Business Intelligence Concepts and its Applications

Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization‘s data and communicate the information to relevant users.

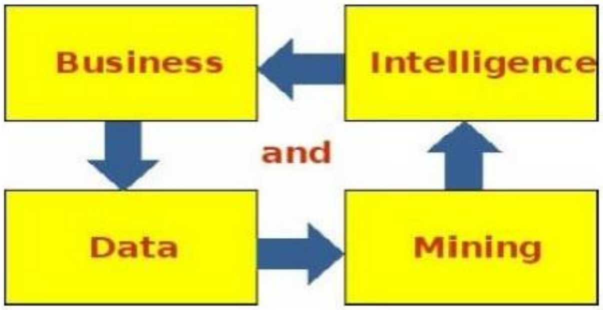


Fig 1 : BIDM cycle

The essence of life and business is to grow. Information is the lifeblood of your business. Business is more effective than one based on emotion alone. Actions that leverage new insights, based on accurate data, information, knowledge, experimentation and testing, are more likely to succeed and lead to sustainable growth. Your own data can be your most effective teacher. Therefore, organizations must collect, sift, analyze and evaluate data, generate insights, and then incorporate these insights into their operations. There is a new recognition of the importance and urgency of data as it is seen as the new natural resource. It can be mined for value, insight and competitive advantage.

In a hyper-connected world where everything is potentially connected to everything and potentially infinitely correlated, data represents natural impulses in the form of specific events and attributes. Veteran businessmen want to use this data store

BI for better decisions

The future is inherently uncertain. Risk is the result of a probabilistic world where certainty does not exist and complexity abounds. People use crystal balls, astrology, palmistry, groundhogs, and even math and numbers to de-risk their decisions. The goal is to make effective decisions while reducing risk. Organizations calculate risk and make decisions based on a variety of facts and insights. Reliable knowledge of the future helps managers make the right decisions with less risk

The spread of the Internet has dramatically improved the speed of action. In a competitive world, speed of decision-making and consistent action are key advantages. The Internet and mobile technology have made it possible to make decisions anytime, anywhere. Ignoring the rapid pace of change can put your company's future in jeopardy. According to research,

Companies and their products on social media shouldn't be left alone for long. In 2013, banks had to pay hefty fines for complaints filed on the US Consumer Financial Protection Bureau (CFPB) website. On the other hand, positive sentiment expressed on social media should be leveraged as potential sales and promotional opportunities.

Decision types

There are two main types of decisions: strategic decisions and operational decisions. BI can help improve both. Strategic decisions are those that influence the direction of the company. The decision to reach new customer groups becomes a strategic decision. Operational decisions are more routine, tactical decisions aimed at increasing efficiency. Updating an old website with new functionality is an operational decision.

In strategic decision making, the goal itself may or may not be clear, and the same is true of the path to the goal. The result of the decision will become clear after some time. Therefore, we are always looking for new opportunities and new ways to achieve our goals. BI helps with what-if analysis for many possible scenarios. BI can also help generate new ideas based on new patterns found through data mining.

By analyzing historical data, you can make operational decisions more efficiently. Classification systems can be created and modeled using data from past instances to develop a good model of the domain. This model will help improve future operational decisions. BI helps improve efficiency by automating operational-level decision-making and model-driven millions of micro-level operational decisions. For example, banks want to use data-driven models to make financial lending decisions in a more scientific way. Decision tree-based models can provide consistently accurate credit decisions. Developing such decision tree models is one of the main uses of data mining techniques.

Effective BI has components that evolve as your business model evolves. New facts (data) are created by the actions of people and organizations. Current business models can be tested against new data, but these models may not perform well. Decision models must then be revised to incorporate new findings. An endless process of generating fresh new insights in real time can help you make better decisions, which in turn can represent a significant competitive advantage.

BI Tools

BI includes a variety of software tools and techniques to provide managers with the information and insights they need to run their business. It can provide current situational information with the ability to drill down into details, as well as insight into emerging patterns that can predict the future. BI tools include data warehousing, online analytical processing, social media analytics, reports, dashboards, queries, and data mining.

