CHAPTER – I

DATA MINING AND KDD PROCESS FOR IMPUTATION
1.1 Introduction

Modern science and engineering are based on using first-principle models to describe physical, biological, and social systems. Such an approach starts with a basic scientific model, such as Newton's Laws of Motion or Maxwell's Equations in electromagnetism, and then builds upon various applications in mechanical engineering and electrical engineering. In this approach, experimental data is used to verify the underlying first-principle models and to estimate some of the parameters that are difficult or sometimes impossible to measure directly. However, in many domains the underlying first principles are unknown, or the systems under study are too complex to be mathematically formalized. With the growing use of computers, there is a great amount of data being generated by such systems. In the absence of first-principle models, such readily available data can be used to derive models by estimating useful relationships between a system's variables (i.e., un-known input—output dependencies). Thus currently there is a paradigm shift from classical modeling and analyses based on first principles to developing models and the corresponding analyses directly from data.

We are gradually accepting to the fact that there are tremendous volumes of data filling our computers, networks, and lives. Government agencies, scientific institutions, and businesses have all dedicated enormous resources to collecting and storing data. In reality, only a small amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analyzed effectively. How could this happen? The primary reason is that the original effort to create a data set is often focused on issues such as storage efficiency; it does not include a plan as to how the data will eventually be used and analyzed.

The need to understand large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering. In the
business world, corporate and customer data are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining [50, 112].

Data mining is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, knowledge acquisition, information retrieval, high performance computing, and data visualization. Data mining emerged during the late 1980's, has made great strides during the 1990's, and is expected to continue to nourish into the new millennium.

The recent advances in data mining have produced algorithms for extracting knowledge contained in large data sets. This knowledge can be explicit, e.g., represented as decision rules, and utilized for decision making in areas where decision models do not exist [7]. The machine learning algorithms construct associations among various parameters (called features in data mining and attributes in computer database literature) that are important in modeling processes. Selection of algorithms is an essential context for a data mining framework design. The widespread use of algorithmic trading has led to the question of whether the most suitable algorithm is always being used [34, 47, 66, 109]. Evolutionary phenomena have propounded a most ardent need in the development of algorithms based on the transactions of nature and evolution [2, 47].

Data mining refers to extracting or “mining” knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, “data mining” should have been more appropriately named “knowledge mining from data”, which is unfortunately somewhat long.
"Knowledge mining", a shorter term, may not reflect the emphasis on mining from large amounts of data [50].

There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Rational decisions are usually based on the analysis and evaluation of problem-specific data. Today there is an increasing amount of such data available due to its extensive generation e.g. by surveys or simulation studies and an advanced information infrastructure [39]. At the same time many decisions e.g. in business or politics are embedded in a growing system of dependencies and lead to far-reaching consequences so that not only the possibility but also the need for substantiated data analysis rises. This means many decisions have to be made within increasingly complex and wide information spaces. To discover relevant facts and patterns therein which support or oppose an argument a wealth of methods has been developed, e.g. in the fields Knowledge Discovery in Databases or data mining.

"Knowledge Discovery in Databases", or KDD. Alternatively, some view data mining as simply an essential step in the process of knowledge discovery in databases.

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.
In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into any one of the two following categories:

i. Predictive data mining, which produces the model of the system described by the given data set, or

ii. Descriptive data mining, which produces new, nontrivial information based on the available data set.

As, hard problems force innovative approaches and attention to detail, their exploration often contributing beyond the area initially attempted. Data mining process has also resulted in developing predictors for numerical series [32].

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably. The goals of prediction and description are achieved by using data-mining techniques for the following primary data-mining tasks:

i. Classification – discovery of a predictive learning function that classifies a data item into one of several predefined classes.

ii. Regression – discovery of a predictive learning function, which maps a data item to a real-value prediction variable.
iii. Clustering – a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.

iv. Summarization – an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.

v. Dependency Modeling – finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.

vi. Change and Deviation Detection – discovering the most significant changes in the data set.

The success of a data-mining engagement depends largely on the amount of energy, knowledge, and creativity that the designer puts into it. In essence, data mining is like solving a puzzle. The individual pieces of the puzzle are not complex structures in themselves. Taken as a collective whole, however, they can constitute very elaborate systems. As you try to unravel these systems, you will probably get frustrated, start forcing parts together and generally become annoyed at the entire process; but once you know how to work with the pieces, you realize that it is not really that hard in the first place. The same analogy can be applied to data mining. In the beginning, the designers of the data-mining process probably do not know much about the data sources; if they did, they would most likely not be interested in performing data mining. Individually, the data seem simple, complete, and explainable. But collectively, they take on a whole new appearance that is intimidating and difficult to comprehend, like the puzzle. Therefore, being an analyst and designer in a data-mining process requires, besides thorough professional knowledge, creative thinking and a willingness to see problems in a different light.

Data mining is one of the fastest growing fields in the computer industry. Once a small interest area within computer science and statistics, it has quickly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets. Since data mining is the entire data
warehousing, data-mart, and decision-support community, encompassing professionals from such industries as retail, manufacturing, telecommunications, healthcare, insurance, and transportation. In the business community, data mining can be used to discover new purchasing trends, plan investment strategies, and detect unauthorized expenditures in the accounting system. It can improve marketing campaigns and the outcomes can be used to provide customers with more focused support and attention. Data-mining techniques can be applied to problems of business process reengineering, in which the goal is to understand interactions and relationships among business practices and organizations.

Many law enforcement and special investigative units, whose mission is to identify fraudulent activities and discover crime trends, have also used data mining successfully. For example, these methodologies can aid analysts in the identification of critical behavior patterns in the communication interactions of narcotics organizations, the monetary transactions of money laundering and insider trading operations, the movements of serial killers, and the targeting of smugglers at border crossings. Data-mining techniques have also been employed by people in the intelligence community who maintain many large data sources as a part of the activities relating to matters of national security.

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and scientific exploration.

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving
data warehousing and data mining). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, data analysis and understanding has naturally become the next target.

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [50]. A hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it.

Many organizations have a pressing need to manipulate all the data from their different branches rapidly and reliably. This need is very difficult to satisfy when the data is stored in many independent databases, and the data is all of importance to an organization. Formulating and implementing queries requires data from more than one database [97].

