**Data Mining Techniques/Algorithms**

Data mining refers to filtering, sorting, and classifying data from larger datasets to reveal subtle patterns and relationships, which help enterprises identify and solve complex business problems through data analysis. Data mining software tools and techniques allow organizations to foresee future market trends and make business-critical decisions at crucial times. Various techniques used in data mining are:

1. Association Rules

2. Clustering

3. Classification

4. Regression

5. Anomaly detection

6. Prediction

**Association rules:** It is a data mining technique that is trying to find the relationship of common patterns of the objects within a data set. Which products are sold together? Which DNA profiles are sensitive to new drug? For example, while tracking a customer’s behavior when purchasing online items, an observation is made that the customer generally buys cookies when purchasing a coffee pack. In such a case, the association rule establishes the relation between two items of cookies and coffee packs, thereby forecasting future buys whenever the customer adds the coffee pack to the shopping cart. With this technique, questions are proposed. In this model, significant dependencies between variables are defined. Though it is a very simple method to be used but it is capable of providing a lot of insight and information related to the day to day business. This information can be used to generate the required revenue and even improve the efficiency of the business. There are far fledged applications related to this method which can help various industries and business to increase their value. Here are some examples: Up-selling and cross-selling of products, physical organization of items, network analysis, and marketing and management. E,g. in the nappy and beer model. This method was used for many years in the industry for the market basket analysis but now new recommendations have been made by the engineers, which have overpowered the traditional methods.

**Clustering:** Clustering data mining techniques group data elements into clusters that share common characteristics. We can cluster data pieces into categories by simply identifying one or more attributes. Some of the well-known clustering techniques are k-means clustering, hierarchical clustering, and Gaussian mixture models.

Clustering is an important technique through which object grouping can be done (like the different groups of customers). The objects belonging to the same cluster are similar but those which are in the different groups are different. In this descriptive task a finite set of clusters are determined which identify or describe the data. The process of clustering can be defined in such a way that if you have a group of data points which have attributes of their own and have some kind of similarity then they should be clustered in such a way that the data points in that cluster are much alike each other. Data points in separate clusters are likely to be dissimilar to one another. To find how close or far one cluster is from the another, we can use the Euclidean distance, which can be applied only if attributes are continuous or other similarity measures that is relevant to the specific problem. A useful application of clustering is marketing segmentation, in which distinct set of customers are made in the market and distinct marketing strategies are applied to each of the subsets. It is possible to do this by analysing the lifestyle related and geographical information of each customer and make their clusters. This will help in finding out the clustering quality of the customers by observing the difference in the buying patterns of the customers in one cluster to the customers in the other cluster. Clustering is important for grouping of similar objects in database. For example, determination of different customer profile in markets, while making city planning determination of house prices according to the geographical grouping. Clustering is unsupervised learning.

**Classification**: Identifies which class the incoming information belongs to. Before digging into the hectic modelling phase of the analysis of data the primary step we have to take is classification. This classifies the data item in anyone of the predefined classes. For example, we can classify vehicles into different categories, such as sedan, hatchback, petrol, diesel, electric vehicle, etc., based on attributes such as the vehicle’s shape, wheel type, or even number of seats. When a new vehicle arrives, we can categorize it into various classes depending on the identified vehicle attributes. One can apply the same classification strategy to classify customers based on their age, address, purchase history, and social group.

Some of the examples of classification methods include[decision trees](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-decision-tree/)**,**Naive Bayes classifiers, logistic regression, and so on.

Assume you have a set of records which have their own set of attributes and one of the present attribute is our class (as per the letter grades). Our main motive is to find a model for the class that will be able to predict the undiscovered records (from external similar data sources) accurately which will be similar to the known label of the class, provided all values of other attributes. We usually divide the data set into two subsets, to train the model in a particular manner for a specific task: training set and test set. The model will be built with the help of training set and the test set will do the validation. It is the test set which determines the accuracy and performance of the model. Classification is supervised learning.

**Regression**: Regression is a statistical modeling technique using previous observations to predict new data values. In other words, it is a method of determining relationships between data elements based on the predicted data values for a set of defined variables. This category’s classifier is called the ‘Continuous Value Classifier’. [Linear regression](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/)**,**multivariate regression, and decision trees are key examples of this type. It is the data mining algorithm which determines relationship between more than one variable. As an example can be given as the determination of expenses of potential customers whose income and profession are known. Regression can be simply called as the "predictive power". Assuming a linear or non-linear model of dependency, regression analysis can be used by us to predict the value of given (continuous) features based on the other features in the data. The data item is mapped into a real valued prediction variable. Here are some examples: The revenue of new products are predicted depending upon the complementary products. Based on the amount of food and cigarette consumed by a person and his age the prediction of cancer can be done. "Logistic regression" is such a term which appears in almost every aspect of this field and regression techniques are also found to be useful in this science. Another illustration is estimating a customer's age based on past purchases. We may also forecast costs based on factors like consumer demand; for instance, if demand for vehicles in the US increases, prices on the secondary market would rise. These techniques are especially used in the case of neural network which can be used to create such complex functions which help in imitating the functionalities of the brain.

**Anomaly detection (change and deviation detection):** Anomaly detection technique helps to determine the most significant data change that has taken place in the database. It is used to find outliers (values that deviate from the norm). For instance, it can identify unexpected sales at a store location during a specific week in e-commerce information. It can be used among other things, to find credit or debit fraud and spot network attacks or disruptions. This involves identifying deviations in a dataset. These might either represent data errors or informative outliers, depending on the context. This is calculated and identified on the previously determined data.

