Chapter

The Concept of Data Mining

Julius Olufemi Ogunleye

Abstract

Data mining is a technique for identifying patterns in large amounts of data and information. Databases, data centers, the internet, and other data storage formats; or data that is dynamically streaming into the network are examples of data sources. This paper provides an overview of the data mining process, as well as its benefits and drawbacks, as well as data mining methodologies and tasks. This study also discusses data mining techniques in terms of their features, benefits, drawbacks, and application areas.

Keywords: Data mining techniques, Data mining process, Regression Analysis, Statistical techniques, Clustering techniques, Neural networks, Nearest Neighbors, Decision trees, Rule induction

1. Introduction

1.1 Data Mining

Data mining may be thought of as a natural progression of information technology. It can be simply defined as a procedure for searching, gathering, filtering, and analyzing data. It's the method of extracting useful knowledge from vast volumes of data kept in databases, data centers, or other data repositories. Database technology, statistics, artificial intelligence, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis are all techniques used in data mining. Data mining allows for the extraction of interesting knowledge, regularities, or high-level information from databases, which can then be viewed or browsed from various perspectives.

It deals with the secondary study of massive databases in order to uncover previously unknown relationships that are of interest or benefit to database owners. It can be thought of as computer-assisted exploratory data analysis of massive, complex data sets from a statistical standpoint. Data mining is having a big effect in business, industry, and science right now. It also opens up a lot of possibilities for new methodological advances in science. New issues occur, partly as a result of the sheer scale of the data sets in question, and partly as a result of pattern matching issues.

Decision making, process control, information management, and query processing are only a few of the applications for the newly discovered experience. As a result, data mining is recognized as one of the most exciting modern database technologies in the information industry, as well as one of the most important frontiers in database systems [1, 2]. This chapter will go into the basics of data mining as well as the data extraction techniques. Mastering this technology and its techniques will have significant advantages as well as a competitive edge.

1.2 The importance of data mining

Data mining has gotten a lot of attention in the information industry in recent years because of the widespread availability of massive quantities of data and the pressing need to transform the data into valuable information and knowledge. Business management, quality control, and market research, as well as engineering design and science discovery, will all benefit from the information and expertise acquired. Governments, private corporations, large organizations, and all industries are interested in collecting a large amount of data for business and research purposes [3, 4]. The following are some of the reasons why data mining is so important:

- Data mining is the process of collecting vast amounts of data in order to extract information and dreams from it. The data industry is currently experiencing rapid growth, which has resulted in increased demand for data analysts and scientists.

- We interpret the data and then translate it into useful information using this technique. This enables an organization to make more accurate and better decisions. Data mining aids in the creation of wise business decisions, the execution of accurate campaigns, the prediction of outcomes, and many other tasks.

- We can evaluate consumer habits and insights with the aid of data mining. This results in a lot of growth and a data-driven business.

It’s important to remember that which data mining approach to utilize is mostly determined by the amount of data accessible, the type of data, and the dimensions. Although there are evident differences in the types of challenges that each data

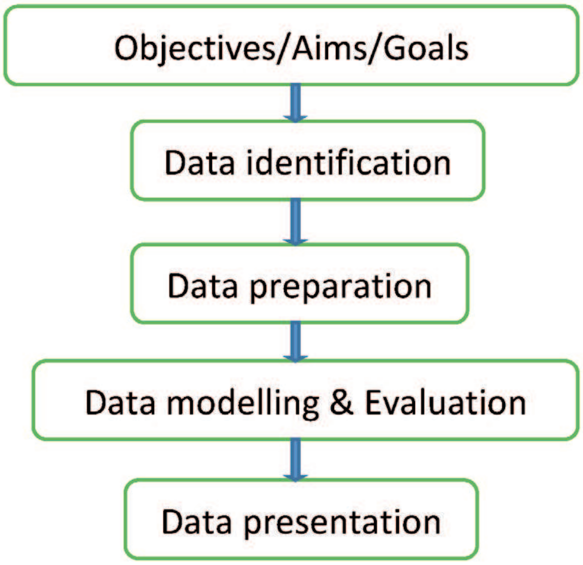

Figure 1.
An overview of data mining process.

mining technique is best suited for, the nature of data from the real world, as well as the complicated ways in which markets, customers, and the data that represents them, means that the data is always altering. As a result, there is no obvious law that favors one technique over another. Decisions are sometimes made depending on the availability of experienced data mining analysts in one or more techniques. The preference of one technique over the others depends more on getting good resources and good analysts (**Figure 1**) [5].