The increasing use of multi-database technology, such as computer communication networks and distributed, federated and homogeneous multi-database systems, has led to the development of many multi-database systems for real world applications. For decision-making, large organizations need to mine the multiple databases distributed throughout their branches. In particular, as the Web is rapidly becoming an information flood, individuals and organizations can take into account low-cost information and knowledge on the Internet when making
decisions. The data of a company is referred to as internal data whereas the data collected from the Internet is referred to as external data. KDD algorithms in multidimensional databases are often based on similarity queries which are performed for a high number of objects [19].

A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. An application of clustering techniques to develop a hybrid methodology that imputes data by overcoming the limitations faced by the existing methods has been discussed in [18, 105]. The selection process of the donors implemented in our methodology is significant as it liberates the methodology from the hurdles caused by the inherent characteristics of the data set, a hybrid methodology to overcome the limitations in most existing imputation methods. It was designed by taking into aspect the missing mechanism, data set size, missing percentage and the pattern in which the data are missing. We implemented it on six different real-time software project data sets and evaluated its performance.

Moreover, it is not computationally demanding and can work extremely well with all sizes of data sets. Due to the very nature of the method to form homogenous clusters, the missing pattern or the missing mechanism causes no degradation in its performance. However, the only factor that can influence its performance is the high percentage of missing data. Though the results drawn from the six data sets gave us confidence on the conclusions, it needs to be tested extensively on a larger number of data sets having high missing percentages (> 40 percent). Also, the criterion used in the clustering algorithm and its complexity may be improved. Even Knowledge Discovery has ushered into various types of databases that are historical and related to time.

Knowledge discovery in temporal databases has been an active field of study. Many existing algorithms focus on temporal association rule mining. However, little work has been done on discovering infrequent episodes, and it is not easy to discover association rules with low support but high confidence by
existing algorithms. This work presents a new approach for discovering infrequent (consequent) episodes in time series databases [22].

The evolution of database technology is an essential prerequisite for understanding the need of knowledge discovery in databases (KDD). Data mining is a pivotal step in the KDD process – the extraction of interesting patterns from a set of data sources (relational, transactional, object-oriented, spatial, temporal, text, and legacy databases, as well as data warehouses and the World Wide Web). The patterns obtained are used to describe concepts, to analyze associations to build classification and regression models, to cluster data, to model trends in time-series, and to detect outliers. Since the patterns which are present in data are not all, equally useful, interestingness measures are needed to estimate the relevance of the discovered patterns to guide the mining process [73, 86].

1.2. Data Mining and Decision Making

The study of differences between Data Mining and Decision Making Systems would be an impulsive attempt for Researchers in the Database Technology. Data Mining systems are mixed systems comprising of principles from various disciplines like, Mathematics, Relational Algebra, Neural Networks, Simulated Annealing, Optimization Techniques, Genetic Algorithms, Biological and Evolutionary Phenomena. Decision Making Systems are derivative of traditional database systems just by having variants in the types of databases like knowledge databases, deductive databases, analytical databases, etc.

Scientific analysis of decision problems aims at giving the decision maker (DM) a recommendation concerning a set of objects (called also alternatives, solutions, acts, actions, cases, candidates) evaluated from the point of view of a plurality of their characteristics considered relevant for the problem at hand, and called attributes [94].
For example, a decision can regard:

i. Diagnosis of pathologies for a set of patients, being the objects of the decision, and the attributes are symptoms and results of medical examinations.

ii. Assignment to classes of risk for a set of enterprises, being the objects of the decision, and the attributes are ratio indices and other economical indicators such as the market structure, the technology used by the enterprises, the quality of the management and so on.

iii. Selection of a car to be bought from among a given set of cars, being the objects of the decision, and the attributes are maximum speed, acceleration, price, fuel consumption and so on.

iv. Ordering of students applying for a scholarship, being the objects of the decision, and the attributes are scores in different disciplines.

The volume of information available for decision-making is growing at an unprecedented rate. Models, Algorithms, and tools offered by decision-making theory, operations research, mathematical programming, and other disciplines have met some of these growing needs for data processing. The expanding scope of decision-making problems calls for a new class of computational tools. Data mining satisfies some of these needs by offering capabilities to process data of different types (e.g., qualitative and quantitative) and originating at different sources. None of the traditional approaches is able to cover such a wide spectrum of data diversity.

1.2.1 Data Mining Framework for Decision Making

Models and algorithms for effective decision-making in a data-driven environment are discussed in earlier works. To enhance the quality of the extracted knowledge and decision-making, the data sets are transformed; the knowledge is extracted with multiple algorithms [5].
The problems considered in this area differ from most data mining tasks where knowledge is extracted and used to assign decision values to the new objects that have not been included in the training data. For example, the equipment fault is recognized (i.e., the value of the fault number is assigned) based on the failure symptoms. There are many applications prevailing in this area, where a subset of rules, in particular a single rule, is selected from the extracted knowledge. The parameter values corresponding to the conditions of the rules in this subset are called a decision signature. The decision signature is used to control the process under consideration. One of the questions posed in this area of research is how to construct the most promising decision signatures for large-scale rule bases. The construction of such decision signatures becomes a challenge due to the temporal nature of the processes from which the data sets considered in this research have been collected.

The process of decision making through data mining is structured into six phases. They are 1) Data transformation, 2) Rule extraction with alternative
algorithms, 3) Decision signature selection, 4) Decision signature validation, 5) A return to the data transformation phase, 6) The acceptance of the decision signature.

The complexity of decision-making in manufacturing, business, and medical applications is rapidly increasing, as the world is becoming data-driven. To cope with this increasing complexity in a changing environment, new modeling and computing paradigms are needed.

The salient features of the new modeling and computing paradigm are:

- Adaptability of decision-making models to the evolving decision environment.
- Ability to handle changing qualitative and quantitative data.
- Short decision response time.
- Large and overlapping problem domains.
- Interpretability of the results.
- Process rather than problem orientation.

Decision making often depends on the analysis and evaluation of large amounts of data for which information visualization proved to be a valuable approach [39].

