International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211
Volume 9, Issue 8, August -2021, Impact Factor: 7.429, Available online at: www.ijaresm.com

A Study of Data Mining and Data Science

Megha Sagar Patil[1], Vijay Bapuso Pujari[2], Sairaj Satish Suryavanshi[3], Satyajeet Sanjay Mandale[4]

[1,2,3,4]BCA Department, Vivekanand College (Autonomous) Kolhapur Maharashtra, India

----------**********----------

ABSTRACT

Data mining is about finding trends in data sets and using these trends to determine future patterns. This is an important step in the discovery process. This usually involves analyzing large amounts of historical data that were previously ignored. A research field, including big data analysis, data mining, predictive modeling, data visualization, mathematics, and statistics. Data science is called the fourth paradigm of science. (The other three are theoretical, empirical, and computational). The college often conducts exclusive research in the field of data science.. Data science, also known as data-driven science, is a broad field or field that includes methods for extracting, analyzing, and extracting information from data. Data mining is also called data discovery. It is a method and technique involving data analysis. The goal is to discover useful information in the data set and use it to discover the patterns covered.

Keywords: Data, Data Mining, Information, Science, technology, applications etc.

INTRODUCTION

Data mining and data science are the two most important topics in technology. Both are related to data, but they are used in different ways. In addition, the skills required to perform operations in these areas are also different, so we understand the concepts of these two areas and analyze their main differences. Data science is a field or field that involves processing large amounts of data and using it to create predictive, normative, and normative analysis models. It is about mining, collecting (building models), analyzing (validating models), and using data (showing the best model).

This is the intersection of data and calculation. It is a combination of computer science, economics, and statistics.

Data mining is a method of extracting key and important information and knowledge from a large amount of data/library. It extracts information through rigorous mining, analysis, and processing of big data to identify patterns and correlations that may be important to the organization. This is similar to gold mining, which is extracted from rocks and sand.

WHAT IS DATA MINING?

Mining just extracts valuable minerals. In the 21st century, dates are the most expensive mineral. In order to extract useful data from a specific set of raw data, we use data mining. Through data mining, we extract useful information from a given data set to identify patterns and relationships. The data mining process is a complex process involving intensive data storage and powerful computer technology. In addition, data mining is not limited to extracting data, it is also used to transform, clean, integrate, and analyze data patterns. Another term for data mining is knowledge discovery. has several important data mining options, such as association rules, sorting, clustering, and prediction. Some key features of data mining:

- Forecast patterns based on data trends.
- Calculation of profit forecast.
- Create information in response to the analysis.
- focuses on larger databases.
- Grouped visualization data

DATA MINING STEPS

Knowledge discovery is an important part of data mining. The most important steps in data mining are:

Step 1: Clean up the data-This step cleans up the data so that it is not affected by noise or interference.

Step 2: Data Integration-In data integration, we combine multiple data sources into one.

Step 3: Get data: In this step, we get data from the database.

Step 4: Transform the data: In this step, we transform the data to perform pivot analysis and aggregation operations.

Step 5: Data Mining: In this step, we extract useful data from the existing data set.

Step 6: Evaluation Mode: We analyze the different patterns that exist in the data.

Step 7: Knowledge display: In the last step, we display knowledge to users in the form of trees, tables, graphs, and matrices.

DATA MINING APPLICATIONS
There are various data mining applications such as:

- Market and stock analysis
- Fraud detection
- Risk management and business analysis
- Total customer value analysis

DATA MINING TOOLS
Some of the most popular data mining tools:

RapidMiner
This is one of the most popular data mining tools. It is written in Java but does not require any coding to work. In addition, it also provides various data mining functions, such as data processing. , Data presentation, filtering, grouping, etc.

Weka
Weka is an open-source data mining software developed by the University of Wichita. Like RapidMiner, it has an easy-to-use graphical interface without programming.

With Weka, you can directly call machine learning algorithms or import them using Java code. Provides multiple tools such as preview, preprocessing, sorting, grouping, etc.

KNime
KNime is a powerful data mining package, mainly used for data preprocessing, namely ETL: extraction, transformation, and loading. In addition, it also integrates various components of machine learning and data mining, providing an integrated platform for all related operations.

Apache Mahout
Apache Mahout is an extension of the Hadoop big data platform. Apache developers created Mahout to meet the growing demand for data mining and analysis operations in Hadoop. Therefore, it includes various machine learning functions such as classification, regression, grouping, etc., and programs.

TeraData
Data transmission requires storage. TeraData, also known as TeraData database, provides storage services composed of data mining tools. It can store data according to usage, that is, store infrequently accessed data in its "slow" part to provide fast access to frequently accessed data.

Orange
Orange software is known for integrating machine learning and data mining tools. It is written in Python and provides users with interactive and beautiful visualization.

