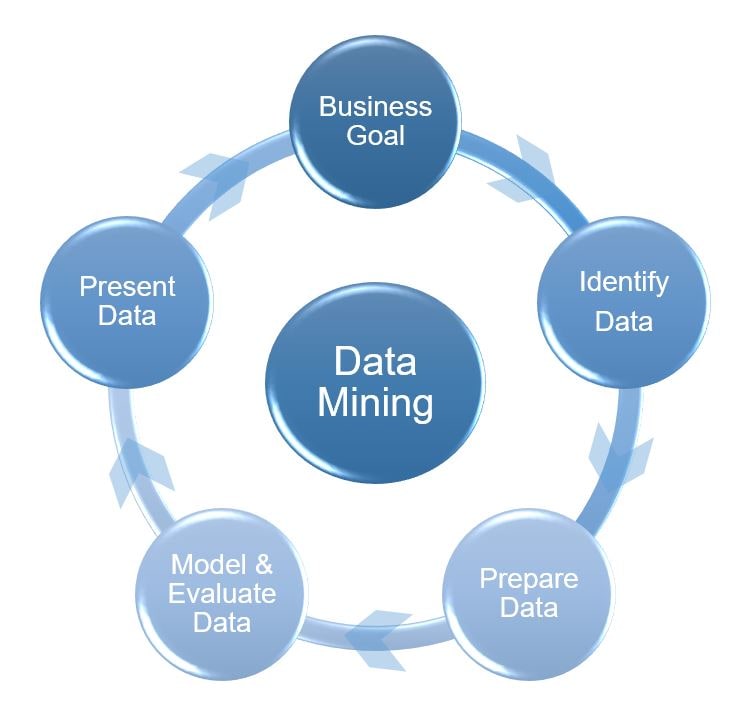
**What is Data Mining?**

Data mining consists of two primary functions, description, for interpretation of a large database and prediction, which corresponds to finding insights such as patterns or relationships from known values. Before deciding on data mining techniques or tools, it is important to understand the business objectives or the value creation using data analysis. The blend of business understanding with technical capabilities is pivotal in making big data projects successful and valuable to its stakeholders.



**Defining your business need for big data analysis**

Businesses should emphasize on clear objectives of their big data strategy, such as, identifying high value customers to offer specific products or services, process improvements for optimizing cost, and so on.

**Data identification and acquisition**

 Is the foremost step for successful implementation. Understanding the business challenges that you are trying to solve helps in determining the source and types of data to utilize. Data can be in any form - it can be a subset of variables, or can be data samples from a larger database. The key data should directly correlate to the business objective. We will further discuss this in detail while selecting data mining techniques.

**Data cleansing or cleaning**

Is done on the target data set to improve its effectiveness for fulfilling data mining objectives. This process is primarily to identify inaccurate, incorrect or incomplete data and then replacing, modifying or removing it. This process ensures that your data is complete and error free, thus making it more relevant and effective.

**Data exploration**

Is at the core of data mining activity. The main objective of this step is to identify the correct data mining techniques or methods and selecting the best suited algorithms for those techniques. Some of the most known data mining techniques include association, classification, regression, segmentation, link analysis, etc. Selecting data mining techniques among the pool is one of the most difficult decisions. However, the selection parameter should consider the objective of business and the available data sets. Mostly, two or more combinations of data mining techniques are used, but it depends on the scale of the project. The last leg of this activity is to perform data mining, i.e., to search for patterns hidden in the data.

**Data presentation**

Is the final activity which includes interpretation and evaluation of patterns and presenting them to users in a logical and understandable way. The main objective here is to present useful patterns and discard any redundant or irrelevant ones. For any inconsistency, revision of data exploration step is required. The success is attributed to its usefulness in providing insights to decision makers and its relevance to the business objectives.

## ****Identifying the data mining operations****

**Classification -**This data mining function is used to classify data into different buckets/classes based on constraints. The technique is used in large data sets to predict category of class labels based on training data sets. Some of the business cases which utilize these techniques are diagnosis of patient’s medical condition to select medical treatment, classifying individuals in different credit groups based on their financial data and segregating individual loan applicants in different credit risk parameters. Most prominently used classification algorithms are Naive Bayes, SVM (Support Vector Machines), k-nearest neighbor classifier and ANN (Artificial Neural Network). Determining the classification algorithm is crucial and confusing at times; it requires experts to evaluate the best for a given project. For example, Naive Bayes algorithm, though simple to implement, required a large data set for training. ANN can be used with fewer parameters but requires high processing time.

**Regression** – This operation is used to predict the real value variable. Traditional data models are developed using statistical methods like linear and logistics regression. When compared with classification, both are used for prediction. However, output is categorized in classification, and it is numeric output in regression. Some prominent examples for regression operation are determining the crime rate of a city based on different parameters, property valuation based on factors like location, floor area, etc., insured scoring systems (like in auto insurance) to predict likelihood of an insured meeting with an accident, etc. A few popular regression algorithms are Generalized Linear Models (GLM) for Linear and Support Vector Machines (SVM) for both linear and nonlinear regression.

**Statistics**

Statistics is the science of data that involves:

* Collecting
* Classifying
* Summarizing
* Organizing and Interpretation

Statistics deals with every aspect of data including the planning of data collection in term of design of surveys and experiments.

Scope of Statistics

* Used in Business
* Used in Economics
* Used in Banking
* Used in Accounting
* Used in Astronomy
* Used in Research Work