BI tools range from very simple tools that can be considered end-user tools to very sophisticated tools that offer a very wide and complex range of functionality. An executive can be her own BI expert or rely on a BI specialist to set up her BI mechanism. That's why large organizations invest in expensive and sophisticated her BI solutions that deliver great information in real time.

A spreadsheet program like Microsoft Excel can serve as a simple and effective BI tool in its own right. You can download the data, save it in a spreadsheet, analyze it to generate insights, and display it in charts and tables. This system offers limited automation with macros and other features. Analytical features include basic statistical and financial functions. Pivot tables are useful for advanced what-if analysis. Add-on modules can be installed to enable moderately advanced statistical analysis.

Dashboard systems such as IBM Cognos and Tableau can provide a sophisticated set of tools for collecting, analyzing, and presenting data. On the user side, modular dashboards can be easily designed and redesigned with a graphical user interface. The backend data analysis functionality includes many statistical functions. Dashboards are linked to a back-end data warehouse so that dashboard tables, charts, and other elements are updated in real time.

Data mining systems such as IBM SPSS Modeler are industrial-grade systems that provide the ability to apply a wide range of analytical models to large amounts of data. Open source systems like Weka are popular platforms designed to sift through large amounts of data and discover patterns.

BI Skills

As data grows and exceeds our capacity to make sense of it, the tools need to evolve, and so should the imagination of the BI specialist. –Data Scientist has been called as the hottest job of this decade.

The more seasoned and experienced BI specialists are able to think outside the box, open the door, and see a broader perspective with more dimensions and variables to find the patterns and insights that matter. Must be open minded. This issue needs to be looked at from a wider perspective to include more angles that may not be immediately apparent. We need to propose imaginative solutions to problems to produce interesting and useful results.

A good data mining project starts with an interesting problem to solve. Choosing the right data mining problem is an important skill. The problem must be worth the time and expense to solve it. Collecting, organizing, cleaning, and preparing data for mining and other analysis requires a lot of time and energy. Data miners should continue to explore patterns in the data. Your skill level must be deep enough to engage with the data and derive useful new insights from it.

BI Applications

BI tools are needed in almost every industry and function. Although the type of information and speed of response varies from company to company, every manager today needs access to her BI tool for the most up-to-date indicators on company performance. Organizations must incorporate new insights into their operational processes and evolve operations with more efficient practices. Some of the application areas of BI and data mining are listed below.

Customer Relationship Management

A company exists to serve its customers. Satisfied customers become regular customers. Businesses need to understand their customers' needs and emotions, sell more products to existing customers, and expand the customer base they serve. BI applications can affect many aspects of marketing.

1. Maximize revenue from marketing campaigns: Understanding customer pain points from data-driven analytics allows marketing to fine-tune her messaging to increase customer empathy.

2. Improved Customer Retention (Churn Analysis):

Acquiring new customers is more difficult and costly than retaining existing ones. Evaluating each customer based on their likelihood of churn can help companies design effective interventions, such as discounts and complimentary services, to retain profitable customers in a cost-effective manner. increase.

3. Maximize Customer Benefits (Cross, Upselling): Every customer contact should be viewed as an opportunity

to assess the customer's current needs. Offering new products and solutions to customers based on these envisioned needs helps increase revenue per customer. Customer complaints can also be seen as an opportunity to motivate customers. By knowing the history and values of their customers, companies can decide to sell premium services to them.

4. Identify and delight your valued customers. By segmenting your customers, you can identify your best customers. They are more likely to be in touch, satisfied with more attention and better service.You can manage your loyalty program more effectively.

5. Management of brand image. Businesses can create listening posts to hear social media gossip about their company. You can then perform sentiment analysis on the text to understand the nature of the comments and respond appropriately to your prospects and customers.

Healthcare and Wellness

Healthcare is one of the largest sectors in developed countries. Evidence-based medicine is the latest trend in data-driven healthcare. BI applications help apply the most effective diagnoses and prescriptions for various diseases. It also helps address public health issues and reduce waste and fraud.