1.2.2 Potential Applications

This theory of decision making through data mining is potential for certain applications of:

i. Aluminum processing plants are interested in production of high quality aluminum products. Though numerous neural controllers are involved in the production process, the values of most of the parameters are manually set. The decision signatures to be generated by the algorithms proposed in this research will provide the values for these parameter settings.
ii. *Semiconductor manufacturers* are interested in improving the quality of wafers. So far, due to unknown reasons under seemingly similar manufacturing conditions, some wafers attain perfect quality and others are not acceptable. Even within individual plants product quality varies. The concept of decision signatures is to determine the ranges of control parameters leading to the production of wafers of the desired quality.

iii. *Electronic assembly processes* face quality problems with printed circuit boards where assemblies rather than components fail the quality test for unknown reasons. The management is not satisfied with the current process control or with other tools that provide solutions for a "population" of products rather than an "individual" product. The managers would like to predict circumstances under which an individual product (object) might fail, and thereby prevent this failure. The ideas proposed in this research will utilize the rules extracted from a population of objects to generate knowledge, which will determine conditions preventing the production of faulty products.

iv. *DNA manufacturing* is the most recent addition to the topics of interest to the approach discussed in this paper. The DNA manufacturing process has evolved from a biological laboratory to an industrial-scale process and it involves many unknowns. Finding ways of improving product quality by generating decision signatures that would determine process control parameters are of interest to DNA manufacturers.

v. *Medical applications* call for analysis of vast amount of data and recommend medical actions, e.g., for critically ill infants after open-heart surgery on a minute-to-minute basis. Even the most skilled health professionals have difficulty understanding the relationships between dozens of parameters and translating these relationships into consistent treatment.

Framework defined in [3] and [110] have also been studied at the key interest of developing the proposed framework in this research.
Besides the hard (formally defined constraints), the algorithms in data mining framework consist of soft constraints leading to the following knowledge properties:

i. \textit{Knowledge diversity}: The knowledge to be used for structuring will be diverse due to the data transformation schemes applied before knowledge extraction and the use of algorithms of different types for knowledge extraction.

ii. \textit{Knowledge interestingness}: Knowledge is interesting if it is unexpected (surprising to the user) and actionable (the user can do something with it).

iii. \textit{Knowledge comprehensibility}: A user should be able to comprehend and have confidence in the extracted knowledge regardless of his/her own background.

iv. \textit{Customer Satisfaction research} is one of the fastest growing segments of the marketing field. Marketing and management sciences, nowadays, are focusing on the coordination of all the organization's activities in order to provide goods or services that can satisfy best specific needs of existing or potential customers [77]. To reinforce customer orientation on a day-to-day basis, a growing number of companies choose customer satisfaction as their main performance indicator. However, it is almost impossible to keep an entire company permanently motivated by a notion as abstract and intangible as customer satisfaction. Therefore, customer satisfaction must be translated into a number of measurable parameters directly linked to people's job-in other words factors that people can understand and influence. Nowadays, every company faces to an issue how to focus management resources to a core business to maintain a growing position in the global market [108].

\textbf{1.3. Statistical Methods and Data Mining}

Since antiquity the intuitive notions of continuous change, growth, and motion, have challenged scientific minds. Yet, the way to the understanding of
continuous variations in systems and perceiving various dimensions of data was found only in late 1980s when modern computer science emerged and rapidly developed in close conjunction with electronics and allied sciences. One of the key issues in data mining is size [61]. A typical data mining problem involves a large database from where one seeks to extract useful knowledge.

Data come in many forms and the place to develop a complete taxonomy persists with data preservation, and data retrieval, data mining. Indeed, it is not even clear that a complete taxonomy can be developed, since an important aspect of data in one situation may be unimportant in another. The basic notions of data and their derivatives: the derivative projects the rate of change meaning with increasing dimensions in the data. A precise understanding of these concepts and their overwhelming fruitfulness rests upon these concepts of scope of the data mining experiments and that of the experiments which in turn depend upon an understanding of the statistics and data model. Nonetheless, only gradually, by penetrating more and more into the substance of primitive statistics and data modeling, can one appreciate its power and beauty.

Early studies of data mining include incorporation of statistical methods which generate the cryptic statistical inference that only projects a syntactic structure than expressing sensible semantic features hidden within data. Database Management Systems have contributed much of statistical functionalities in varied versions that contain techniques of different flavors for data storage, data retrieval, data presentation etc. Relational database systems provide five built-in aggregate functions: count(), sum(), avg(), max(), and min(). These functions can also be computed efficiently (in incremental and distributed manners) in data cubes. Thus, there is no problem in including these aggregate functions as basic measures in the descriptive mining of multidimensional data.

However, for many data mining tasks, users would like to learn more data characteristics regarding both central tendency and data dispersion. Measures of central tendency include mean, median, mode, and midrange, while measures
of data dispersion include quartiles, outliers, variance, and other statistical measures. These descriptive statistics are of great help in understanding the distribution of the data. Such measures have been studied extensively in the statistical literature. However, from the data mining point of view, we need to examine how they can be computed efficiently in large, multidimensional databases.

The analysis and modeling of incomplete data poses special challenges, for example, in the estimation of covariance matrices. Covariance matrices are important because every commonly used multivariate analysis be it regression analysis, principal component analysis, discriminant analysis, or canonical correlation analysis issues from estimates thereof. Yet, their estimation from incomplete data is not straightforward. Covariance matrices estimated from all available data, leaving out missing values in sums of products and cross products of variables, may not be positive semi-definite, potentially causing problems in multivariate analyses. Covariance matrices estimated as the usual sample covariance matrices from a dataset with missing values filled in with imputed values usually are biased: if the imputed values come from the center (e.g., the mean) of a distribution of possible values, the variation of the missing values about the center of the distribution is ignored. Since the estimation of statistics from incomplete data and the imputation of missing values are closely related problems given the statistics and available data, expected values of the missing values can be calculated any inaccuracy in the estimation of statistics such as covariance matrices translates into inaccuracies in imputed values. Missing Values are values of instances that are not observed [58, 113, 116].

As in other fields, many heuristics have been developed to deal with incomplete data in atmosphere, ocean, and climate science, largely without reference to a unifying framework such as that of maximum likelihood estimation that would establish generally, for example, when estimates of variances are unbiased or how to estimate confidence intervals for missing values. No textbook covers the specific challenges that analysis of incomplete data poses in our field,
in which datasets are typically large, variables are highly correlated in space and in time, and the number of variables often exceeds the sample size, such that sample covariance matrices are singular. However, statistical methods for dealing with incomplete data have developed rapidly in the past decades, and there are excellent surveys of the subject in the statistics literature.

Little and Rubin’s *Statistical Analysis with Missing Data, Second Edition* (2002, Wiley) is a classic text written by authors who worked out much of the theoretical foundation for analyses of incomplete data and contributed several practical methods. The book covers fundamental concepts and methods clearly and thoroughly, and the substantially expanded second edition also covers in depth more recent developments such as Bayesian [11, 91] methods and multiple imputation [117]. It begins with a discussion of the concept of values that are missing at random, which means that the probability that a value is missing is independent of the missing value the central necessary condition for mechanisms responsible for missingness to be ignorable in analyses of incomplete data. (This is the assumption in question in the recent discussions of how well historic temperatures can be reconstructed from sparse measurements or proxies, given that temperatures and the availability of data may be correlated.) The conceptual foundations serve as a point of departure for discussions of heuristic methods for estimating statistics such as means and covariance matrices and for imputing missing values, of re-sampling methods (bootstrap and jackknife) for estimating uncertainties in imputed values, and of maximum-likelihood methods and their properties. The expectation–maximization (EM) algorithm and variants for the computation of maximum likelihood estimates of statistics and missing values are discussed extensively, including discussions of properties (e.g., convergence rates) that are important in applications.

What may turn out to be particularly relevant for atmosphere, ocean, and climate science are Bayesian and closely related multiple imputation [81] methods, to which Little and Rubin devote a chapter. If the number of variables in a dataset exceeds or is only marginally smaller than the sample size, covariance
matrices are singular or nearly singular, and standard estimates of missing values are not unique or not stable. Additional information needs to be introduced to regularize the estimates to make them unique or stable which can be done in a Bayesian framework by introducing prior information. (Bayesian methods for normal data can also be typically justified on geometric or regularity grounds, as is common in the applied mathematics literature; the books discussed here focus on the Bayesian perspective.) Because second and higher-order statistics cannot be reliably estimated from a completed dataset with a single imputed value filled in for each missing value (as mentioned above, possible variations of the missing values about the imputed values would be ignored), multiple imputation methods in which several completed datasets are generated with missing values drawn from a posterior distribution of possible values are attractive if the completed datasets are to be archived for use by other researchers.

Subsequent analyses can be performed on each of the completed datasets, and, as in other ensemble methods, the results can be combined to obtain inferences that reflect uncertainties in the imputed values. Often, only a few completed datasets are necessary to obtain reliable estimates, for example, of variances, so archiving a few completed datasets can be much more efficient than archiving one completed dataset plus often large covariance matrices and, possibly, information about higher-order statistics. Little and Rubin’s survey of Bayesian and multiple imputation methods provide a good introduction to make these methods fruitful for our field.

1.3.1 Measuring the central tendency

The most common and most effective numerical measure of the “center” of a set of data is the (arithmetic) mean. Let \( x_1, x_2, \ldots, x_n \) be a set of \( n \) values or observations. The mean of this set of values is

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
This corresponds to the built-in aggregate function, average (avg() in SQL), is provided in relational database systems. In most data cubes, sum and count are saved in precomputation. Thus, the derivation of average is straightforward, using the formula \( \text{average} = \frac{\text{sum}}{\text{count}}. \)

Sometimes, each value \( x_i \) in a set may be associated with a weight \( w_i \), for \( i = 1, \ldots, n \). The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

\[
X = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}
\]

This is called the weighted arithmetic mean or the weighted average.

Although the mean is the single most useful quantity that we use to describe a set of data, it is not the only, or even always the best, way of measuring the center of a set of data. For skewed data, a better measure of center of data is the median, \( M \). Let us assume that the values forming a given set of data are in numerical order.

The median is the middle value of the ordered set if the number of values \( n \) is an odd number; otherwise (i.e., if \( n \) is even), it is the average of the middle two values \([93, 102]\).

Although it is not easy to compute the exact median value in a large database, an approximate median can be computed efficiently. For example, for grouped data, the median, obtained by interpolation, is given by

\[
\text{median} = L_1 + \left( \frac{n/2 + \sum f_i}{f_{\text{median}}} \right) c
\]

19
where $L_j$ is the lower class boundary of (i.e., lowest value for) the class containing the median, $n$ is the number of values in the data, $(\Sigma f_i)l$ is the sum of the frequencies of all of the classes that are lower than the median class, and $f_{median}$ is the frequency of the median class, and $c$ is the size of the median class interval.

Another measure of central tendency is the mode. The mode for a set of data is the value that occurs most frequently in the set. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. If a data set has more than three modes, it is multimodal. At the other extreme, if each data value occurs only once, then there is no mode.

For unimodal frequency curves that are moderately skewed (asymmetrical), we have the following empirical relation:

$$mean \approx mode = 3 \times (mean \approx median)$$

This implies that the mode for unimodal frequency curves that are moderately skewed can easily be computed if the mean and median values are known.

The midrange, that is, the average of the largest and smallest values in a data set, can be used to measure the central tendency of the set of data. It is trivial to compute the midrange using the SQL aggregate functions, max() and min().

### 1.4. Data Mining appreciates in Various Domains

#### 1.4.1 Data Mining For the Telecommunications Industry

The telecommunication industry has quickly evolved from offering local and long-distance telephone services to providing many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and Web-data transmission, and other data traffic. The integration of telecommunications, computer networks, Internet, and numerous others means
of communication and computing is underway. The U.S. Telecommunication Act of 1996 allowed Regional Bell Operating Companies to enter the long-distance market as well as offer "cable-like" services. The European Liberalization of Telecommunications Services has been effective since the beginning of 1998. Besides deregulation, there has been a sale by the Federal Communication Commission (FCC) of airwaves to companies pioneering new ways to communicate. The cellular industry is rapidly taking on a life of its own. With all this deregulation of the telecommunication industry, the market is expanding rapidly and becoming highly competitive.

The hypercompetitive nature of the industry has created a need to understand customers, to retain them, and to model effective ways to market new products. This creates a great demand for data mining to help understand the new business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services. In general, the telecommunications industry is interested in answering some strategic questions through data-mining applications such as:

i. How does one retain customers and keep them loyal as competitors offer special offers and reduced rates?

ii. Which customers are most likely to churn?

iii. When is a high-risk investment, such as new fiber optic lines, acceptable?

iv. How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services?

v. What characteristics differentiate our products from those of our competitors?

