | Monitoring                                      |

3. Major Issues in Data Mining
With the increasing growth of data in any application, data mining satisfies an imminent need for accurate, scalable, and flexible data analysis in our society. Data mining can be seen as a natural IT creation and a convergence of several disciplines and associated fields of application. Although data mining is very effective, it faces numerous challenges throughout its implementation. Data mining issues and challenges can involve performance, data, techniques, etc. Data mining is successful if issues or problems are correctly detected and properly sorted. Data can be collected on any data as long as data is essential for the target app, such as database data, data warehouse data, transaction information, and advanced data forms. Figure 2 demonstrates the data mining problems.
4. Data mining techniques

Organizations and institutions today have more access than ever to data. However, it can be very difficult to grasp large volumes of structured and unstructured data to strengthen the organization and at other levels, due to the sheer volume of information. If not properly treated, this challenge will lessen the benefits of all data. Data mining is how businesses discover data trends to obtain knowledge that is important to their business needs. For both business intelligence and data science, it is important. There are many techniques for data mining that organizations can use to convert raw information into actionable insights. This encompasses everything from advanced artificial intelligence to fundamental data planning, which are the secret to optimizing the value of data investment.

4.1. Classification

It’s a task of data analysis, i.e. the process of finding a model that describes and differentiates data classes and concepts. This data extraction approach helps to classify data into different categories.

4.2. Clustering

Cluster analysis is a form of data extraction to classify related data. This method helps to consider gaps between data and similarities.

4.3. Regression

Regression analysis is a data extraction process in which the relationship between variables is defined and analyzed. It is used to evaluate a given variable’s likelihood, since there are other variables.

4.4. Association rules

This data extraction technique helps to find the connection between two or more objects. Detects a pattern in the dataset.
4.5. Outer detection
This type of data extraction technique refers to the observation of data elements that do not fit the predicted behavior pattern in the data collection. This technology can be used in different areas including intrusion, tracking, fraud, detection of bugs, etc. Offshore analysis or offshore mining is also known as external detection.

4.6. Sequential patterns
This technique of data collection helps to detect or discover similar patterns or trends in transaction data over a given timeframe.

4.7. Prediction
The predictions used numerous other techniques for data mining, such as patterns, sequences, clustering, grouping, etc. It analyses past events or circumstances in the right order to predict a future occurrence.

5. Conclusion and future works
Data mining is important in pattern finding, prediction, discovery of knowledge, etc. in different fields of industry. Data mining applications use a range of data types, from text to photos, warehouses, and different databases and data structures.
Different data mining techniques for extracting patterns and hence knowledge from these various databases. Data collection and methods Data mining is an essential activity and domain awareness is crucial in this process. A range of data needs to be collected in the particular problem area to collect data, pick data from the identified data for data mining, clean and process data, extract patterns to generate information and finally interpret pattern and generate knowledge.
Data mining is used in medical sciences, malicious executables tracking, sports associations, trend recognition, sales forecasting, basket analysis etc.
There were still numerous unresolved security problems, social concerns, user interface issues, performance issues, and so on before data mining became a conventional, mature and trusted file. While data mining is very efficient, during its implementation it faces numerous challenges. Problems and challenges related to data mining may be efficiency, data, techniques, etc.
When challenges or problems are correctly identified and sorted properly, data mining is successful. We would like to propose potential recommendations for data creation, including the development of efficient and successful data prepared algorithms and systems for single and multiple data sources, taking all internal data into account. And external knowledge. Create the environment for immersive and automated data extraction.

6. References
[1] Roshan Gangurde, Binod Kumar, and SD Gore. Building prediction model using market basket analysis. Int. J. Innov. Res. Comput. Commun. Eng, 5(2):1302–1309, 2017.
[2] Meisam Gordan, Zubaidah Ismail, Zainah Ibrahim, and Huzaifa Hashim. Data mining technology for structural control systems: Concept, development, and comparison. In Recent Trends in Artificial Neural Networks-from Training to Prediction. IntechOpen, 2019.
[3] Manoj Kumar Gupta and Pravin Chandra. A comprehensive survey of data mining. International Journal of Information Technology, pages 1–15, 2020.
[4] Veronika Plotnikova, Marlon Dumas, and Fredrik P Milani. Data mining methodologies in the banking domain: A systematic literature review. In International Conference on Business Informatics Research, pages 104–118. Springer, 2019.
[5] R Ragavi, B Srinithi, and VS Anitha Sofia. Data mining issues and challenges: A review,”. International Journal of Advanced Research in Computer and Communication Engineering, (7):4–7, 2018.
[6] Dimple Tiwari and Manoj Kumar. Social media data mining techniques: A survey. In Information and Communication Technology for Sustainable Development, pages 183–194. Springer, 2020.
[7] Kochetov Vadim. Overview of different approaches to solving problems of data mining. Procedia computer science, 123:234–239, 2018.
[8] Wissam Nazeer Wassouf, Ramez Alkhatib, Kamal Salloum, and Shadi Balloul. Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study. Journal of Big Data, 7:1–24, 2020.
[9] Radhwan HA Alsagheer, Abbas FH Alharan, and Ali SA Al-Haboobi. Popular decision tree algorithms of data mining techniques: A review. International Journal of Computer Science and Mobile Computing, 6(6):133–142, 2017.
[10] Dursun Delen, Enes Eryarsoy, and S.adi S.eker. Introduction to data, text and web mining for business analytics minitrack. In Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.
[11] Sandra Milena Merchan Rubiano and Jorge Alberto Duarte Garcia. Analysis of data mining techniques for constructing a predictive model for academic performance. IEEE Latin America Transactions, 14(6):2783–2788, 2016.
[12] Yang Zhao, Chaobo Zhang, Yiwen Zhang, Zihao Wang, and Junyang Li. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. Energy and Built Environment, 1(2):149–164, 2020.
[13] A Mostafa. Review of data mining concept and its techniques. DOI, 10:207–216, 2016.
[14] Mohini Chakarverti, Nikhil Sharma, and Rajiva Ranjan Divivedi. Prediction analysis techniques of data mining: A review. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), 2019.
[15] P Naga Deepthi, Raju Anitha, and K Swathi. A review on bioinformatics using data mining techniques. In Journal of Physics: Conference Series, volume 1228, page 012023. IOP Publishing, 2019.
[16] Snjeˇzana Kri´zani’c. Educational data mining using cluster analysis and decision tree technique: A case study. International Journal of Engineering Business Management, 12:1847979020908675, 2020.
[17] Anna Doroshenko. Applying artificial neural networks in construction. In E3S Web of Conferences, volume 143, page 01029. EDP Sciences, 2020.
[18] Shiju George and Ajit Kumar Santra. Traffic prediction using multifaceted techniques: A survey. Wireless Personal Communications, pages 1–60, 2020.
[19] Abolfazl Mollalo, Kiara M Rivera, and Behzad Vahedi. Artificial neural network modeling of novel coronavirus (covid-19) incidence rates across the continental united states. International Journal of Environmental Research and Public Health, 17(12):4204, 2020.
[20] Xiaoyu Liu, Zhe Huang, and Bo Tong. Review on the data mining technology and the applications on financial analysis area. In 2016 International Conference on Communication and Electronics Systems (ICCES), pages 1–7. IEEE, 2016.