iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33790

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Data Mining

Anubhav Sharma[1], Simple Sharma[2]
[1]*Student, Associate Professor, Department of Computer Science Engineering, Faculty of Engineering and Technology*
[2]*Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India*

I. INTRODUCTION

Data mining can be defined as the process used to extract usable data or the data we need on the largest set of any raw data. Basically, data mining is a way of analyzing data patterns with large data sets stored in a system or anything, using one or more software.
Data mining in many areas is also called Knowledge Discovery in Data (KDD).
There are many descriptions of data mining out there. According to the individual mine data the data can have a unique meaning.
Data mining can also be described as a process of large-scale data analysis to acquire business intelligence that helps companies solve problems, reduce risks, and seize new opportunities. It is very similar to searching for important information in a large database and digging a mountain of ore.

II. KDD- KNOWLEDGE DISCOVERY IN DATABASES

KDD basically refers to a large process of obtaining information from data and demonstrating the use of specific data mining techniques. It is a field of interest for researchers in a variety of fields including machine learning, artificial intelligence, pattern recognition, mathematics, data, data recognition and much more.
The main purpose of the KDD process is to extract facts or data from data in the context of big data. It does this by using data mining algorithms to identify what is considered information.

A. KDD Process

The KDD process is repetitive and interactive; it has nine steps. The process is repeated at each stage; means that a return to previous actions may be required. The process has hundreds of aspects to think about in the sense that one cannot present a single formula or perform a complete scientific class with the appropriate decision for each step. Therefore, it is necessary to understand the process and all the different types of needs and opportunities at all stages.

The following is a description of the nine-step process for KDD: -

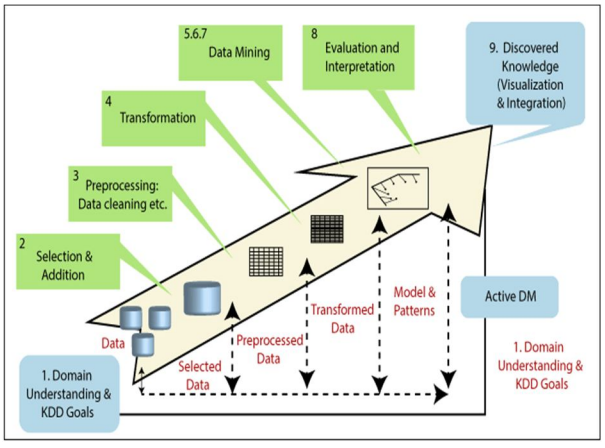


Figure 1: NINE-STEP KDD Process

The process begins with the determination of objectives in KDD and ends with the implementation of the information obtained. At that point, the loop is closed, and the active data mine, begins. After that, changes will need to be made in the application area. For example, offering a variety of features to mobile phone users to reduce bubble. This eventually closes the loop, and the effects that are made are measured in the newly created data caches, as well as the KDD process as well.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

III. DATA MINING TECHNIQUES

There are 7 Data Mining techniques

A. Classification

This method is used to retrieve important and relevant information about data, as well as metadata. This particular data mining method helps us to classify data into different types of classes.

B. Clustering

This data mining method is used to identify data that is similar to each other. This process helps us to understand the differences and similarities between different types of data.

C. Regression

Undoing is a method of digging data to identify and analyze the relationships between different variables. It is used to indicate the tendency for certain variations in the presence of other variables.

D. Outer

This type of data mining strategy refers to the testing of different types of data objects on a particular database that often differ from the expected pattern. This method can be used in a variety of fields, such as discovery, fraud or error detection, intrusion, etc. External acquisition is also known as Outlier Analysis or Outlier mining in many respects.

E. Sequential Patterns

Sequential patterns is a data mining process that helps to identify related patterns or patterns of transaction data over a period of time.

F. Prediction

Guessing a different type of data mining techniques using a combination of other data mining methods such as regression, sequential patterns, integration, separation, etc. It basically analyzes all past events or conditions in the correct order to predict the future event.

G. Association Rules

This method of data mining helps us to find relationships or connections between two or more objects. Detects a hidden pattern in a specific data set. The organizational governance process has many functions or functions and is often used to facilitate sales integration into data or medical data sets.