Companies like AT&T, AirTouch Communications, and AMS Mobile Communication Industry Group have announced the use of data mining to improve their marketing activities. There are several companies including Lightbridge and Verizon that use data-mining technology to look at cellular fraud...
for the telecommunications industry. Another trend has been to use advanced visualization techniques to model and analyze wireless-telecommunication networks. Selected examples of data-mining applications in the telecommunication industry follow.

1.4.2 Data Mining For the Retail Industry

Slim margins have pushed retailers into data warehousing earlier than other industries. Retailers have seen improved decision-support processes leading directly to improved efficiency in inventory management and financial forecasting. The early adoption of data warehousing by retailers has allowed them a better opportunity to take advantage of data mining. The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer-shopping history, goods transportation, consumption patterns, service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing availability and popularity of business conducted on the Web, or e-commerce. Today, even many stores also have Web sites where customers can make purchases online. A variety of sources and types of retail data provide a rich source for data mining.

Retail data mining can help identify customer-buying behaviors, discover customer-shopping patterns and trends, improve the quality of customer services, achieve better customer retention and satisfaction, enhance goods consumption, design more effective goods transportation and distribution policies, and, in general, reduce the cost of business and increase profitability. In the forefront of applications that have been adopted by the retail industry are direct-marketing applications. The direct-mailing industry is an area where data mining is widely used. Almost every type of retailer uses direct marketing, including catalogers, consumer retail chains, grocers, publishers, B2B marketers, and packaged goods manufacturers. The claim could be made that every Fortune 500 company has used some level of data mining in their direct-marketing campaigns. Large retail chains and grocery stores use vast amounts of sale data that is
"information rich". Direct marketers are mainly concerned about customer segmentation, which is a clustering or classification problem.

Retailers are interested in creating data-mining models to answer questions such as:

i. What are the best types of advertisements to reach certain segments of customers?

ii. What is the optimal timing at which to send mailers?

iii. What is the latest product trend?

iv. What types of products can be sold together?

v. How does one retain profitable customers?

vi. What are the significant customer segments that buy products?

Data mining helps to model and identify the traits of profitable customers, and it also helps to reveal the "hidden relationship" in data that standard query processes have not found. IBM has used data mining for several retailers to analyze shopping patterns within stores based on point-of-sale (POS) information. For example, one retail company with $2 billion in revenue, 300,000 UPC codes, and 129 stores in 15 States arrived at some interesting results: "...we found that people who were coming into the shop gravitated to the left-hand side of the store for promotional items, and they were not necessarily shopping the whole store". Such information is used to change promotional activities and provide a better understanding of how to lay out a store in order to optimize sales. Additional real-world examples of data-mining systems in retail industry follow.

1.4.3 Data Mining In Healthcare and Biomedical Research

Owing to amount of information and issues in the healthcare industry, not to mention the pharmaceutical industry and biomedical research, opportunities for data-mining applications are extremely widespread, and benefits from the results are enormous. Storing patients' records in electronic format and the development in medical-information systems cause a large amount of clinical data
to be available online. Regularities, trends, and surprising events extracted from these data by data-mining methods are important in assisting clinicians to make informed decisions, thereby improving health services.

Clinicians evaluate a patient's condition over time. The analysis of large quantities of time-stamped data will provide doctors with important information regarding the progress of the disease. Therefore, systems capable of performing temporal abstraction and reasoning become crucial in this context. Although the use of temporal-reasoning methods requires an intensive knowledge-acquisitions effort, data mining has been used in many successful medical applications, including data validation in intensive care, the monitoring of children's growth, analysis of diabetic patient's data, the monitoring of heart-transplant patients, and intelligent anesthesia monitoring.

Data mining has been used extensively in the medical industry. Data visualization and artificial neural networks are especially important areas of data mining applicable in the medical field.

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome. The logic behind investigating the genetic causes of disease is that once the molecular bases of diseases are known, precisely targeted medical interventions for diagnostics, prevention, and treatment of the disease itself can be developed. Much of the work occurs in the context of the development of new pharmaceutical products that can be used to fight a host of diseases ranging from various types of cancers to degenerative disorders such as Alzheimer's Disease.

A great deal of biomedical research has focused on DNA-data analysis, and the results have led to the discovery of genetic causes for many diseases and disabilities. An important focus in genome research is the study of DNA sequences since such sequences form the foundation of the genetic codes of all
living organisms. What is DNA? Deoxyribonucleic acid, or DNA, forms the foundation for all living organisms. DNA contains the instructions that tell cells how to behave and is the primary mechanism that permits us to transfer our genes to our offspring. DNA is built in sequences that form the foundations of our genetic codes, and that are critical for understanding how our genes behave. Each gene comprises a series of building blocks called nucleotides. When these nucleotides are combined, they form long, twisted, and paired DNA sequences or chains. Unraveling these sequences has become a challenge since the 1950s when the structure of the DNA was first understood. If we understand DNA sequences, theoretically, we will be able to identify and predict faults, weaknesses, or other factors in our genes that can affect our lives. Getting a better grasp of DNA sequences could potentially lead to improved procedures to treat cancer, birth defects, and other pathological processes. Data-mining technologies are any one weapon in the arsenal used to understand these types of data, and the use of visualization and classification techniques is playing a crucial role in these activities.

It is estimated that humans have around 100,000 genes, each one having DNA that encodes a unique protein specialized for a function or a set of functions. Genes controlling production of haemoglobin, regulation of insulin, and susceptibility to Huntington's chorea are among those that have been isolated in recent years. There are seemingly endless varieties of ways in which nucleotides can be ordered and sequenced to form distinct genes. Any one gene might comprise a sequence containing hundreds of thousands of individual nucleotides arranged in a particular order. Furthermore, the process of DNA sequencing used to extract genetic information from cells and tissues usually produces only fragments of genes. It has been difficult to tell using traditional methods where these fragments fit into the overall complete sequence from which they are drawn. Genetic scientists face the difficult task of trying to interpret these sequences and form hypotheses on which genes they might belong to, and the disease processes that they may control. The task of identifying good candidate gene sequences for
further research and development is like finding a needle in a haystack. There can be hundreds of candidates for any given disease being studied.

Therefore, companies must decide which sequences are the most promising ones to pursue further development. How do they determine which ones would make good therapeutic target? Historically, this has been a process based largely on trial and error. For every lead that eventually turns into a successful pharmaceutical intervention that is effective in clinical settings, there are dozens of others that do not produce the anticipated results. This is a research area that is crying out for innovations that can help to make these analytical processes more efficient. Since pattern analysis, data visualization, and similarity-search techniques have been developed in data mining, this field has become a powerful infrastructure for further research and discovery in DNA sequences.