1. Diagnosing a patient's illness: Diagnosing the cause of an illness is an important first step in any medical practice. Accurately diagnosing cancer and diabetes is a matter of life and death for patients. In addition to the patient's own current situation, many other factors may be considered, including the patient's medical history, medication history, family history, and other environmental factors. This makes diagnostics as much an art as it is a science. Systems like IBM Watson take all previous medical research and present probabilistic diagnoses in the form of decision trees, along with full explanations of recommendations. These systems take most of the guesswork out of diagnosing illnesses for doctors.

2. Treatment Effectiveness: Prescribing drugs and treatments is also a difficult choice with so many options. For example, there are more than 100 drugs for hypertension (hypertension) alone. There are also interactions in terms of which drugs work better with other drugs and which do not. Decision trees help doctors learn and prescribe more effective treatments. Patients can therefore return to health faster, with less risk and cost of complications.

3. Wellness Management: This includes tracking patient records, analyzing clients' health trends, and providing proactive advice to take necessary preventative measures.

4. Fight Fraud and Abuse: Unfortunately, we know that some doctors perform unnecessary tests and overcharge governments and health insurance companies. The exception reporting system can identify such providers and act against them.

5. Public Health Management: Public health management is one of the key tasks of government. By using effective forecasting tools and technologies, governments can more accurately predict disease outbreaks in specific regions in real time. Prepare to fight disease. Google is known to predict the prevalence of certain diseases by tracking search terms (flu, vaccines, etc.) used in different parts of the world.

Data Warehousing

A data warehouse (DW) is an organized collection of subject-oriented, consolidated databases designed to support decision support functions. DW is organized at the right level of granularity to provide clean, enterprise-wide data in a standardized format for reporting, querying, and analysis. DW is physically and functionally separated from operational and transactional databases. Creating a DW for analytics and queries takes a lot of time and effort. To be useful, it should be constantly updated. DW offers many business and technical advantages.

DW supports annual reports and data mining activities. Facilitate distributed access to the latest business knowledge across departments and functions to improve business efficiency and customer service. DW can provide a competitive advantage by driving decision-making and supporting business process innovation.

DW gives you a unified view of your enterprise data, all cleaned up and organized. So you get a unified view of your entire organization. In this way DW provides more relevant and timely information. This simplifies data access and enables extensive end-user analytics. Improve overall IT performance by offloading operational databases used by Enterprise Resource Planning (ERP) and other systems.

Design Considerations for DW

DW's goal is to provide business knowledge that supports decision making. These decisions must be aligned in order for DW to achieve its goals. It should be comprehensive, easily accessible, and up-to-date. Here are the requirements for a good DW:

1. Theme-based: To be effective, DW must be designed around a theme. H. Helps solve certain categories of problems.

2. Integration: DW should contain data from many features that can shed light on a particular topic. Organizations can therefore benefit from a comprehensive view of the topic.

3. Time Series (time series): The data in DW should increase daily or another selected interval. This allows for current comparisons over time.

4. Non-Volatile: DW must be persistent. In other words, it should not be voluntarily created from a production database. This means DW is continuously available across the enterprise and for analysis over time.

5. Aggregation: DW contains data aggregated at the appropriate level for querying and analysis. The process of summarizing data helps create consistent granularity for effective comparison. It also helps reduce the number of variables or dimensions in your data to make it more meaningful
decision maker.

5. Denormalization: DWs often use star schemas. A star schema is a central rectangular table surrounded by several lookup tables. A single table view greatly speeds up queries.

6. Metadata: Many of the variables in the database are calculated from other variables in the operational database. For example, total daily sales can be a calculated field. How each variable is calculated should be effectively documented. Each element of the DW should be well defined and well defined.

8. Near Real Time and/or Right Time (Active): DW should be used in many industries for mass production. B. Airlines are updated in near real time. However, the cost of implementing and updating DW in real time can be daunting. Another drawback of real-time DW is that reports can be inconsistent at intervals of just a few minutes.

