# 4 Data Mining

Learning Objectives

* Outline the concept of Data Mining
* Understand the process of gathering, selecting, cleansing, and otherwise preparing the data for mining
* Identify the outputs of data mining
* Learn to evaluate the effectiveness of various data mining models
* Understand the CRISP-DM data mining process
* Summarize the range of data mining techniques
* Appreciate the wide range of toolsets available for data mining and their respective strengths

### INTRODUCTION

Data Mining is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data. Patterns must be valid, novel, potentially useful, and understand- able. The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.

Data mining is a multidisciplinary field that borrows techniques from a variety of fields. It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

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

For example, “customers who buy cheese and milk also buy bread 90 percent of the time,” would be a useful pattern for a grocery store, which can then stock the products appropriately. Similarly, “people with blood pressure greater than 160 and an age greater than 65 were at a high risk of dying from a heart stroke,” is of great diagnostic value for doctors, who can then focus on treating such patients with urgent care and great sensitivity.

Past data can be of predictive value in many complex situations, especially where the pattern may not be so easily visible without the modeling technique. Here is a dramatic case of a data-driven decision-making system that beats the best of human experts. Using past data, a decision tree model was developed to predict votes for Justice Sandra Day O’Connor, who had a swing vote in a 5–4 divided US Supreme Court. All her previous decisions were coded on few variables. What emerged from data mining was a simple four-step decision tree that was able to accurately predict her votes 71 percent of the time. In contrast, the legal analysts could at best predict correctly 59 percent of the time. (*Source:* Martin et al., 2004)

#### Caselet: Target Corp – Data Mining in Retail

*Target is a large retail chain that crunches data to develop insights that help target marketing and advertising campaigns. Target analysts managed to develop a pregnancy prediction score based on a customer’s purchasing history of 25 products. In a widely publicized story, they figured out that a teenage girl was pregnant before her father did. The targeting can be quite successful and dramatic as this example published in the New York Times illustrates.*

*About a year after Target created their pregnancy-prediction model, a man walked into a Target store and demanded to see the manager. He was clutching coupons that had been sent to his daughter and he was angry, according to an employee who participated in the conversation. “My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?”*

*The manager didn’t have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man’s daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.*

*On the phone, though, the father was somewhat subdued. “I had a talk with my daughter,” he said. “It turns out there’s been some activities in my house I haven’t been completely aware of. I owe you an apology.” (Source: New York Times).*

1. *Do Target and other retailers have full rights to use their acquired data as it sees fit, and to contact desired consumers with all legally admissible means and messages? What are the issues involved here?*
2. *Facebook and Google provide many services for free. In return they mine our email and blogs and send us targeted ads. Is that a fair deal?*

### GATHERING AND SELECTING DATA

The total amount of data in the world is doubling every 18 months. There is an ever-growing avalanche of data coming with higher velocity, volume, and variety. One must quickly use it or lose it. Smart data mining requires choosing where to play. One must make judicious decisions about what to gather and what to ignore, based on the purpose of the data mining exercises. It is like deciding where to fish, as not all streams of data will be equally rich in potential insights.

To learn from data, quality data needs to be effectively gathered, cleaned and organized, and then efficiently mined. One requires the skills and technologies for consolidation and integration of data elements from many sources. Most organizations develop an enterprise data model (EDM) to organize their data. An EDM is a unified, high-level model of all the data stored in an organization’s databases. The EDM is usually inclusive of the data generated from all internal systems. The EDM provides the basic menu of data to create a data warehouse for a decision-making purpose. DWs help organize all this data in an easy and usable manner so that it can be selected and deployed for mining. The EDM can also help imagine what relevant external data should be gathered to provide context and develop good predictive relationships with the internal data. In the United States, the various federal and local governments and their regulatory agencies make a vast variety and quantity of data available at data.gov.

Gathering and curating data take time and effort, particularly when it is un- structured or semi-structured. Unstructured data can come in many forms like databases, blogs, images, videos, audios, and chats. There are streams of un- structured 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 so on. Eventually the data should be *rectangularized,* that is, put in rectangular data shapes with clear columns and rows, before submitting it to data mining.

Knowledge of the business domain helps select the right streams of data for pursuing new insights. Only the data that suits the nature of the problem being solved should be gathered. The data elements should be relevant, and suitably address the problem being solved. They can directly impact the problem or be a suitable proxy for the effect being measured. Select data can also be gathered from the data warehouse. Every industry and function will have its own requirements and constraints. The healthcare industry provides a different type of data with different names. The HR function provides different kinds of data. There may be different issues of quality and privacy for these data.