IV. DATA MINING ALGORITHMS

A. C4.5 Algorithm

C4.5 is an advanced data mining algorithm developed by Ross Quinlan in the early 20s. C4.5 is used to improve the distinction in the form of a decision tree from a set of data that has already been split. Classifier is basically a data mining tool that takes the data we need to split or share it and try to anticipate a new data category. Each data point will have its own symbols. Decision trees are always easy to understand and explain what makes C4.5 so fast and popular as compared to the other data mining algorithms out there.

B. K-means Algorithm

K-methods builds the amount of groups from a set of data or objects based on similarities between that data or objects. It cannot be assumed that the members of the group will be exactly the same, but those members of the group will be more similar than those who are not members of the group. For general fulfillment, k-means is an algorithm for free reading as it reads the collection itself without external data.

C. Support Vector Machines

The vector support (SVM) machine works almost identical to the C4.5 algorithm. The only difference is that SVM does not use decision trees at all. Basically, SVM learns from data sets and specifies hyperplane to organize data into two phases. Hyperplane is the number line “y = mx + c”. SVM expands it to predict your data at higher rates. Once predicted, SVM describes the leading hyperplane for dividing data into two categories.

D. EM Algorithm

EM is known as Expectation-Maximization. Similar to the k-means algorithm, it is also used as an integration algorithm for obtaining information. The EM algorithm works repeatedly to improve the chances of seeing the data obtained. In the next step, we measure the parameters of the mathematical model by the non-available variables, thus producing the obtained data. The Expectation-Maximization (EM) algorithm is also an algorithm for free learning.

E. Naive Bayes Algorithm

Naive Bayes seems to work well as a single algorithm. Naive Bayes is a collection of programs organized together. The assumption used by a group of algorithms is that all the elements of the split data are independent of all the other elements present in the class. The Naive Bayes is made up of a training database with a table-building label. Therefore, it was developed as a supervised learning algorithm.

F. Apriori Algorithm

The Apriori algorithm works by studying the rules of the organization. After learning the organization's rules, the apriori algorithm is applied to a database that contains a large number of transactions. This algorithm is used to find interesting patterns and interrelated relationships. Therefore, it is treated as a way to learn freely. Although this algorithm works very well, it consumes a lot of memory and consumes a lot of disk space and consumes a lot of time.

G. PageRank Algorithm

PageRank is often used by search engines such as Google, Internet Explorer etc. It is a link analysis algorithm that identifies the relative importance of an object associated with a network of objects. Link analysis is similar to network analysis that examines organizations between objects. Google's search engine uses this algorithm by considering the background links between different web pages.

H. AdaBoost Algorithm

This algorithm is a type of promotion algorithm used to create distinctions. Promoting an algorithm as a whole is a learning algorithm that uses many learning algorithms and integrates them.
Strengthening algorithms takes a family of weak students and combines them to form one strong student. Weak reader analyzes data with minimal accuracy. Adaboost is a fully managed learning algorithm as it works on repetition and repetition, targeting weak students and labeled databases. Adaboost is a simple and straightforward algorithm to be implemented.

I. kNN Algorithm

KNN is a lethal learning algorithm used as a partition algorithm. A loose student will not do much during training. But that student will be storing training data. Lethargic students begin to be segregated only when new non-labeled data is provided as input. C4.5, SVN and Adaboost, on the other hand, are aspiring students who often begin to model for differentiation during self-training. Just because kNN is provided with a training database with a label, it is treated as a managed learning algorithm.

J. CART Algorithm

The full CART form says; Separating From Trees. It is a learning algorithm for deciduous trees that provides trees for deceleration or separation as extraction. In the CART algorithm, the node of the decision tree will have 2 direct branches. THE CAR also divides, similar to C4.5. The split tree model is designed using a labeled training database provided by the user. Therefore, it is also treated as a managed learning process.

V. CONCLUSION

Data mining carries together distinct methods from a variations of methods which includes machine learning, data visualization, statistics, database management and others. These techniques can be made to work together to undertake compound problems. Usually, data mining software or systems make use of one or more than one of these methods to compromise with distinct application areas, types of data, data requirements, and mining tasks.