2. Related works

2.1 An overview of efficient data mining techniques

Data mining, according to Sandeep Dhawan, is the act of uncovering relationships within large data sets, as well as data trends, anomalies, changes, and significant statistical structures. Forming a hypothesis and then testing it against the dataset are two common data analysis strategies. Data mining techniques, on the other hand, find significant patterns in data automatically, and these patterns can be utilized to construct algorithms. The result or pattern detected should be genuine, intelligible, and valuable, which is a crucial issue when mining large data sets. It goes without saying that the efficiency of robust and intelligent data mining algorithms is essential for data warehousing and sustaining massive datasets. Data mining techniques are being used in practically every sector of the business world. There is rarely any sphere of life that does not have an input and integration of these data mining tools, from the music industry to film maintenance and medicine to sports. The study gave an overview of some of the most widely used data mining techniques, as well as their applications [3].

2.2 An overview of data mining techniques

The study offered an overview of some of the most widely used data mining algorithms. They were divided into two portions, each with its own theme:

- Statistics, Neighborhoods, and Clustering are examples of traditional techniques.
- Trees, Networks, and Rules: Next-Generation Techniques.

The authors discussed a variety of data mining methods so that the reader may see how each algorithm fits into the larger picture of data mining approaches. There were six different types of data mining algorithms presented in all. The Authors noted that, although there are a number of other algorithms and many variations of the techniques that were described, one of the algorithms is almost always used in real-world deployments of data mining systems [4].

3. Methods

Data mining is a multi-stage process that is accomplished in stages. Data mining is now widely employed in a variety of fields. Data mining has applications and uses in almost every aspect of life, and there are a variety of commercial data mining solutions available today [6–8].

3.1 Data mining process

Data mining is a collaborative effort that includes the following steps:

1. Collecting requirements

   The collection and understanding of requirements is the first step in any data mining project. With the vendor's business viewpoint, data mining analysts or users determine the requirement scope.

2. Data investigation

   This move entails identifying and converting data patterns using data mining statistics. It necessitates collecting, assessing, and investigating the requirement or project. Experts comprehend the issues and challenges and translate them into metadata.

3. Data collection and planning

   For the modeling phase, data mining experts translate the data into meaningful information. They use the ETL (Extract, Transform, and Load) method. They're also in charge of inventing new data attributes. Various methods are employed here to view data in a structural format while preserving the value of data sets.

4. Modeling

   Data experts use their best tools for this phase because it is so important in the overall data processing. To filter the data in an acceptable way, all modeling methods are used. Modeling and assessment are intertwined steps that must be completed at the same time to ensure that the criteria are correct. After the final modeling is completed, the accuracy of the final result can be checked.

5. Assessment or Evaluation

   After efficient modeling, this is the filtering method. If the result is not acceptable, it is then passed back to the model. After a satisfactory result, the requirement is double-checked with the provider to ensure that no details are overlooked. At the end of the process, data mining experts evaluate the entire outcome.

6. Deployment

   This is the final stage in the entire process. Data is presented to vendors in the form of spreadsheets or graphs by experts.

The following functions can be performed with data mining services [9, 10]:

- ***Knowledge extraction:*** This is the procedure for finding useful trends in data that can be used in decision-making [11]. This is because decisions must be made on the basis of correct/accurate data and evidence.

- ***Data collection:*** By scraping through linked websites and databases, it is possible to collect information about investors, portfolios, and funds using the web scraping process.

- ***Web data:*** Web data is notoriously difficult to mine. This is due to the essence of the situation. Web data, for example, can be considered dynamic, meaning it changes over time. As a result, the data mining process should be replicated at regular intervals.

- ***Data pre-processing:*** Typically, the data gathered is stored in a data center. This information must be pre-processed. Data mining experts should manually delete any data that is considered unimportant during pre-processing.

- ***Market research, surveys, and analysis:*** Data mining can be used for product research, surveys, and market research. It is possible to collect data that would be useful in the creation of new marketing strategies and promotions.

- ***Scanning of data:*** Data obtained and processed would be useless until it is scanned. Scanning is essential for detecting trends and similarities in the data.

- ***Customer feedback:*** A company's operations are heavily influenced by customer feedback and suggestions. Customers can easily find the details on forums, journals, and other sites where they can openly express their opinions.

- ***News:*** With nearly all major newspapers and news outlets sharing their news online these days, it is easy to collect information on developments and other important topics. It is possible to be in a better place to compete in the market this way.

- **Up-to-date data:** Keeping data up to date is important. The information gathered would be useless unless it is modified. This is to ensure that the data is valid before making decisions based on it.

- ***Internet research:*** The internet is well-known for its vast amount of knowledge. It is obvious that it is the most important source of data. It is possible to collect a great deal of knowledge about various businesses, consumers, and company clients. Frauds can be detected using online resources.

- ***Study of competitors:*** It's important to know how your competitors are doing in the business world. It is important to understand both their strengths and weaknesses. Their methods of marketing and distribution can be mined, including their methods of reducing overall costs.

3.2 Advantages of data mining

Data mining and its features have many advantages. It raises the need for a data-driven market as it is combined with analytics and big data. Some of the benefits are as follows:

1. Manufacturing industries benefit from data mining by detecting defective devices and goods using engineering data. This aids in the removal of defective goods from the stock list.

2. It assists government agencies in analyzing financial data and transactions in order to model them into usable data.

3. Data mining is useful not only for making forecasts, but also for developing new services and goods.

4. Predictive models are used in the retail sector for products and services. Better quality and consumer insights are possible in retail stores. Historical data is used to calculate discounts and redemption.

5. Data mining aids financial gains and alerts for banks. They create a model based on consumer data that aids in the loan application process, among other things.

6. Customers gain confidence in companies, which leads to an increase in the number of clients.

7. Marketing firms use data mining to create data models and forecasts based on historical data. They manage promotions, marketing strategies, and so on. This leads to fast growth and prosperity.

8. Data mining results in the creation of new revenue sources, resulting in the expansion of the company.

9. Data mining aids in the improvement of strategy and decision-making processes in organizations.

10. When competitive advantages are found, data mining can help reduce production costs.

3.3 Data mining techniques and tasks

Understanding the types of tasks, or problems, that data mining can solve is the best way to learn about it. The majority of data mining tasks can be classified as either prediction or summary at a high level. Predictive tasks allow you to forecast the value of a variable based on previously collected data. Predicting when a customer will leave a business, predicting whether a transaction is fraudulent, and recognizing the best customers to receive direct marketing offers are all examples of predictive tasks. Descriptive tasks, on the other hand, attempt to summarize the information. Automatically segmenting customers based on their similarities and differences, as well as identifying correlations between products in market-basket data, are examples of such tasks [12].

Organizations now have more data at their disposal than they have ever had before. However, due to the sheer volume of data, making sense of the massive amounts of organized and unstructured data to enact organization-wide changes can be exceedingly difficult. This problem, if not properly handled, has the potential to reduce the value of all the data.

Data mining is the method by which businesses look for trends in data to gain insights that are important to their needs. Both business intelligence and data science need it. Organizations may use a variety of data mining strategies to transform raw data into actionable insights [13]. These range from cutting-edge artificial intelligence to the fundamentals of data planning, all of which are critical for getting the most out of data investments.

a. Cleansing and preparing data

b. Pattern Recognition

c. Classification

d. Association

e. Detection of Outliers

f. Clustering

g. Regression

h. Prediction

i. Sequential trends

j. Decision Trees

k. Statistical techniques

l. Visualization

m. Neural Networks

n. Data warehousing

o. Machine Learning and Artificial intelligence

a. Cleansing and preparing data

Cleaning and preparing data is a vital part of the data mining process. Raw data must be cleansed and formatted in order to be useful in various analytic approaches. Different elements of data modeling, transformation, data migration, ETL, ELT, data integration, and aggregation are used in data cleaning and planning. It's a necessary step in determining the best use of data by understanding its basic features and attributes. Cleaning and preparing data has obvious business value. Data is either useless to an entity or inaccurate due to its accuracy if this first phase is not completed. Companies must be able to trust their data, analytics results, and the actions taken as a result of those findings. These measures are also needed for good data quality and data governance.

b. Pattern Recognition

A basic data mining technique is pattern recognition. It entails spotting and tracking trends or patterns in data in order to draw informed conclusions about business outcomes. When a company notices a pattern in sales data, for example, it has a reason to act. If it's determined that a certain product sells better than others for a specific demographic, a company may use this information to develop similar goods or services, or simply better stock the original product for this demographic [14].

c. Classification

The various attributes associated with different types of data are analyzed using classification data mining techniques. Organizations may categorize or classify similar data after identifying the key characteristics of these data types. This is essential for recognizing personally identifiable information that organizations may wish to shield or redact from records, for example.