### DATA CLEANSING AND PREPARATION

The quality of data is critical to the success and value of the data mining project. Otherwise, the situation will be of the kind of garbage in and garbage out (GIGO). The quality of incoming data varies by the source and nature of data. Data from internal operations is likely to be of higher quality, as it will be ac- curate and consistent. Data from social media and other public sources is less under the control of business and is less likely to be reliable.

Data almost certainly needs to be cleansed and transformed before it can be used for data mining. There are many ways in which data may be cleansed – filling missing values, reigning in the effects of outliers, transforming fields, binning continuous variables, and much more – before it can be ready for analysis. Data cleansing and preparation is a labor-intensive or semi-automated activity that can take up to 60-80 percent of the time needed for a data mining project.

*Duplicate Data Needs to be Removed* The same data may be received from multiple sources. When merging the datasets, data must be de-duped.

*Missing Values Need to be Filled in* Missing values can be filled in with average or modal or default values; or those rows should be removed from analysis.

*Data Elements Should be Comparable* Data elements may need to be (a) transformed from one unit to another. For example, total costs of healthcare and the total number of patients may need to be reduced to cost/patient to allow comparability of that value. They may also need to be adjusted to make them

(b) comparable over time. For example, currency value needs to be adjusted for inflation; it requires to be converted to the same base year for comparability. Also, it may require to be converted to a common currency. Data should be (c) stored at the same granularity to ensure comparability. For example, sales data may be available daily, but the salesperson compensation data may only be avail- able monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.

*Continuous Values May Need to be Binned* Into a few buckets to help with some analyses. For instance, work experience can be binned as low, medium, and high.

*Outlier Data Elements Need to be Removed* After careful review, to avoid the skewing of results, outlier data elements are removed. For example, one big donor can skew the analysis of alumni donors in an educational setting.

*Ensure that the Data is Representative of the Phenomena* Under analysis by correcting for any biases in the selection of data. For example, if the data

includes more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.

*Data May Need to be Selected to Increase Information Density* Some data may not show much variability, because it was not properly recorded or for other reasons. This data may dull the effects 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 can serve different types of objectives. The outputs of data mining will reflect the objective being served. There are many ways of rep- resenting the outputs of data mining.

One popular form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually follow the steps to make a model-based decision. The tree may have certain attributes, such as probabilities assigned to each branch. A related format is a set of business rules, which are if-then statements that show causality. A decision tree can be mapped to business rules. If the objective function is prediction, then a decision tree or business rules are the most appropriate mode of representing the output.

The output can be in the form of a regression equation or mathematical function that represents the best fitting curve to represent the data. This equation may include linear and nonlinear terms. Regression equations are a good way of representing the output of classification exercises. These are also a good representation of forecasting formulae.

Population “centroid” is a statistical measure for describing central tendencies of a collection of data points. These might be defined in a multidimensional space. For example, a centroid may be “middle-aged, highly educated, high-net worth professionals, married with two children, living in the coastal areas”. Or a population of “20-something, ivy-league-educated, tech entrepreneurs based in Silicon Valley”. Or it may be a collection of “vehicles more than 20 years old, giving low mileage per gallon, which failed environmental inspection”. These are typical representations of the output of a cluster analysis exercise.

Business rules are an appropriate representation of the output of a market basket analysis exercise. These rules are if-then statements with some probability parameters associated with each rule. For example, those that buy milk and bread will also buy butter (with 80 percent probability).

### EVALUATING DATA MINING RESULTS

There are two primary kinds of data mining processes – supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.

### Predictive Accuracy = Correct Predictions/Total Predictions

Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created. The model is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the other hand, when a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). Similarly, when a true- negative data point is classified as positive, that is classified as a false positive (FP). This is represented using the confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix | | True Class | |
| *Positive* | *Negative* |
| Predicted Class | *Positive* | True Positive (TP) | False Positive (FP) |
| *Negative* | False Negative (FN) | False Positive (FP) |

FIGURE: Confusion Matrix

Thus, the predictive accuracy can be specified by the following formula. Predictive Accuracy = (TP + TN)/(TP + TN + FP + FN)

All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100 percent. In practice, predictive models with more than 70 percent accuracy can be considered usable in business domains, depending upon the nature of the business.