Data mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data. In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available [12]. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. A major objective of this thesis is to evaluate data mining tools in medical and health care applications to develop a tool that can help make timely and accurate decisions.

1.4.4 Data Mining in Science and Engineering

We live in a database culture. Our libraries are data warehouses. Vendor software tabulates which of our databases are accessed and which articles are requested. Our online catalog records the circulation of items in the collection [27]. Our server tracks the navigation of visitors to our library’s Web pages. We passively accumulate more data than we use.
Enormous amounts of data have been generated in science and engineering, e.g., in cosmology, molecular biology, and chemical engineering. In cosmology, advanced computational tools are needed to help astronomers understand the origin of large-scale cosmological structures as well as the formation and evolution of their astrophysical components (galaxies, quasars, and clusters). Over three terabytes of image data have been collected by the Digital Palomar Observatory Sky Survey, which contain on the order of two billion sky objects.

In molecular biology, recent technological advances are applied in such areas as molecular genetics, protein sequencing, and macro-molecular structure determination as was mentioned earlier. Artificial neural networks and some advanced statistical methods have shown particular promise in these applications. In chemical engineering, advanced models have been used to describe the interaction among various chemical processes, and new tools too have been developed to obtain a visualization of these structures and processes. Let us have a brief look at a few important cases of data-mine applications in engineering problems. Pavilion Technologies' Process Insights, an application-development tool that combines neural networks, fuzzy logic, and statistical methods has been successfully used by Eastman Kodak and other companies to develop chemical manufacturing and control applications to reduce waste, improve product quality, and increase plant throughput. Historical process data is used to build a predictive model of plant behavior and this model is then used to change the control set points in the plant for optimization.

DataEngine is another data-mining tool that has been used in a wide range of engineering applications, especially in the process industry. The basic components of the tool are neural networks, fuzzy logic, and advanced graphical user interfaces. The tool has been applied to process analysis in the chemical, steel, and rubber industries, resulting in a saving of input materials and improvements in quality and productivity. Successful data-mining applications in some industrial complexes and engineering environments follow.
Boeing

To improve its manufacturing process, Boeing has successfully applied machine-learning algorithms to the discovery of informative and useful rules from its plant data. In particular, it has been found that it is more beneficial to seek concise predictive rules that cover small subsets of the data, rather than generate general decision trees [9, 10]. A variety of rules were extracted to predict such events as when a manufactured part is likely to fail inspection, or when a delay will occur at a particular machine. These rules have been found to facilitate the identification of relatively rare but potentially important anomalies.

WebWatcher

Despite the best effort of Web designers, we all have had the experience of not being able to find a certain Web page we want. A bad design for a commercial Web site obviously means the loss of customers. One challenge for the data-mining community has been the creation of "adaptive Web sites"; Web sites that automatically improve their organization and presentation by learning from user-access patterns. One early attempt is WebWatcher, an operational tour guide for the WWW. It learns to predict which links users will follow on a particular page, highlight the links along the way, and learn from experience to improve its advice-giving skills. The prediction is based on many previous access patterns and the current user's stated interests. It has also been reported that Microsoft is to include in its electronic-commerce system a feature called Intelligent Cross Sell that can be used to analyze the activity of shoppers on a Web site and automatically adapt the site to that user's preferences.

1.4.5 Data Mining for Designing Products

In [12, 13, 27], a methodology for using data-mining algorithms in the design of product families was introduced. These were introduced in two steps. In the first step of this methodology, data mining algorithms were used for customer segmentation. Once a set of customers was selected, an analysis of the
requirements for the product design was performed and association rules extracted. The second step created a functional structure that identified the source of the requirements' variability. Options and variants are designed to satisfy the diversified requirements based on a common platform. The last step elaborated on a product structure and distinguished modules to support the product variability. Finally, the paper showed that data-mining techniques could be applied efficiently in the design of product families.

1.5 Stages in Knowledge Discovery in Databases

The application of a data mining algorithm to a data set can be considered the core step of a broader process, often called the knowledge discovery process. In addition to the data mining step itself, this process also includes several other steps. For the sake of simplicity, these additional steps can be roughly categorized into data preprocessing and discovered-knowledge post processing. Data Preprocessing includes the following steps.

i. Data Integration
ii. Data Cleaning
iii. Data Reduction
iv. Data Selection
v. Data Transformation
vi. Data Mining
vii. Evaluation
viii. Interpretation

i. Data Integration:

The KDD process is not currently integrated into normal data processing activities. KDD requests may be treated as special, unusual, or one-time needs. This makes them inefficient, ineffective and not general enough to be used on an
ongoing basis. Integration of data mining function into traditional DBMS systems is certainly a desired goal.

Data integration is the process which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or at files. There are a number of issues to be considered during data integration.

Schema integration

It is the process of matching the structure of the real-world entities. It is rarely possible for any two real-world match with their structure, but during integration the entities should match with their structure then only several relevant external data sources are collected together in the data store where the data mining takes place. If schemas do not match in their structure and still data represented by the schemas is relevant to the integration, then attribute oriented integration should take place. Entity identification problem persists when the entities do not match. Databases and data warehouses typically have metadata - that is, data about the data. Such metadata can be used to help avoid errors in schema integration. Redundancy is another important issue. An attribute may be redundant if it can be “derived” from another table. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

Attribute oriented integration

Attribute oriented integration or attribute level integration, is the process of integrating the data from the selected attributes irrespective of the structure of the schema that they are present. The attribute values of different external data sources are integrated based on their domain values present and their data types, irrespective of the schema structure where they are composed in.
ii. Data Cleaning:

Owing to certain facts data can be noisy, having incorrect attribute values. The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. There may be technological limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes used. Duplicate tuples also require data cleaning.

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion for the mining procedure. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. Therefore, a useful preprocessing step is to run the data through some data cleaning routines.

iii. Data Reduction:

This step consists of transforming a continuous attribute into a categorical (or nominal) attribute, considering on only a few discrete values - e.g., the real-valued attribute Salary can be discretized to take on only three values, say "low", "medium", and "high". This step is particularly required when the data mining algorithm cannot cope with continuous attributes.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.
Strategies for data reduction include the following.

a. **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube.

b. **Dimension reduction**, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

c. **Data compression**, where encoding mechanisms are used to reduce the data set size.

d. **Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which requires storing only the model parameters instead of the actual data), or nonparametric methods such as clustering, sampling, and the use of histograms.

e. **Discretization and concept hierarchy generation**, where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining.