d. Association

The statistical technique of association is a data mining technique. It denotes that some data (or data-driven events) are linked to other data. It's similar to the machine learning concept of co-occurrence, where the existence of one data-driven event indicates the probability of another. Correlation and association are two statistical concepts that are very similar. This means that data analysis reveals a connection between two data occurrences, such as the fact that hamburger purchases are often followed by French fries purchases.

e. Detecting of Outliers

Outlier detection is used to identify the deviations in datasets. When companies discover anomalies in their records, it becomes easier to understand why they occur and plan for potential events in order to achieve business goals. For example, if there is an increase in the use of transactional systems for credit cards at a certain time of day, businesses can use this information to maximize their income for the day by finding out the cause of it.

f. Clustering

Clustering is an analytics methodology that employs visual approaches to data interpretation. Graphics are used by clustering mechanisms to demonstrate where data distribution is in relation to various metrics. Different colors are used in clustering techniques to represent data distribution. When it comes to cluster analytics, graph-based methods are perfect. Users can visually see how data is distributed and recognize patterns related to their business goals using graphs and clustering in particular.

g. Regression

The essence of the relationship between variables in a dataset can be determined using regression techniques. In some cases, such connections may be causal, and in others, they may only be correlations. Regression is a simple white box technique for revealing the relationships between variables. In areas of forecasting and data modeling, regression methods are used (**Figure 2**).

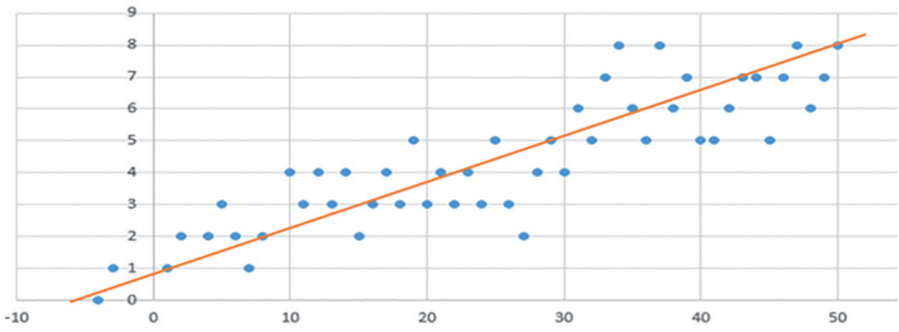

Figure 2.
Illustration example of linear regression on a set of data [15].

h. Prediction

One of the four branches of analytics is prediction, which is a very important feature of data mining. Patterns observed in current or historical data are extended into the future using predictive analytics. As a result, it allows businesses to predict what data patterns will emerge next. Using predictive analytics can take a variety of forms. Machine learning and artificial intelligence are used in some of the more advanced examples. Predictive analytics, on the other hand, does not have to rely on these methods; simpler algorithms can also be used.

i. Sequential Trends

This data mining technique focuses on identifying a sequence of events. It's particularly useful for transactional data mining. For example, when a customer buys a pair of shoes, this technique will show which pieces of clothing they are more likely to buy. Understanding sequential trends may assist businesses in recommending additional products to consumers in order to increase sales.

j. Decision trees

Decision trees are a form of predictive model that enables businesses to mine data more effectively. A decision tree is technically a machine learning technique, but because of its simplicity, it is more often referred to as a white box machine learning technique. Users can see how the data inputs influence the outputs using a decision tree. A random forest is a predictive analytics model that is created by combining different decision tree models. Complicated random forest models are referred to as "black box" machine learning techniques because their outputs are not always easy to comprehend based on their inputs. However, in most cases, this simple form of ensemble modeling is more effective than relying solely on decision trees (**Figure 3**).

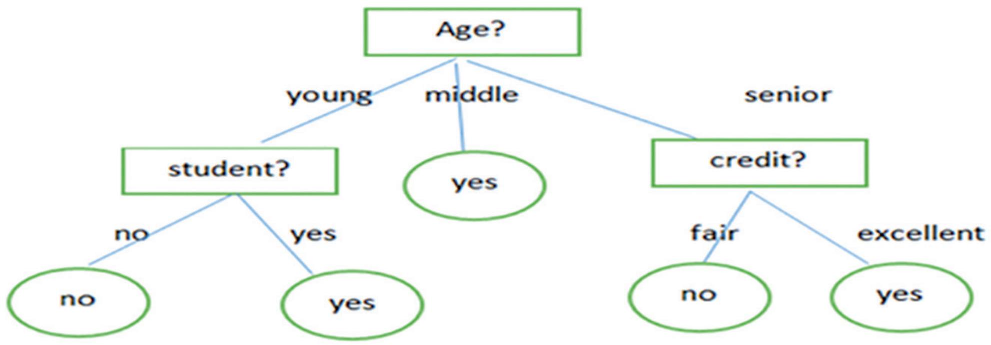

Figure 3.
Example of a decision tree [15].