There are no good objective measures to judge the accuracy of unsupervised learning techniques such as Cluster Analysis. There is no single right answer for the results of these techniques. For example, the value of the segmentation model depends upon the value the decision-maker sees in those results.

### DATA MINING TECHNIQUES

Data may be mined to help make more efficient decisions in the future. It may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved.

|  |  |  |
| --- | --- | --- |
| Data Mining Techniques | | |
| Supervised Learning (Predictive ability based on past data) |  | |
| Machine Learning | Decision Trees |
| Neural Network |
| Statistics | Regression |
| Unsupervised Learning (Exploratory analysis to discover patterns) | Clustering Analysis |  |
| Association Rules |

FIGURE: Important Data Mining Techniques

The most important class of problems solved using data mining are classification problems. Classification techniques are called supervised learning as there is a way to supervise whether the model is providing the right or wrong answers. These are problems where data from past decisions is mined to extract the few

rules and patterns that would improve the accuracy of the decision making process in the future. The data of past decisions is organized and mined for decision rules or equations that are then codified to produce more accurate decisions.

Decision trees are the most popular data mining technique for the following reasons

* Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
* Decision trees select the most relevant variables automatically out of all the available variables for decision making.
* Decision trees are tolerant of data quality issues and do not require much data preparation from the users.
* Even non-linear relationships can be handled well by decision trees.

There are many algorithms to implement decision trees. Some of the popular ones are C5, CART and CHAID.

### Regression

It is the most popular statistical data mining technique. The goal of regression is to derive a smooth well-defined curve that best fits the data. Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature. Simply plotting the data may show a non-linear curve. Applying a non-linear regression equation will fit the data very well with high accuracy. Once such regression model has been developed, the energy consumption on any future day can be predicted using this equation. The accuracy of the regression model depends entirely upon the dataset used and not at all on the algorithm or tools used.

### Artificial Neural Networks (ANN)

It is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure – neurons receive stimuli, process them, and communicate their results to other neurons successively, and eventually a neuron outputs a decision. A decision task may be processed by just one neuron and the result may be communicated soon. Alternatively, there could be many layers of neurons involved in a decision task, depending upon the complexity of the domain. The neural network can be trained by deciding repeatedly with many data points. It will continue to learn by adjusting its internal computation and communication parameters

based on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make any intuitive sense to an observer. Thus, the neural networks are considered a black-box system.

At some point, the neural network will have learned enough and begin to match the predictive accuracy of a human expert or alternative classification techniques. The predictions of some ANNs that have been trained over a long period of time with a large amount of data have become decisively more accurate than human experts. At that point, the ANNs can begin to be seriously considered for deployment, in real situations in real time. ANNs are popular because they are eventually able to reach a high predictive accuracy. ANNs are also relatively simple to implement and do not have any issues with data quality. However, ANNs require a lot of data to train it to develop good predictive ability.

### Cluster Analysis

It is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural groupings of things. Data instances that are like (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters. There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number

(K) of clusters from the data.

Clustering is also known as the segmentation technique. It helps divide and conquer large datasets. The technique shows the clusters of things from past data. The output is the centroids for each cluster and the allocation of data points to their cluster. The centroid definition is used to assign new data instances that can be assigned to their cluster homes. Clustering is also a part of the artificial intelligence family of techniques.

### Association Rules

This is a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities. This is the heart of the personalization engine used by e-commerce sites like Amazon.com and streaming movie sites like Net- flix.com. The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form *X* Æ*Y*, where *X* and *Y* are sets of data items. A form of unsupervised learning, it has no dependent variable; there are no right or wrong answers, there are just stronger

and weaker affinities. Thus, each rule has a confidence level assigned to it. A part of the machine learning family, this technique achieved legendary status when a fascinating relationship was found in the sales of diapers and beers.

### Tools and Platforms for Data Mining

Data mining tools have existed for many decades. However, they have recently become more important as the values of data have grown, and the field of big data analytics has come into prominence. There are a wide range of data mining platforms available in the market today. Few are described as follows

*Simple or Sophisticated* There are simple end-user data mining tools such as MS Excel, and there are more sophisticated tools such as IBM SPSS Modeler.

*Stand-alone or Embedded* There are stand-alone tools and those that are embedded in an existing transaction processing or data warehousing or ERP system.