DW Architecture

DW has four key elements. The first element is the data source that provides the raw data. The second element is the process of transforming this data to meet decision-making requirements. The third factor is how to load that data regularly and accurately into the EDW or data mart. The fourth element is the data access and analytics part, where devices and applications use data from his DW to provide insights and other benefits to users.

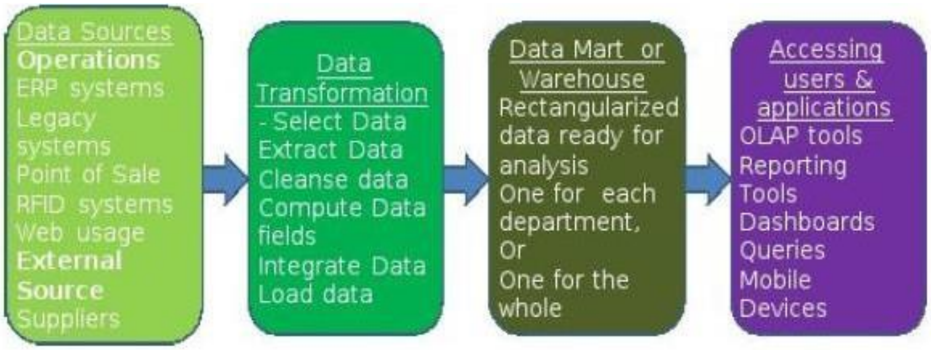


Fig 2: Data Warehousing Architecture

Data Sources

Data warehouses are built from structured data sources. Unstructured data such as text data should be structured before being inserted into the DW.

1. Operational data: This includes data from all business applications, including ERP systems that form the backbone of an organization's IT systems. The data extracted depends on the data warehouse object. For example, for a sales/marketing data mart, only data about customers, orders, customer service, etc. is extracted.

2. Specialized applications: This includes applications such as point-of-sale (POS) terminals and e-commerce applications that also provide customer-facing data. Supplier data can come from your supply chain management system. Planning and budget data should also be added where appropriate for comparison with targets.

3. Externally syndicated data: This includes publicly available data such as weather and economic data. You can also add it to DW if you need to provide decision makers with the right contextual information.

Data Loading Processes

At the heart of any useful DW is the process of entering high-quality data into the DW. This is called an extract-transform-load (ETL) cycle.

1. Data must be routinely extracted from operational (transactional) database sources and other applications.

2. The extracted data should be aligned by major fields and merged into a single dataset. I need to remove irregularities and missing values. They should roll up to the same level of granularity. Desired areas such as B. I need to calculate the total sales for each day. Next, the entire data should be in the same format as the central table in DW.

3. These transformed data should be uploaded to DW. This ETL process should be run on a regular basis. Daily transactional data can be extracted from the ERP, transformed, and uploaded to the database the same night. So the DW is updated every morning, but if you need his DW for near real-time information access, you'll need to run the ETL process more frequently. ETL work is typically performed using automated programming scripts. This script is written, tested and deployed to update the DW regularly.

Data Warehouse Design

A star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There is a lookup table that provides detailed values for the codes used in the central table. For example, the middle table can use numbers to represent salespeople. A lookup table helps provide the name of this salesperson code. Below is an example of a data mart star schema for monitoring sales performance. Other schemes include the Snowflake architecture.

The difference between star and snowflake is that in the latter a lookup table can have other lookup tables of its own. DW development has many technology options. This includes choosing the right database management system and the right data management tools. There are several large and reliable providers of DW systems. A DW can also choose a provider for a production DBMS. Alternatively, you can use a premium DW provider. There are also various tools for data migration, data upload, data retrieval, and data analysis.

DW Access

Data from DW can be accessed from many devices for many purposes by many users.

1. The primary use of DW is to generate periodic management and monitoring reports. For example, a sales performance report shows sales in many dimensions and compares them to plans. A dashboard system uses data from the warehouse and presents analytics to the user. Data from DW can be used to create customized performance dashboards for executives. Dashboards can include drill-down capabilities to analyze performance data for root cause analysis.