Oracle DataMining
Oracle Data Mining is an excellent data classification, analysis, and prediction tool that enables users to perform data

mining on their SQL database to extract views and patterns.

WHAT IS DATA SCIENCE?

Data science is one of the hottest challenges of the 21st century. "Harvard Business Review" called it "the sexiest job of the 21st century." In recent years, it has become a buzzword and has become more and more popular. The advent of advanced computer technology has greatly increased the amount of data. Companies need to analyze data and obtain meaningful information. This special position is for data scientists who are familiar with statistics and calculation tools. With the knowledge of machine learning, data scientists can predict future events.

Therefore, data science is a broad subject that covers various data operations, such as data extraction, data processing, data analysis, and data prediction. Data science originated in many disciplines, such as mathematics, statistics, and computer programming. The industry needs data scientists to help them make effective data-driven decisions. There are many positions in data science. This is because data is everywhere, exponentially growing, and requires analysis.

DATA SCIENCE STEPS

Here are the 5 steps in data science-

Step 1. Extract the data. The first step in data science is data recovery. The recovered data can be in the form of structured and unstructured data. Several databases support these data. Recovery queries, such as SQL and NoSQL.

Step 2: Data preprocessing-This step includes cleaning the data, transforming the data, and replacing any missing values. This is the most important step because it can organize the data and make it useful for further analysis.

Step 3: Data analysis-Data analysis involves the use of various statistical techniques, such as logical statistics and descriptive statistics, to discover patterns and trends in data.

Step 4. Make predictions. The next important step is to use machine learning algorithms to make predictions. There are several types of predictions and rankings performed on historical data to predict future events and identify patterns in the data.

Step 5: Optimize the model. The last step is to optimize the machine learning model to improve its performance and obtain accurate results.

I. DATA SCIENCE TOOLS

Some of the main tools used in data science:

1. Python: Python is the most popular data science and software development programming language, and provides a wide range of libraries that support data manipulation.

2. R-R is an open-source statistical programming language that provides multiple software packages that can visualize and analyze data.

3. SAS-SAS stands for Statistical Analysis System. It is a software package developed by SAS to promote various statistical operations. Because of its stability and reliability, it is a proprietary closed source tool of choice for many companies.

4. Apache Spark: Apache Spark is an advanced big data tool that provides data processing and analysis functions. It is known for its ability to stream rather than batch process legacy platforms.

5. D3.js-D3.js is a Javascript-based library for creating interactive visualizations. This tool allows you to embed beautiful graphics in web applications.

6. Tableau-Tableau is visualization software for creating interactive charts and graphs. It can interact with OLAP, spreadsheets, and SQL databases. In addition, Tableau can display latitude and longitude on the map.

7. TensorFlow-TensorFlow is a powerful machine learning library for implementing deep learning algorithms. It is a fast rendering library that supports graphics processing units (GPU).

DATA MINING VS DATA SCIENCE

Data science is a collection of data operations, including data mining.

- Data scientists are responsible for developing information products for the industry. On the other hand, data mining is responsible for extracting useful data from other unnecessary information.
- Although data science is a quantitative field, data mining is limited to business roles that need to mine specific information.
- Data scientists must perform multiple operations, such as analyzing data, developing predictive models, and discovering hidden patterns. In contrast, data mining involves statistical modeling to generate useful information.
- Data scientists have to deal with structured and unstructured data, and data mining only applies to structured information.

CONCLUSION

The term data mining and data analysis has been around for a long time. Success requires data mining and data analysis. No matter where you work, there is no denying that both are important in data management in the 21st century. Some user groups use them interchangeably, while other user groups make a clear distinction between these two areas. Data mining is usually part of data analysis, and its goal or intent is still to simply define or determine the structure of the data set. On the other hand, data analysis is a complete package for understanding the database, and data mining may or may not involve it. Both fields require different skills, abilities, and experience. In the next few years, these two fields will have considerable data, resources, and work requirements.

ACKNOWLEDGMENT

Authors are thankful to faculty of Bca DepartmentVivekanad College Kolhapur for motivating us also all the friends for motivating us to write this paper. All the references are hereby acknowledged used in this paper.

REFERENCES

[1] https://www.jigsawacademy.com/blogs/data-science/data-mining-vs-data-analysis/
[2] https://data-flair.training/blogs/data-mining-and-data-science/
[3] https://www.geeksforgeeks.org/
[4] https://www.upgrad.com/blog/
[5] https://en.wikipedia.org/wiki/Data_science#:~:text=Data%20science%20is%20related%20to,analyze%20actual%20phenomena%22%20with%20data.&text=However%2C%20data%20science%20is%20different%20from%20computer%20science%20and%20information%20science.
[6] https://en.wikipedia.org/wiki/Data_mining