*Open Source or Commercial* There is an open source and freely available tools such as Weka, and there are commercial products.

*User Interface* There are text-based tools that require some programming skills, and there are GUI-based drag-and-drop format tools.

*Data Formats* There are tools that work only on proprietary data formats and there are those that directly accept data from a host of popular data management tools format.

Here we compare three platforms that we have used extensively and effectively for many data mining projects.

MS Excel is a relatively simple and easy data mining tool. It can get quite versatile once Analyst Pack and some other add-on products are installed on it.

Weka is an open-source GUI based tool that offers a large number of data mining algorithms.

R is an extensive and extensible, versatile open-source statistical programming language with 600+ libraries and 120,000 functions. It is very popular with startup companies, and increasingly so in the large organizations.

IBM’s SPSS Modeler is an industry-leading data mining platform. It offers a powerful set of tools and algorithms for most popular data mining capabilities. It has a colorful GUI format with drag-and-drop capabilities. It can accept data in multiple formats including reading Excel files directly.

Table Comparison of Popular Data Mining Platforms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Excel | Weka | R | IBM SPSS  Modeler |
| Ownership | Commercial | Open-source, free | Open-source, free | Commercial, expensive |
| Data mining features | Limited, extensible with add-on modules | Extensive features, issues with large datasets | Extensive features | Extensive fea- tures, unlimited data sizes |
| Stand-alone | Stand-alone | Stand-alone | Embeddable in other systems | Embedded in BI software suites |
| User skills needed | End-users | Skilled BI analysts | Programming skills | For skilled BI analysts |
| User interface | Text and click, easy | GUI, mostly black & white text output | Integrated devel- opment environ- ment | Drag-and-Drop use, colorful, beautiful GUI |
| Data formats | Industry- standard | Proprietary, CSV | CSV | Variety of data sources accepted |

ERP systems include some data analytic capabilities too. SAP has its Business Objects (BO) software. BO is considered one of the leading BI suites in the industry and is often used by organizations that use SAP.

### DATA MINING BEST PRACTICES

Effective and successful use of data mining activity requires both business and technology skills. The business aspects help understand the domain and the key questions. It also helps one imagine possible relationships in the data and create hypotheses to test it. The IT aspects help fetch the data from many sources, clean up the data, assemble it to meet the needs of the business problem, and then run the data mining techniques on the platform.

An important element is to go after the problem iteratively. It is better to divide and conquer the problem with smaller amounts of data and get closer to the heart of the solution in an iterative sequence of steps. There are several best practices learned from the use of data mining techniques over a long period of time. The data mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM). It has six essential steps (Figure 4.3).

1. *Business Understanding* The first and the most important step in data mining is asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise. In other words, selecting a data mining project is like any other project, in that



Business

Understanding

Data

Understanding

Data

Preparation

Deployment

Modeling

Data

Evaluation

FIGURE: CRISP-DM Data Mining Cycle

it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy. A related important step is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in terms of a proposed model as well in the datasets available and required.

1. *Data Understanding* A related important step is to understand the data available for mining. One needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem. Without relevant data, the hypotheses cannot be tested.
2. *Data Preparation* The data should be relevant, clean and of high quality. It’s important to assemble a team that has a mix of technical and business skills, who understands the domain and the data. Data cleaning can take 60-70 percent of the time in a data mining project. It may be desirable to continue to experiment and add new data elements from external sources of data that can help improve predictive accuracy.
3. *Modeling* This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported. Patience is required in continuously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used. A tool may be tried with different options, such as running different decision tree algorithms.
4. *Model Evaluation* One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques, and conducting many what-if scenarios, to build confidence in the solution. One should evaluate and improve the model’s predictive accuracy with more test data. When the accuracy has reached some satisfactory level, then the model should be deployed.
5. *Dissemination and Rollout* It is important that the data mining solution is presented to the key stakeholders, and is deployed in the organization. Otherwise the project will be a waste of time and a setback for establishing and supporting a data-based decision-process culture in the organization. The model should be eventually embedded in the organization’s business processes.

### MYTHS ABOUT DATA MINING

There are many myths about this area, scaring away many business executives from using data mining. Data Mining is a mindset that presupposes a faith in the ability to reveal insights. By itself, data mining is neither too hard nor too easy. It does require a disciplined approach and some cross-disciplinary skills.