2. Data from DW can be used for ad-hoc queries and other applications that use internal data.

3. Data from DW is used to provide data for mining purposes. A portion of the data is extracted and combined with other relevant data for data mining.

DW Best Practices

Data warehousing projects reflect a large investment in information technology (IT). All best practices should be followed when implementing an IT project.

1. DW projects should be based on corporate strategy. To set goals, you should consult top management. Affordability (ROI) must be established. Projects should be managed by both IT and business professionals. DW designs should be thoroughly tested before development begins. A redesign after development work has begun is often much more expensive.

2. It's important to manage user expectations. A data warehouse should be built in stages. Users should be trained in using the system so that they can absorb the many functions of the system.

3. Quality and adaptability must be built in from the beginning. Only relevant, clean, and high-quality

data should be loaded. The system should be able to adapt to new access tools. As your business needs change, you may need to create new data marts to meet your new needs.

Data Mining

Data mining is the art and science of discovering knowledge, insights and patterns in data. This is the process of extracting useful patterns from an organized collection of data. Patterns should be valid, novel, potentially useful, and understandable. The implicit assumption is that data about the past can reveal patterns of activity that can be predicted in the future.

Data mining is an interdisciplinary field that incorporates techniques from many different disciplines. Uses knowledge of data quality and data organization from the field of databases. It is based on modeling and analytical techniques from the fields of statistics and computer science (artificial intelligence). It also draws decision-making knowledge from the field of business administration.

The field of data mining arose in the context of pattern recognition in defense, such as identifying friend versus foe on the battlefield. Like many other defense-inspired technologies, it has evolved to gain a competitive advantage in business.

For example, "90% of customers who buy cheese and milk also buy bread" is a useful template for grocery stores and can store products accordingly. resemble, “A person whose blood pressure is 160 years old or older and who is 65 years old or older has an increased risk of dying from a heart attack" is of great diagnostic value to physicians, who may refer such patients to urgent care. You can focus on treating with care and highly sensitive treatments.

Historical data can be predictive in many complex situations, especially when patterns are not immediately apparent without modeling techniques. This is a dramatic case of data-driven decision-making systems beating out the best human experts. Using previous data, a decision tree model was developed to predict the vote of Judge Sandra Day O'Connor, who had her 5-4 split vote on the U.S. Supreme Court. All previous decisions were coded based on some variables. Data mining yielded a simple four-level decision tree that correctly predicted votes 71% of the time. In contrast, legal analysts were able to predict correctly at best 59% of the time. (Source: Martin et al. 2004)

Gathering and selecting data

The amount of data in the world doubles every 18 months. There is an ever-growing avalanche of data, with ever-increasing speed, volume, and diversity. You must use it immediately or lose it. Intelligent data mining requires game location selection. You should make wise decisions about what to collect and what to ignore based on the purpose of your data mining exercise. It's like choosing where to fish. This is because not all data streams are equally rich in potential insights.

Learning from data requires effective collection, organization, and organization of high-quality data, and then mining it efficiently. Integrating and integrating data elements from many sources requires skills and technology. Most organizations develop an Enterprise Data Model (EDM) to organize their data. EDM is a unified high-level model of all data stored in an organization's database. EDMs typically include data generated by all internal systems. EDM provides a basic menu of data for creating data warehouses for specific decision-making purposes. DW helps organize all this data in a simple and easy-to-use way so that it can be selectively mined. EDM also helps you visualize what relevant external data you need to collect to provide context and build good predictive relationships with your internal data. In the United States, various federal and local governments and their regulators make a large variety of data available at data.gov.

Collecting and organizing data takes time and effort, especially for unstructured or semi-structured data. Unstructured data comes in a variety of formats, including databases, blogs, images, videos, audio, and chats. It has streams of unstructured social media data from blogs, chats, and tweets. There are streams of machine-generated data from connected machines, RFID tags, the Internet of Things, and more. Finally, I need the data to be rectangular. H. Columns and rows are transformed into a distinct rectangular data shape before being sent to data mining.