k. Statistical techniques

Statistical approaches are at the heart of the majority of data mining analytics. The various analytics models are focused on mathematical principles that produce numerical values that can be used to achieve clear business goals. In image recognition systems, neural networks, for example, use complex statistics based on various weights and measures to decide if a picture is a dog or a cat. Statistical models are one of artificial intelligence's two primary branches. Some mathematical methods have static models, while others that use machine learning improve over time.

l. Visualization

Another essential aspect of data mining is data visualization which uses sensory impressions that can be seen to provide users with access to data. Today's data visualizations are interactive, useful for streaming data in real-time, and distinguished by a variety of colors that show various data trends and patterns. Dashboards are a valuable tool for uncovering data mining insights using data visualizations. Instead of relying solely on the numerical results of mathematical models, organizations may create dashboards based on a variety of metrics and use visualizations to visually illustrate trends in data.

m. Neural Networks

A neural network is a type of machine learning model that is frequently used in AI and deep learning applications. Among the most accurate machine learning models used today is neural network. They are named for the fact that they have multiple layers that resemble how neurons function in the human brain. While a neural network can be a powerful tool in data mining, companies should exercise caution when using it because some of these neural network models are extremely complex, making it difficult to understand how a neural network calculated an output (**Figure 4**).

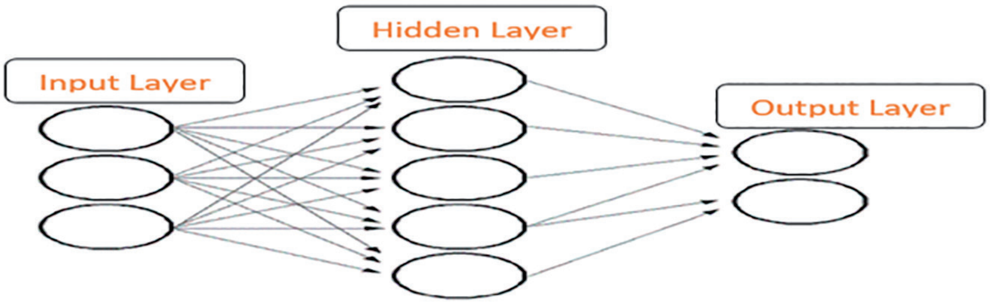

Figure 4.
Example of a neural network [15].

n. Data warehousing

Data warehousing used to imply storing organized data in relational database management systems so that it could be analyzed for business intelligence, reporting, and simple dashboarding. Cloud data centers and data warehouses in semi-structured and unstructured data stores, such as Hadoop, are available today. Although data warehouses have historically been used to store and analyze historical data, many new approaches can now provide in-depth, real-time data analysis.

o. Machine Learning and Artificial Intelligence

Some of the most advanced advances in data mining are machine learning and artificial intelligence (AI). When operating with large amounts of data, advanced machine learning techniques such as deep learning provide extremely accurate predictions. As a result, they can be used to process data in AI applications such as computer vision, speech recognition, and advanced text analytics using Natural Language Processing. These data mining techniques work well with semi-structured and unstructured data to determine meaning.

3.4 Data mining software for optimization

With so many methods to use during data mining, it's important to have the right resources to get the most out of your analytics. For proper implementation, these methods usually necessitate the use of many different tools or a tool with a broad set of capabilities.

While organizations can use data science tools like R, Python, or Knime for machine learning analytics, it's critical to use a data governance tool to ensure compliance and proper data lineage. Additionally, in order to conduct analytics, companies would need to collaborate with repositories such as cloud data stores, as well as dashboards and data visualizations to provide business users with the knowledge they need to comprehend analytics. All of these features are available in tools, but it's critical to find one or more that meet your company's requirements [16].

4. Discussions

4.1 The cloud and data mining's future

The development of data mining has been accelerated by cloud computing technology. Cloud systems are ideally adapted for today's high-speed, massive amounts of semi-structured and unstructured data that most businesses must contend with. The elastic capabilities of the cloud will easily scale to meet these big data demands. As a result, since the cloud can carry more data in a variety of formats, more data mining techniques are needed to transform the data into insight. Advanced data mining techniques such as AI and deep learning are now available as cloud services.