*Myth #1* Data Mining is about algorithms. It is used by businesses to answer important and practical questions. Formulating the problem statement correctly and identifying imaginative solutions for testing are far more important before the data mining algorithms gets called in. Understanding the relative strengths of various algorithms is helpful but not mandatory.

*Myth #2* Data Mining is about predictive accuracy. While important, predictive accuracy is a feature of the algorithm. As in myth#1, the quality of output is a strong function of the right problem, right hypothesis, and the right data.

*Myth #3* Data Mining requires a data warehouse. While the presence of a data warehouse assists in the gathering of information, sometimes the creation of the data warehouse itself can benefit from some exploratory data mining. Some data mining problems may benefit from clean data available directly from the DW, but a DW is not mandatory.

*Myth #4* Data Mining requires large quantities of data. Many interesting data mining exercises are done using small or medium sized datasets, at low costs, using end-user tools.

*Myth #5* Data Mining requires a technology expert. Many interesting data mining exercises are done by end-users and executives using simple everyday tools like spreadsheets.

### DATA MINING MISTAKES

Data mining is an exercise in extracting non-trivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide. Much domain knowledge, tools and skills are required to find such patterns. Here are some of the more common mistakes in doing data mining and should be avoided.

*Mistake #1 Selecting the Wrong Problem for Data Mining* Without the right goals or having no goals, data mining leads to a waste of time. Getting the right answer to an irrelevant question could be interesting, but it would be point- less from a business perspective. A good goal would be one that would deliver a good ROI to the organization.

*Mistake #2 Buried Under Mountains of Data without Clear Metadata* It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought. There may be insufficient knowledge about the data, or metadata. Examine the data with a critical eye and do not naively believe everything you are told about the data.

*Mistake #3 Disorganized Data Mining* Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy. This can come from being sloppy about keeping track of the data mining procedure and results. Not leaving enough time for data acquisition, selection and preparation can lead to data quality issues, and GIGO. Similarly, not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

*Mistake #4 Insufficient Business Knowledge* Without a deep understanding of the business domain, the results would be gibberish and meaningless. Don’t make erroneous assumptions, courtesy of experts. Don’t rule out anything when observing data analysis results. Don’t ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even

when insights emerge at one level, it is important to slice and dice the data at other levels to see if more powerful insights can be extracted.

*Mistake #5 Incompatibility of Data Mining Tools and Datasets* All the tools from data gathering, preparation, mining, and visualization, should work together. Use tools that can work with data from multiple sources in multiple industry standard formats.

*Mistake #6 Looking only at Aggregated Results and not at Individual Records/Predictions* It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level. Diving into the data at the right angle can yield insights at many levels of data.

*Mistake #7 Not measuring your results differently from the way your sponsor measures them* If the data mining team loses its sense of business objectives, and beginning to mine data for its own sake, it will lose respect and executive support very quickly. The BIDM cycle (Figure 1.1) should be remembered.

## Conclusion

Data Mining is like diving into the rough material to discover a valuable finished nugget. While the technique is important, domain knowledge is also important to provide imaginative solutions that can then be tested with data mining. The business objective should be well understood and should always be kept in mind to ensure that the results are beneficial to the sponsor of the exercise.

## Questions

1. What is data mining? What are supervised and unsupervised learning techniques?

The art and science of discovering knowledge, insights, and patterns in data.

It is a multidisciplinary field that borrows techniques from a variety of fields.

1. Describe the key steps in the data mining process. Why is it important to follow these processes?
2. What is a confusion matrix?
3. Why is data preparation so important and time consuming?
4. What are some of the most popular data mining techniques?
5. What are the major mistakes to be avoided when doing data mining?
6. What are the key requirements for a skilled data analyst?

## True/False

1. Data mining seeks to identify patterns in the data.
2. Data mining is about using the algorithms with the highest predictive ac- curacy to model and solve a business problem.
3. Data cleansing and preparation for data mining takes about one-third of the time of the entire data mining project.
4. The first and most important practice in data mining is to understand the sources and quality of data.
5. The goal of supervised learning is to analyze the historical data stored in a database and automatically generate a model that can predict future behavior.
6. Regression equations are a good way of representing the output of clustering exercises.
7. Predictive Accuracy = Correct Predictions/Total Predictions
8. CRISP-DM is the name of a data modeling tool.
9. Neural networks are a versatile data mining technique.
10. Decision trees can generate knowledge from a few instances that can then be applied to a broad population.