Business domain knowledge helps you select the right data streams to gain new insights. You should collect only data that is appropriate to the nature of the problem you are solving. Data elements should be relevant and appropriately address the problem to be solved. They may have a direct impact on the problem or be a suitable surrogate for measured effects. Selected data can also be collected from a data warehouse. Every industry and function have its own requirements and limitations. The healthcare industry provides different types of data with different data names. HR functionality provides different types of data. This data has various aspects of quality and data protection.

Data cleansing and preparation

Data quality is critical to the success and value of any data mining project. Otherwise, the situation is garbage in and garbage out (GIGO) type. The quality of input data depends on the source and type of data. Data from internal operations may be of higher quality because they are accurate and consistent. Data from social media and other public sources are not under the control of the company and may be unreliable.

You will almost certainly need to clean and transform your data before using it for data mining. There are many ways data must be cleaned up before it is ready for analysis, such as imputing missing values, controlling for the influence of outliers, transforming fields, and grouping continuous variables. Data cleaning and preparation are labor-intensive or semi-automated activities that take up to 60-70% of the time required for data mining projects.

1. Duplicate data should be removed. You can receive the same data from multiple sources. When merging datasets, the data should be deduplicated.
2. Missing values should be filled or these rows should be removed from the analysis. Missing values can be filled with mean or modal or normal values.
3. Data items should be comparable. They may (a) need to be transformed from one entity to another; For example, the total cost of healthcare and the total number of patients may need to be reduced to cost/patient to allow for comparability of this value. Data items may (b) need to be adjusted for comparison over time. For example, you may need to adjust currency values for inflation. Must be converted to the same base year for comparability. They may need to be converted into a common currency. (c) data should be stored at the same granularity to ensure comparability; For example, sales data may be available daily, but compensation data for salespeople may be available only monthly. To relate these variables, the data must be aligned to the lowest common denominator (months in this case).

4. Continuous values may need to be split into several buckets to aid in several analyses. For example, work experience can be categorized as low, medium, and high.
5. Outlier data items should be removed after careful consideration to avoid biasing the results. For example, large donors can skew the analysis of alumni donors in educational settings.
6. Ensure that the data are representative of the phenomenon being analysed by correcting for data selection biases. For example, data should be adjusted if it contains more members of each gender than is typical for the population of interest.
7. Data may need to be selected for greater information density. Some data may not fluctuate much due to improper recording or other reasons. This data can weaken the impact of other differences in the data and should be removed to improve the information density of the data.

Outputs of Data Mining

Data mining techniques are useful for many kinds of goals. Data mining results reflect the goals pursued. There are many ways to present the results of data mining.

A common form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually track the steps for making model-based decisions. Trees can have certain attributes such as probabilities associated with each branch. A related form is a set of business rules that are if-then statements that indicate causal relationships. Decision trees can be associated with business rules. If the objective function is prediction, decision trees or business rules are the most appropriate ways to represent the output.

The output can be in the form of a regression equation or a mathematical function representing a curve that best represents the data. This equation can contain linear and nonlinear terms. A regression equation is a good way to show the results of a classification exercise. They are also good representations of the prediction formula.

A population "centroid" is a statistical measure used to represent the central tendency of a collection of data points. They can be defined in multidimensional space. For example, focus is "Middle-aged, highly educated, wealthy professionals, married, two children, live in a coastal area" Or the population of "20-year-old, highly-skilled Silicon Valley-based tech entrepreneurs." Or it could be a collection of "her 20+ year old vehicles with low mileage per gallon that didn't pass environmental testing." These are typical representations of the results of the cluster analysis exercise. A business rule is a rational representation of the results of a market basket analysis. These rules are if-then statementing with some probability parameter associated with each rule. For example, buying milk and bread also buys butter (80% chance).

The output can be in the form of a regression equation or a mathematical function representing a curve that best represents the data. This equation can contain linear and nonlinear terms. A regression equation is a good way to show the results of a classification exercise. They are also good representations of the prediction formula.