### iv. Data Selection:

The data needed for the mining process may be obtained from many different and heterogeneous data sources. A Selection process, more importantly includes, identifying the data from the external data sources relevant to the data analysis. Selected data is said to be the most relevant to the mining process and analysis, such data which is very relevant to the mining process is called *Task-relevant-data*.

The data selection avoids the incidental mistakes caused in the algorithms while processing. The algorithms do not report not relevant result if the data selected for the input is not relevant to the task described in the algorithm.
v. Data Transformation

Transformation eases the algorithm complexity and simplifies the interpretability of the data by the algorithm. This improves the performance of the algorithm. Data Transformation includes various methods of normalization, aggregation and additional data preprocessing procedures that would contribute towards the success of the mining process.

Data from different sources must be converted into a common format for processing. Some data may be encoded or transformed into more usable formats. Data reduction may be used to reduce the number of possible data values being considered. Data is consolidated into forms appropriate for mining by performed summary or aggregation operations.

♦ **Smoothing**: To remove noise from data. Using binning, clustering and regression smoothing is done.

♦ **Normalization**: Where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.

♦ **Aggregation**: Where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.

♦ **Generalization**: Generalization of the data, where low level or 'primitive' (raw) data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher level concepts, like city or county. Similarly, values for numeric attributes, like age, may be mapped to higher level concepts, like young, middle-aged, and senior.
vi. Data Mining:

Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database.

vii. Evaluation:

Identify the truly interesting patterns that express the interesting knowledge and endorses the productivity in supporting the decision making process.

A data mining system has the potential to generate thousands or even millions of patterns, or rules. Are all of the patterns interesting? Typically not, only a small fraction of the patterns potentially generated would actually be of interest to any given user. Evaluating the result of the data mining process is to validate the result by means of both structure and interestingness measures of the pattern. Several objective measures of pattern interestingness exist. These are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form \( X \Rightarrow Y \) is rule support, representing the percentage of data samples that the given rule satisfies. Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association.

viii. Interpretation:

Data mining results are presented to the users, are extremely important because the usefulness of the results is dependent on it. Right way of presenting the knowledge mined from the mining process will benefit the analyst to interpret the meaning and can be purposefully used for decision making process. Knowledge extracted from the mining processes is not useful for the analysis and
decision making processes, only the knowledge that is rightly interpreted will have value and can be used in decision making processes.

Data visualization refers to the visual presentation of data. The old maxim "a picture is worth more than thousand words" certainly is true when examining the structure of data. The use of visualization technique allows users to summarize, extract and grasp more complex results than more mathematical or text type descriptions of the results.

1.6 Technological Elements of Data Mining

Technological elements of Data Mining are the building blocks for a data mining system. The core group of Technological elements of Data Mining refers to classification, clustering, association analysis and prediction. The technological elements contain methodologies which do not necessarily require formulation in terms of a probabilistic model. In fact, many of these methodologies were invented and developed in the field of computer science rather than in statistics. Recently, however, statisticians have also made use of these methodologies because of their proven usefulness in solving data mining problems. Various technological elements of Data Mining derived from statistics include; measures of inferences, measure of distance, cluster analysis, hierarchical methods, regression, tree models, entropy analysis and pruning. Various technological elements of Data Mining derived from computer science include; genetic algorithms, neural networks, machine learning, soft computing, artificial intelligence, database technologies and so on.

Data Mining is also performed on some undirected subjects who are mostly found in the medical field. At the outset, there are several data mining techniques, such as Market Basket Analysis, Memory-Based Reasoning (MBR), Cluster Detection, Link Analysis, Decision Trees, Artificial Neural Networks (ANNs), Genetic Algorithms, and On-Line Analytic Processing (OLAP) [12, 74].

35
Genetic Algorithms

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination).

Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype of the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

Neural Networks and Artificial Intelligence

In general, a biological neural network is composed of a group or groups of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic microcircuits and
other connections are possible. Apart from the electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion, which have an effect on electrical signaling. As such, neural networks are extremely complex.

Artificial intelligence and cognitive modeling try to simulate some properties of neural networks. While similar in their techniques, the former has the aim of solving particular tasks, while the latter aims to build mathematical models of biological neural systems.

In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots. Most of the currently employed artificial neural networks for artificial intelligence are based on statistical estimation, optimization and control theory.

The cognitive modeling field involves the physical or mathematical modeling of the behaviour of neural systems; ranging from the individual neural level (e.g. modeling the spike response curves of neurons to a stimulus), through the neural cluster level (e.g. modeling the release and effects of dopamine in the basal ganglia) to the complete organism (e.g. behavioural modeling of the organism's response to stimuli).

**Artificial Neural Networks**

An artificial neural network (ANN), also called a simulated neural network (SNN) or commonly just neural network (NN) is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network.
In more practical terms neural networks are non-linear statistical data modeling or decision making tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

Machine Learning

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.

Machine learning algorithms are organized into a taxonomy, based on the desired outcome of the algorithm. Common algorithm types include:

- Supervised learning — in which the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate) the behavior of a function which maps a vector \( X_1, X_2, \ldots, X_N \) into one of several classes by looking at several input-output examples of the function.
- Unsupervised learning — an agent which models a set of inputs: labelled examples are not available.
- Semi-supervised learning — which combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- Reinforcement learning — in which the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- Transduction — similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and test inputs which are available while training.
- Learning to learn — in which the algorithm learns its own inductive bias based on previous experience.

**Soft Computing**

Soft computing refers to a collection of computational techniques in computer science, machine learning and some engineering disciplines, which study, model, and analyze very complex phenomena: those for which more conventional methods have not yielded low cost, analytic, and complete solutions. Soft Computing uses soft techniques contrasting it with classical artificial intelligence hard computing techniques. Hard computing is bound by a Computer Science concept called NP-Complete, which means, in layman's terms, that there is a direct connection between the size of a problem and the amount of resources needed to solve the problem (there are problems so large that it would take the lifetime of the Universe to solve them, even at super computing speeds). Soft computing aims to surmount NP-complete problems by using inexact methods to give useful but inexact answers to intractable problems.