**Prediction:** The prediction data mining technique is typically used for predicting the occurrence of an event, such as the failure of machinery or a fault in an industrial component, a fraudulent event, or company profits crossing a certain threshold. Prediction techniques can help analyze trends, establish correlations, and do pattern matching when combined with other mining methods. Using such a mining technique, data miners can analyze past instances to forecast future events. Prediction can be thought of as the creation and use of a model to determine the class of an unlabeled item or the value or ranges of a particular attribute that an object is likely to possess.

**Data Mining Software Tools and Applications**

**1. Rapid Miner**

It is developed by Rapid Miner Company; hence the name of this tool is a rapid miner. It is written [using java programming language](https://www.educba.com/java-programming-language-features/). The fast miner can be used for predictive analysis, business application, education and research, commercial applications, etc. It increases the speed of delivery as it follows the template framework. It not only increases the delivery speed but also reduces errors while transforming data. It is a free and open source data science platform. There are three types of rapid miner – Rapid Miner Studio, Rapid Miner Server, and Rapid Miner Radoop.

1. Rapid Miner Studio: Workflow design, prototyping, validation, etc., are done in this module.
2. Rapid Miner Server: This module is used for operating predictive data models.
3. Rapid Miner Radoop: For simplification of predictive analysis, this module executes a process in Hadoop.

**2. Orange**

It is open-source software written in python language. Orange is the best software for analyzing data and machine learning. These components are called widgets. These widgets are used for reading data, analyzing components, allowing users to select the features, and showing the data. With orange, data formatting and moving them with the help of widgets become fast and easy.

**3. Weka**

The University of Waikato developed Weka. It is an open-source software used for predictive modelling and analysis of data. Weka [has a GUI](https://www.educba.com/course/graphic-user-interface/) that provides easy and interactive access to users. It [supports SQL](https://www.educba.com/what-is-sql/) and allows a user to connect to the database, and perform operations by firing query. It stores data in a flat-file format.

**4. Oracle Data Mining**

The Oracle Data Mining tool is a part of ‘Oracle Advanced Analytics’ that creates predictive models and comprises multiple algorithms essential for tasks such as classification, regression, prediction, and so on.

Oracle Data Mining allows businesses to identify and target prospective audiences, forecast potential customers, classify customer profiles, and even detect frauds as and when they occur. Moreover, the programmer community can integrate the analytics model into BI applications using a Java API to see complex trends and patterns. Oracle provides a 30-day free trial to potential buyers

**5. KNIME**

It is an open-source developed by KNIME.com AG used for data analytics. It is built by combining data mining and machine learning components. It has been used for pharmaceutical research, business intelligence, and financial analysis.

**6. Sisense**

It is not open-source software; it is licensed software, and we have to purchase the license to use this. Small and large organizations use Sisense to handle the data. As it also supports widgets like orange, it is easy to move data and create reports by dragging and dropping. Even non-technical persons can work with Sisense as its GUI based. With the help of widgets, Sisense generated words are in the form of bar chart, pie chart, line chart, etc.

**7. Apache Mahout**

The Apache foundation developed it. Apache Mahout aims at creating algorithms for machine learning and focuses on regression, clustering classification of data. As it is written in a well-known language like java and contains java libraries that support mathematics operation, it is [used for statistical analysis](https://www.educba.com/top-10-free-statistical-analysis-software/). The tool is an open-source platform designed for researchers and professionals who intend to implement their own algorithms.

**8. SSDT**

SSDT is short for SQL Server Data Tools. It is used to expand the database development phases in a visual studio. It is widely used for data analysis and provides solutions to solve business intelligence problems. SSDT provides a table designer to perform table operations like create a table, adding table data, deleting table data, modifying table content. It allows a user to connect to the database as it supports SQL.

**9. Rattle**

The Rattle is an open-source developed using the R language. It provides a GUI interface. The inbuilt log close tab enables Rattle to generate duplicate for every activity.

**10. DataMelt**

It is also known as DMelt. It is used to analyze and visualize data. It is designed for students, engineers, and scientists. It is platform-independent, which means it can run on any operating system [which contains JVM](https://www.educba.com/what-is-jvm/) (Java Virtual Machine). It is used to create 2D or 3D plots, random numbers, mathematical operations, algebra equations.

**11. IBM Cognos**

It is suited for Business Insider intelligence. It is used for analyzing data, data reporting.

Components of IBM Cognos

1. **Report Studio**: It is used to generate reports.
2. **Query Studio:** Contains query operation to get desired results.
3. **Analysis Studio:** It is used to handle a large amount of data and analyzing the relation between data
4. **Event Studio:** It is used to give the event notifications.
5. **Cognos Connection:** It is a web portal to summarize the large volumes of data and give the reports.

**12. SAS**

It is developed for managing a large amount of data. It allows a user to modify the data, store data from different locations into one space. As it provides a GUI interface, a non-technical person can also use this quickly and handles their data efficiently. Offers free software trial and customized pricing packages for advanced features.

**13. Teradata**

It contains data warehouse tools as well as [data mining software](https://www.educba.com/data-mining-software/). It is widely used for business analytics. Teradata is used to give information about data like the available products, number of products sold, inventory, etc. Teradata offers a flexible pricing model, which refrains from charging any upfront cost.

**14. Dundas**

It is a dashboard, analytics, and reporting tool. With Dundas, unlimited data transformation is possible. It provides features to create attractive data like charts, tables styles, graph, text formatting, etc.

**15. H2O**

The H2O data mining tool brings AI technology into data science and analysis, making it accessible to every user. The tool is suitable for running several ML algorithms with features that support auto ML functions for the build and faster deployment of ML models. H2O is a free-to-use and open-source tool.

H2O offers integration features through APIs available in standard programming languages and is suitable for managing complex datasets. The tool provides fully-managed options and the facility to deploy it in a hybrid setting.