Evaluating Data Mining Results

There are two main types of data mining processes: supervised learning and unsupervised learning.

Supervised learning allows you to build a decision-making model based on past data and use that model to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many classification techniques, but decision trees are the most common. Each of these techniques can be implemented with many algorithms. A common metric for all classification methods is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Data Mining Techniques

You can mine your data to make more efficient decisions in the future. Alternatively, it can be used to examine data and find interesting association patterns. The correct technique depends on the type of problem you are solving. The most important class of problems solved by data mining is the classification problem. Classification techniques are called supervised learning because there are ways to monitor whether the model is providing correct or incorrect answers. These are problems that involve evaluating data from previous decisions to extract some rules and patterns that improve the accuracy of future decision-making processes. It cleans up historical decision data and searches for decision rules or equations. These are codified to produce more accurate decisions.

Decision trees are the most popular data mining technique, for many reasons.

1. Decision trees are easy to understand and use for both analysts and executives. Also, the prediction accuracy is high.
2. Decision Tree automatically selects the most relevant variables from all available variables for decision making.
3. Decision trees are tolerant of data quality issues and do not require much data preparation by the user.
4. Decision trees can handle non-linear relationships well. There are many algorithms for implementing decision trees. The most common are C5, CART, and CHAID.

Regression is one of the most common statistical data mining techniques. The goal of regression is to derive a smooth, well-defined curve and get the most out of your data. For example, regression analysis techniques can be used to model and predict energy consumption as a function of daily temperature. A simple plot of the data may show a nonlinear curve. After applying the nonlinear regression equation, the data fit very well. Once such a regression model is developed, this formula can be used to predict future energy consumption. The accuracy of a regression model is completely dependent on the data set used and not the algorithms or tools used. Artificial Neural Networks (ANNs) are advanced data mining techniques that emerged from the artificial intelligence trend in computer science. This mimics the behavior of human neural structures. A neuron receives a stimulus, processes it, transmits its result to other neurons in turn, and finally he one neuron decides. A decision-making task can be handled by just one neuron, and the results can be communicated immediately. Alternatively, many layers of neurons may be involved in decision tasks, depending on the complexity of the domain. Neural networks can be trained by repeatedly making decisions using a large number of data points. It continues to learn by adjusting its internal calculations and communication parameters based on feedback on past decisions. Intermediate values passed within layers of neurons may not be intuitive to the observer. Neural networks are therefore considered black box systems.

Cluster analysis is an exploratory learning technique that helps identify large numbers of similar groups in data. This is a technique used to automatically identify natural groups of things. Data instances that are similar (or close) to each other are grouped into one cluster, and data instances that are significantly different (or far) from each other are grouped into another cluster. There is no limit to the number of clusters that can be generated by your data. The K-Means technique is a popular technique that provides guidance to the user in choosing the correct number (K) of clusters from the data. Clustering is also called segmentation technique. It helps you share and conquer massive amounts of data. This technique shows clusters of things from historical data. The output is the centroid of each cluster and the mapping of data points to clusters. The centroid definition is used to assign new data instances to cluster homes. Clustering is also part of artificial intelligence technology.

Association rules are a data mining technique commonly used in businesses, especially when it comes to sales. Also known as market basket analysis, it helps answer questions about cross-selling opportunities. It's the heart of the personalization engine used by e-commerce sites like Amazon.com and streaming movie sites like Netflix.com. This technique helps find interesting relationships (affinities) between variables (items or events). These are expressed as rules of the form X ® Y. where X and Y are sets of data items. It is a form of unsupervised learning and has no dependent variable. And there is no right or wrong answer. There are only strong affinities and weak affinities. Each rule is therefore associated with a trust level. Part of the machine learning family, this technique attained legendary status when an interesting relationship was discovered in the sale of diapers and beer.

Tools and Platforms for Data Mining

Data mining tools have been around for decades. However, they are gaining importance recently as the value of data increases and the field of big data analytics gains prominence. There are various data mining platforms on the market today.

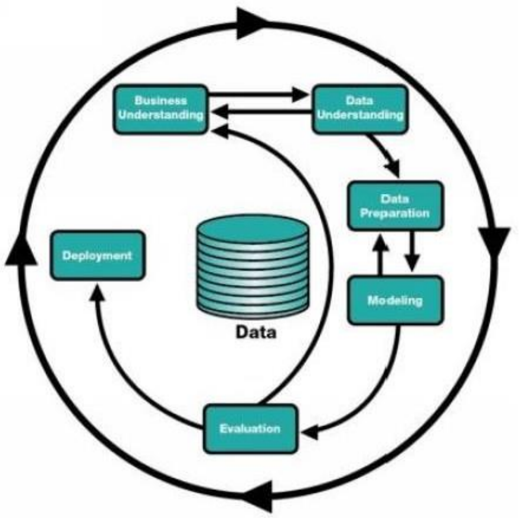
1. Simple or Sophisticated: There are simple end-user data mining tools like MS Excel and more sophisticated tools like IBM SPSS Modeler.
2. Standalone or Embedded: There are standalone tools and tools that are embedded into existing transaction processing, data warehousing, or ERP systems. 3. Open source and commercial: There are open source and freely available tools like Weka, as well as commercial products.
4. User Interface: There are text-based tools that require programming knowledge, and GUI-based drag-and-drop tools.
5. Data Formats: There are tools that only work with their own data formats, and there are tools that accept data in a variety of popular formats directly from data management tools.

Data Mining Best Practices

Effective and successful use of data mining activity requires both business and technology skills.

Fig 3: CRISP-DM Data Mining cycle

The business aspect will help you understand your domain and key questions. It also helps you imagine possible relationships in your data and create hypotheses to test them. The IT side helps in getting data from many sources, cleaning the data, assembling the data to meet the needs of the business problem, and running data mining techniques on the platform. The key element is to approach the problem iteratively. We encourage sharing and tackling problems with smaller amounts of data, and getting closer to the heart of the solution in an iterative series of steps. There are some best practices that we have learned from using data mining techniques over time. The data mining industry has proposed a cross-industry standard process for data mining (CRISP-DM). It has six essential steps:



1. *Business Understanding*: The first and most important step in data mining is asking the right business questions. A question is good if answering it leads to significant financial and other benefits for the organization. In other words, choosing a data mining project, like any other project, should pay off if the project succeeds. Data mining projects require strong administrative support. This means that the project is well aligned with business strategy. A related key step is to be creative and open in proposing imaginative hypotheses for solutions. It is important to think outside the box regarding both the proposed model and the available and required datasets.

2. *Data Understanding*: A related and important step is understanding the data available for mining. You have to be resourceful when sifting through a lot of data from many sources to work on a hypothesis to solve your problem. Without relevant data, hypotheses cannot be tested.

3. *Data Preparation*: Data must be relevant, clean and of high quality. It's important to assemble a team that has both technical and business skills and understands the domain and data. In a data mining project, data cleaning can take 60-70% of the time. It may be desirable to continue experimenting and adding new data items from external data sources to help improve prediction accuracy.

4. *Modeling*: This is the actual task of running a number of algorithms using the available data to see

if the hypothesis is supported. It takes patience to continuously work with data until good insights come out of it. A variety of modeling tools and algorithms should be used. The tool can be tried with different options such as: B. Running various decision tree algorithms.

5. *Model Evaluation*: Don't just accept what the data show. We recommend triangulating your analysis by applying multiple data mining techniques and running many what-if scenarios to build confidence in your solution. We need to evaluate and improve the predictive accuracy of our model with more test data. Once the accuracy reaches a sufficient level, the model should be used.

6. *Dissemination and rollout*: It is important that the data mining solution is presented to key stakeholders and used within the organization. Otherwise, projects are a waste of time and a hindrance to establishing and supporting a data-driven decision-making culture within your organization. The model should ultimately be incorporated into the organization's business processes.