Soft Computing became a formal Computer Science area of study in the early 1990's. Earlier computational approaches could model and precisely analyze only relatively simple systems. More complex systems arising in biology, medicine, the humanities, management sciences, and similar fields often remained intractable to conventional mathematical and analytical methods. That said, it should be pointed out that simplicity and complexity of systems are relative, and many conventional mathematical models have been both challenging and very productive.
Components of soft computing include:

- Neural networks (NN)
- Fuzzy systems (FS)
- Evolutionary computation (EC), including:
  - Evolutionary algorithms
  - Harmony search
- Swarm intelligence
- Ideas about probability including:
  - Bayesian network
- Chaos theory

Generally speaking, soft computing techniques resemble biological processes more closely than traditional techniques, which are largely based on formal logical systems, such as sentential logic and predicate logic, or rely heavily on computer-aided numerical analysis (as in finite element analysis). Soft computing techniques are intended to complement each other.

Unlike hard computing schemes, which strive for exactness and full truth, soft computing techniques exploit the given tolerance of imprecision, partial truth, and uncertainty for a particular problem. Another common contrast comes from the observation that inductive reasoning plays a larger role in soft computing than in hard computing.

**Rough Set Theory**

Rough set theory was introduced by Z. Pawlak in the early 80's and has currently reached a level of high visibility and maturity. Originally, Roughs Sets, whose main philosophy is based simply on discernibility between objects, were presented as an approach to concept approximation under uncertainty [8, 94]. This brilliantly simple idea has been successively expanded in the last twenty years. Many effective methods for data analysis have been developed on the basis of rough set theory. In recent years, a growth of interest in rough set theory and its
applications can be seen in the number of research papers submitted to international workshops, conferences, journals and edited books, including two main biannual conferences on rough sets and the special sub-line of LNCS series. A large number of efficient applications of rough sets in Knowledge Discovery from various types of databases have been developed. Rough sets are applied in many domains, such as, e.g., instance, medicine, finance, marketing, telecommunication, conflict resolution, text mining, intelligent agents, image analysis, pattern recognition, and bioinformatics.

**Decision Trees**

Decision trees are a way of representing a series of rules that lead to a class or value. Therefore, they are used for directed data mining, particularly classification. One of the important advantages of decision trees is that the model is quite explainable since it takes the form of explicit rules. This allows the evaluation of results and the identification of key attributes in the process. The rules, which can be expressed easily as logic statements, in a language such as SQL, can be applied directly to new records.

**Cluster Detection**

Cluster detection consists of building models that find data records similar to each other. This is inherently undirected data mining, since the goal is to find previously unknown similarities in the data. Clustering data may be considered a very good way to start any analysis on the data. Self-similar clusters can provide the starting point for knowing what is in the data and for figuring out how to best make use of it.

1.7 Data Processing and Preprocessing

"Data Processing is a post processing stage for a preprocessing task". Post Processing and Preprocessing are definitely needed for huge data analyses
assignments such as data mining, knowledge discovery. Since the knowledge extracted from the databases through expensive mining processes is valuable in its structure and nature (semantically), the process of discovery should incorporate stages which meticulously churn the data. The knowledge discovery processes often consume longer time to make data ripen during the process and produce as valuable knowledge. Before allowing the data to enter into the valuable knowledge discovery process, the data is ensured of its consistency. As the maxim spells “Garbage In is Garbage Out”. The data given to the process need to be taken care ensuring its consistency so as to enable the knowledge discovery process to extract more valuable knowledge.

Data Processing is a typical data mining activity or knowledge discovery activity, often that is also a set of activities that are involved in the intermediate stages of the data mining and knowledge discovery process. In general terms, data processing includes all the logic and modules that are required to process the input data to produce the required output. The knowledge discovery process itself is a process consisting of several steps may be called as data processing.

1.8 Data Cleansing

Data preparation is a fundamental stage of data analysis. While a lot of low-quality information is available in various data sources, many organizations or companies are interested in how to transform the data into cleaned forms which can be used for high-profit purposes [99]. As Real-world data tend to be incomplete, noisy, and inconsistent, data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

*Missing values*

In a typical data analysis task, a database may consist of tuples with no recorded value for several attributes, such as customer income. How can you go
about filling in the missing values for this attribute? Let's look at the following methods.

Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.

Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown", or $-\infty$. If missing values are replaced by, say, "Unknown", then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common that of "Unknown". Hence, although this method is simple, it is not recommended.

Use the attribute mean to fill in the missing value: For example, suppose that the average income of a customers table in the database contains $28,000. Use this value to replace the missing value for income.

Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for use the most probable value to fill in the missing value: This may be determined with inference-based tools using a Bayesian formalism or decision tree induction.
1.9 Data Cleaning is very essential.

Fig: 1.2 Data Cleaning

Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

1.10 Central Map of the Work

This section locates the area where exactly the work which has been carried over in this thesis is mapped with respect to data mining. Data mining is a key processing activity in Knowledge Discovery Process. The Knowledge Discovery Process consists of various stages. As explained by Han & Kamber, Dunhalm, and others, The KDD Process consists of stages: Cleaning and Integration, Selection and Transformation, Data Mining, Evaluation and Presentation. As noted in the previous section data cleaning is a very important state for knowledge discovery process, Data cleansing activities have been tremendously developed by statisticians and programmers. Cleansing process particularly includes “filling missing values”. Many methods from statistics and
probability have been introduced to fill the missing value with a plausible value. The methods from statistics and probability derive the probable values as plausible values to replace, however the plausibility may not be correct. The hypothetical frameworks and algorithms as carried out in this research uses corpus of digest values of tuples that are rich with clues to replace the plausible values. The figure shown below reflects the actual map of the research work with respect to the knowledge discovery process.

**The KDD Process**

![Diagram of KDD Process]

**Fig: 1.3 Central Map of the Work**

**Summary**

In this chapter preliminary knowledge of data mining and the decision making processes is discussed. The motto of the research is elaborated and the initiative directions have been proposed. Data mining as a radical element contributes to decision support analysis. Missing value analysis had been identified as a pivotal topic of data preprocessing and the core interest of the research. Various statistical methods are looked upon as important basis for elimination of missing values. Imputation methods are found to be the best statistics based methods in the process of missing values elimination. Methods of imputation have been introduced in this chapter.