Future advancements in cloud computing would undoubtedly increase the need for more powerful data mining software. AI and machine learning will become even more commonplace in the next five years than they are now. The cloud is the most suitable way to both store and process data for business value, given the exponentially growing pace of data growth on a daily basis. As a result, data mining methods can depend much more on the cloud than they do now.

Currently, data scientists use a variety of data mining techniques, which differ in precision, efficiency, and the type and/or volume of data available for analysis. Classical and modern data mining techniques are two types of data mining techniques. Statistical approaches, Nearest Neighbors, Clustering, and Regression Analysis are examples of Classical techniques, while Modern techniques include Neural Networks, Rule Induction Systems, and Decision Trees.

4.2 Data mining techniques – advantages and disadvantages

1. Statistical Techniques

Advantages

- Because secondary data is normally inexpensive and requires less time to compile because it has already been done.

- Because the patterns and correlations are obvious and reliable.

- Broad samples were used to ensure high generalizability.

- It can be used several times to test various variables.

- It is possible to assess changes that enhance reliability and representativeness.

Disadvantages

- The researcher is only able to draw patterns and correlations from the data and cannot assess the validity or consider a causal theory process.

- Statistical data is often secondary data, making misinterpretation simple.

- Statistical evidence can be manipulated, and it can be skewed and phrased to support the researcher's point (effects objectivity).

- It's difficult to view and validate this data because it's always secondary.

2. Nearest Neighbors

Advantages

- It's easy and intuitive to use

- It does not make any assumptions

- No Education Transfer

- It is constantly growing.

- It is a simple multi-class issue to enforce.

- Regression and classification are also possible applications.

- Selecting the first hyper parameter can take some time, but once done, the rest of the parameters are compatible with it.

- Provides a variety of distance parameters to choose from (Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance).

Disadvantages

- Irrelevant characteristics can influence the distance between neighbors.
- While the implementation can be simple, the efficiency (or speed of the algorithm) decreases rapidly as the dataset grows.
- Can handle a small number of input variables, but as the number of variables increases, the algorithm has trouble predicting the performance of new data points.
- Characteristics must be consistent.
- When classifying new data, the problem of determining the optimal number of neighbors to consider is frequently encountered.
- Issues with using data that is unbalanced.
- It is vulnerable to outliers since neighbors are clearly selected based on distance parameters.

3. Clustering

Advantages

- Hierarchical methods enable the end user to choose from a large number of clusters or a small number of clusters.
- Appropriate for data sets of any form and attributes of any kind.
- There are a variety of well-developed models that provide a way to accurately represent the data, and each model has its own unique characteristics that can provide significant advantages in some specific areas.

Disadvantages

- Cluster numbers must be preset.
- The assumption is not completely right, and the clustering result is dependent on the parameters of the chosen models.

4. Regression Analysis *(MARS- Multivariate Analysis for Regression Splines, OLS - Ordinary Least Square regression, SVR-Support Vector Regression, Radial Basis Function Networks)*

Advantages

- Linear regression can solve some very simple problems much faster and more easily, since prediction is simply a multiple of the predictors.
- Linear regression: the modeling process is simple, requires few calculations, and runs quickly even when the data is large.

- Linear regression: the factor can provide insight into and interpretation of each variable.
- In linearly separable datasets, linear regression works well.
- Linear regression is easier to implement, evaluate, and apply than other methods.
- In linear regression, dimensionality reduction, regularization (L1 and L2), and cross-validation methods can all be used to prevent over-fitting.
- Multiple regression will assess the relative importance of one or more predictor variables in determining the criterion's significance.
- Outliers, or deviations, can be found using multiple regression.

Disadvantages

- Linear regression: There is a minimum linear association.
- Linear regression: Outliers are affected easily.
- The regression solution would most likely be thick (because there is no regularization)
- Linear regression is vulnerable to noise and overfitting.
- Regression solutions obtained through a variety of approaches (e.g., optimization, less-square, QR decomposition, etc.) are not necessarily unique.
- Vulnerable to multicollinearity: Multicollinearity should be eliminated (using dimensionality reduction techniques) before using linear regression since it means that the independent variables have no relationship.
- Any disadvantage of using a multiple regression model is usually due to the data used, either because there is insufficient data or because the cause is incorrectly assumed to be a correlation.

5. Neural Networks

Advantages

- Artificial Neural Networks (ANN) will model and analyze nonlinear, complex relationships.
- Has highly accurate statistical models that can be used to solve a wide range of problems.
- Information is stored on the network as a whole, not in a database, and the network will run even though a few pieces of information are missing from one location.

- The ability to work with limited knowledge
- Has fault tolerance, which means that contamination of one or more ANN cells will not stop development.
- Is endowed with a memory.
- Gradual corruption: As a network ages, it slows down and becomes more vulnerable. The network issue does not seem to be corroding right away.
- Machine-training capability: Artificial neural networks learn events by observing and reflecting on similar events.
- Parallel processing capability: Artificial neural networks have the computing capacity to perform several tasks at the same time.

Disadvantages

- Extraction of features-the issue of determining which predictors are the most suitable and significant in building models that are predictably accurate.
- Hardware reliance: Artificial neural networks, by their very nature, require parallel processing processors. This is the foundation on which the equipment realization is built.
- Assurance of proper network structure: When it comes to artificial neural network design, there are no hard and fast rules. The correct network design is achieved by practice and trial and error.
- Network behavior that is not explained: Even though ANN provides a sampling solution, it does not explain why or how it works.
- Difficulty in demonstrating the problem to the network: ANNs should deal with numerical data. Before integrating into ANN, problems must be translated into numerical values.
- The network's length is unknown: reducing the network to a certain sample error value indicates that the training is complete. This may not result in optimal results.

6. Rule Induction

Advantages

- When dealing with a small number of rules, IF-THEN rules are easy to understand and are meant to be the most interpretable model.
- The decision rules are just as descriptive as decision trees, but they are a lot smaller.

- Since only certain conditional statements must be checked to determine the rules apply, IF-THEN rules are simple to predict.

- Since conditions only shift at the threshold, decision rules will withstand monotonous input function transformations.

- IF-THEN rules produce models with few features. Only the features that are important to the model are chosen.

- Simple rules like OneR can be used to test more complex algorithms.

Disadvantages

- IF-THEN laws are mostly concerned with grouping and almost completely neglect regression.

- Categorical functions are also needed. This means that numerical features must be classified if they are to be included.

- The majority of older rule-learning algorithms are prone to overfitting.

- In the study of linear feature-output relations, decision rules are ineffective.

7. Decision Trees *(CART – Classification and Regression Trees)*

Advantages

- Data is organized into distinct categories, which are therefore simpler to grasp than points on a multidimensional hyperplane, as in linear regression. With its nodes and edges, the tree structure has a natural visualization.

- In real-world problems, the models to be built and the interactions to be detected are usually much more complex.

- CART validates the Tree immediately, implying that the algorithm has the model validation and discovery of the optimally general model (the algorithm) built deep inside it.

- When it comes to missing data, the CART algorithm is fairly reliable.

- There are so many powerful data mining features that decision trees mark so strongly.

Disadvantages

- Can struggle with some very simple problems where prediction is simply a multiple of predictors.

- Trees are incapable of handling linear relationships. Splits must be used to approximate any linear input–output relationship, resulting in a phase function. This is not going to work.

- It had a silky feel to it. Small changes in the input function may have a big effect on the forecast outcome, which is not always a good thing.

- The trees are still very shaky. A few tweaks to the training dataset will result in a completely different tree. Since every split is based on splitting the parent, this is the case.

These methods are best applied to particular tasks in order to achieve the best performance. The **Table 1** below lists the data mining tasks and the techniques that can be used to complete them.

A business analyst's dream is data warehousing. All of the data concerning the organization's actions is centralized and accessible through a single set of analytical tools. A data warehouse system's goal is to give decision-makers the accurate, timely data they need to make the best decisions possible. A relational database management system server serves as the central repository for informational data in the data warehouse architecture. The processing of operational data is kept distinct from the processing of data warehouse data.

The central information repository is surrounded by a number of critical components that work together to make the overall ecosystem functional, manageable, and available to both operational systems and end-user query and analysis tools. The warehouse's raw data is often derived from operational applications. Data is cleansed and turned into an integrated structure and format when it enters the warehouse. Conversion, summarization, filtering, and condensing of data may all be part of the transformation process. Since the data contains a historical component, the warehouse must be capable of holding and managing large volumes of data as well as different data structures for the same database over time.

No	Data mining task	Data mining techniques
1	Classification	Decision trees, Neural networks, K-nearest neighbors, Rule induction methods, SVM-Support vector machine, CBR-Case based reasoning
2	Prediction	Neural networks, K-nearest neighbors, Regression Analysis
3	Dependency Analysis	Correlation analysis, Regression Analysis, Association rules, Bayesian networks, Rule Induction
4	Data description and summarization	Statistical techniques, OLAP (Online Analytical Processing)
5	Segmentation or clustering	Clustering techniques, Neural Networks
6	Consolidation	Nearest neighbors, Clustering

Table 1.
Data mining tasks and the methods used to accomplish them.

The following **Table 2** lists data mining techniques and their areas of applications.

Data mining techniques	Areas of use
Association Analysis	Designing store shelves, marketing, cross-selling of products.
Classification (K-nearest neighbor, etc.)	Banks, marketing campaign designs by organizations.
Decision Trees	Medicine, engineering, manufacturing, and astronomy, to name a few fields. They were used to solve problems ranging from credit card depletion estimation to time series exchange rate estimation for a variety of international currencies.
Clustering Analysis	Image recognition, web search, and security.
Outlier Detection	Detection of credit card fraud risks, novelty detection, etc.
Regression Analysis (K-nearest neighbor, ...)	Marketing and Product Development Efforts comparison.
Artificial Neural networks	Data compression, feature extraction, clustering, prototype formation, function approximation or regression analysis (including prediction time series, fitness approximation, and modeling), classification (including pattern and sequence recognition, novelty detection, and sequential decision making), data processing (including filtering, clustering, blind source separation, and compression), and robotic compression.
Support vector machines regression	Oil and gas industry, classification of images and text and hypertext categorization.
Multivariate Regression algorithm	Retail sector
Linear Regression	Financial portfolio prediction, salary forecasting, real estate predictions and in traffic estimated time of arrivals (ETAs).

Table 2.
Data mining techniques and their areas use.

5. Conclusion

It's worthy of note to state that time is spent on extracting useful information from data. As a result, in order for companies to develop quickly, it is necessary to make accurate and timely decisions that enable them to take advantage of available opportunities. In today's world of technology trends, data mining is a rapidly growing industry. In order to obtain valuable and reliable information, everyone today needs data to be used in the right way and with the right approach. Data mining can be initiated by gaining access to the appropriate resources. Since data mining begins immediately after data ingestion, finding data preparation tools that support the various data structures required for data mining analytics is important. Organizations may also want to identify data in order to use the aforementioned methods to explore it. Modern data warehousing, as well as various predictive and machine learning/AI techniques, are helpful in this regard.

Choosing which approach to employ, and when, is clearly one of the most difficult aspects of implementing a data mining process. Some of the parameters that are critical in deciding the technique to be used are determined by trial and error. There are clear differences in the types of problems that each data mining technique is best suited for. As a result, there is no simple rule that favors one technique over another.

Decisions are often taken based on the availability of qualified data mining analysts in one or more techniques. The choice of a technique over the other is more dependent on the availability of good resources and analysts.

Acknowledgements

This work was supported by the Faculty of Applied Informatics, Tomas Bata University in Zlín, Czech Republic, under Projects IGA/CebiaTech/2021/001.

Author details

Julius Olufemi Ogunleye
Tomas Bata University in Zlin, Zlín, Czech Republic

*Address all correspondence to: juliusolufemi@yahoo.com

References

[1] Software Testing Help (April 16, 2020): Data Mining Techniques: Algorithm, Methods & Top Data Mining Tools.

[2] Silhavy, P., Silhavy, R., & Prokopova, Z. (2019): Categorical variable segmentation model for software development effort estimation. IEEE Access, 7, 9618-9626.

[3] Sandeep Dhawan (2014): An Overview of Efficient Data Mining Techniques.

[4] Alex Berson et al. (2005): An Overview of Data Mining Techniques.

[5] Jiawei H. and Micheline K. (2000): Data Mining: Concepts and Techniques.

[6] ACM SIGKDD (2006-04-30), Retrieved (2014-01-27): Data Mining Curriculum.

[7] Kamber H. Et al. (2011): Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.

[8] Clifton C. (2010): Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[9] Weiss G. M. and Davison B. D. (2010): Data Mining (Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010).

[10] Mehmed K. (2011): Data Mining Concepts, Models, Methods, and Algorithms (Second Edition).

[11] Berson A. et.al (2005): An Overview of Data Mining Techniques (Excerpts from the book by Alex Berson, Stephen Smith, and Kurt Thearling).

[12] Karna H. et al. (2018): Application of data mining methods for effort estimation of software projects.

[13] Sehra S.K. et al. (2014): Analysis of Data Mining techniques for software effort estimation.

[14] Trevor H. Et al.(2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Archived from the original on 2009-11-10. Retrieved 2012-08-07.

[15] Ogunleye J.O. (2020): Review of Data Mining Techniques in Software Effort Estimation.

[16] Dejaeger K., et al.(2012): Data Mining Techniques for Software Effort Estimation: A Comparative Study.