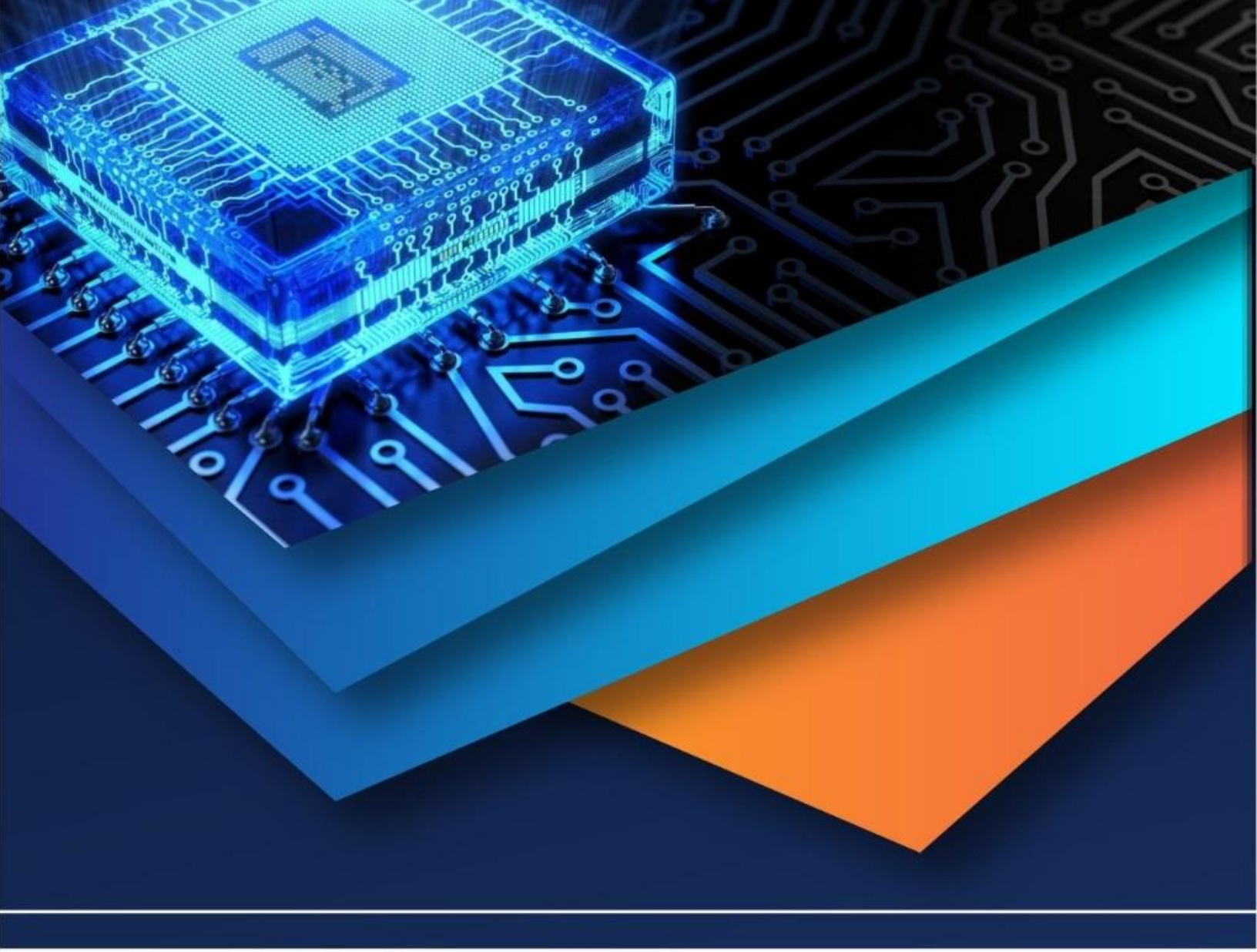
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

VI. ACKNOWLEDGMENT

I would like to thank my teacher Mrs. Simple Sharma who gave me the golden opportunity to do this great work on the topic Data Mining, who also helped me with a lot of research and learning many new things.
Second, I would also like to thank my parents and friends who helped me get the job done on time.

REFERENCES

[1] Jerome H. Friedman, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, CA 94305
[2] D. Hand, American Statistician, 52(2):112-118
[3] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, Knowl Inf Syst (2008) 141-37.
[4] U. Fayyad, G. Piatesky-Shapiro & P. Smyth, AI Magazine, 17(3):37-54, Fall 1996.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY