CHAPTER 2

Review of Data Mining and Human Aspect of Software Engineering

2.1 Introduction

2.2 Human aspect of Software Engineering - What, When, How

  2.2.1 Introduction to HASE
  2.2.2 Evolution of Human aspect of Software Engineering
  2.2.3 People- Capability Maturity Model
  2.2.4 Project personnel - Important Entity of HASE

2.3 Overview of Data mining

  2.3.1 Exploring Data
    2.3.1.1 Knowledge Discovery in Database

2.4 Related work in Data Mining

2.5 Data Mining Tools – WEKA

  2.5.1 Introduction
  2.5.2 Basic Functionality of Weka
  2.5.3 Weka’s main features
  2.5.4 Graphical User Interfaces
  2.5.5 The Weka Library
  2.5.6 Weka Portability

2.6 Summary
2.1 Introduction

In last few years there has been a revolution in the field of software engineering. Practitioners have worked extensively to find the impact of various major elements of software development on quality of software. Practitioners working in software companies have encountered projects which met the goals, schedules and budgets. However some large programming projects did not do so. Therefore software project leaders started looking into factors that make a project successful or the risk factors involved in the software development. It was observed by them that a team of two or three good programmers could achieve success whereas a big team of several programmers could not. While they were looking into framing the process and methodology, it was observed that it is the people who were the main element involved in all processes. Therefore they observed right and capable people should be considered as a important factor for software success. Boehm et al. (2014) considered the quality of project personnel as the most important influence on products when constructing the spiral model for software process. Boehm a software practitioners who also introduced the spiral model, which is risk based process for software development, stated that a good project personnel who performs well is the biggest asset to the company. Boehm has stated that performance of project personnel is good, then it becomes an asset to the company and if the performance of project personnel is bad, then it becomes a big risk for the company. Also while constructing the COCOMO model i.e. the Construction Cost Model , a great weightage was given to good programmers . COCOMO model dealt with main parameters influencing the software project like process documentation, tools, technology and people. Further, authors in Hughes, B. (2006), have strongly pointed out that software professional performance is a strong factor affecting software success. The Constructive Cost Model (COCOMO) is a estimation model for software being developed which has a formula with various parameters and the cost that should be associated with each parameter. This formula is derived taking in account a similar past project. In COCOMO model, Boehm has given maximum importance to capable project personnel. Even though many practitioners have laid emphasis on good programmers, software companies either neglect this important factor or they find it difficult to get the right talent pool. Project Leaders have stated that there is a link between project success and the human aspect and therefore this link of
software engineering which is actually the human component, needs a deeper investigation (Brooks 1987).

Most companies follow the same pattern for selection and retention. It has been observed that companies follow the criteria of stressing on academic scores of the project personnel or experience mainly. It may be so since the earlier projects required such and the pattern were continued without periodic evaluation. Companies have been following the same criteria for a long time. Many of such companies saw a downfall in their products. Their web portal became unpopular and due to it the business got affected. It appeared that though they had all process documented, they did not concentrate much on human factor. Establishments regarding selection and retention of project personnel should be dynamic and based on real data and not on assumptions, old conclusions or beliefs which are formed by glancing at data. The parameters for selection of project personnel should be processed periodically to get the accurate criteria. Data Mining has emerged as a research direction over the past decade.

This chapter explores human aspect of software engineering, data mining techniques and related work in both domains. The complete review of literature starting from beginning of awareness of human component in software development, emerging of data mining as a strong technique for data analysis and the work of researchers in applying data mining in software development and other fields of science and management is described in this chapter.


Human aspect of Software engineering has become a major concern for software companies in recent years. Further introduction, evolution and practitioners perception of HASE is elaborated.

2.2.1 Introduction to HASE

Software Engineering focuses upon development of high quality product through well defined process (Jalote, P. 2002). The development of software is based on an iterative process through which the code structure and functionality are improved gradually. During software development
starting from requirement analysis and planning till testing, project managers are looking into every aspect for producing high quality product. Developing high quality software within scheduled time, cost and resources is one of the major concerns of any software industry. To achieve the same all the processes are well defined and documented. Hardware and the tools are also given priority. All processes are meticulously and diligently followed. However, it is the programmer who transforms documents into knowledge. It is the people working on the project that make things possible. Therefore Quality can be conceived by two dimensions namely through process quality and through people quality (Sangita G. et al. 2013b). Since, people drive the process, the quality of software development process is controlled by the quality level of the people Gerald M. Weinberg, (1997). Consequently, it is vital to carry the software development activities with team consisting of good skill set.

![Diagram of Software Project Components](image)

**Figure 2.1: Software Project Components**

Figure 2.1 indicates that project has two important components which should be given equal importance throughout the development. Project can be completed successfully through cumulative efforts of people and process implementation with equal weight age Stephen H. Kan: (2003). Software companies look into developing the process meticulously though capable people are equally important for the project. In order to complete the project successfully, organizations must focus on two interrelated components—people and process. (Dwyer, Rocky J.,2009).
This study aims to draw the attention of the software community, including management, developing teams, stakeholders, and outsourcing agents, to an important aspect of bringing in a cultural change for the industry to notice and comprehend the connotation of software development as process integration introduced recently by suggesting the need to measure the quality level of process and also measure the competency of the people through few factors specified for good performance during all generic activities. (ASQ, 2011).

In the People, Process and software project system as shown in Fig 2.1, the angle contributing to the Process have been well-understood, documented, and researched. However, the human factor or People angle which makes a significant contribution to any project needs deeper investigation. Even today, the majority of projects are hand-crafted rather than technology-oriented, and the unpredictability of human behavior makes it difficult to bring sophistication in the trade of project engineering (Linberg KR, 1999). Even the portions which are technology-oriented need project personnel with certain skill sets for giving quality output. The Human component, be it the motivation or energy, the emotion, the creativity, and the negligence or error, largely decides success or failure of software projects or engineered products (Humphrey W S., 2006). Now, when the software world is expanding and growing exponentially and more research is being done in the field of software project management, the more it is accepted by the software engineering community that along with process and technology, the people involved in software development processes deserve more attention.

Thus, software quality can be broadly defined as

\[
\text{Software Quality} = \sum (\text{Process Quality} + \text{People Quality}) \quad \text{equ 2.1}
\]

Since project is perceived as having two important components that is people and process, it is necessary to maintain quality in both dimensions for quality software product. By just keeping best process and tools and not good programmers to implement them, the project will certainly not be of the desired quality. Quality can be in terms of innovative, meeting the customer requirements, and completing within cost and time constraints. (Stephen H. Kan, 2003). To
achieve the desired quality companies need to give due importance to people component of the project. Software quality can be achieved by keeping capable people along with maintaining the process quality summation over all complete development process as indicated in equ 2.1.

2.2.2 Evolution of Human aspect of Software Engineering

The beginning of the awareness of the human aspects of software engineering appeared in Brooks’s book titled, “The Mythical Man Month” ,M-MM( Brook,2004). In the preface to the 20th Anniversary Edition, Brooks writes that he is surprised that The Mythical Man-Month is popular even after 40 years. Few books on software project management have been as influential and timeless as The Mythical Man-Month. With a blend of software engineering facts and thought-provoking opinions, M-MM offers insight for anyone managing complex projects. These essays draw from his experience as project manager for the IBM System/360 computer family and then for OS/360, its massive software system. Now, 20 years after the initial publication of his book, Brooks has revisited his original ideas and added new thoughts and advice, both for readers already familiar with his work and for readers discovering it for the first time. Programming managers have long realized wide productivity variation between good software programmers and low ones.

Table 2.1: Positive and Negative impact of major elements during software development.
Bob Hughes (2002) has stated in Software Project Management various factors which affect the software development. As shown in table 2.1 human component affects the development most. If the team is capable it affects the development positively by 100% and a poor programmer will affect it negatively by the same magnitude. Human factor has been recognized as the biggest risk in a software company. Also a skilled team is the biggest asset to the company. The other important factors were process, tools and technology i.e. hardware(SEER for Software). Fig 2.1 shows the positive and negative impact of each element during software development.

As has been mentioned above, the importance of the human aspects of software engineering becomes acknowledged recently. DeMarco et al. (1999) investigated the projects which failed. Failed projects were which failed to complete. They interviewed those team members who were still working in the same company to find out the reasons. For the majority of the failed projects they studied, there was not a single technological issue to explain the failure. Many failures of software systems can be explained by human factors. Brooks, a project manager in IBM has stated in this book M-MM that a small team of capable good programmers rather than a big force of inefficient members. The conclusion was made while working on OS/360, Exec 8, Scope 6600, Multics, SAGE, TSS etc. (IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990). Brooks(2004) concluded that it is better to keep a team of just 25 capable project members rather than a team of 125 inefficient members.

The process to successfully developing quality software suffers from challenges identified over 40 years ago. Upon deeper review, a majority of the factors related to software development failure are human factors. This includes having a good project leader and who can recruit and maintain a good team. This have a positive impact in addressing software engineering challenges as stated by DeMarco, et al. (1999). A project manager focuses not only on what to do with the project but also how. It is essential for project leader to conduct cyclical assessment, informal learning, and dynamics coaching to ensure team harmony and growth and practice sound project management ethics for quality software (Kerzner, 2010).

Unless people are considered as an equally important leg supporting the product–process–people triad of software engineering, the results will remain inconsistent and unstable at best (Townsend, 2007).
However, coming to capability, the complexity comes to finding those factors which make a project personnel capable. On analyzing the human component, many factors like social, economic, skills, certifications and technical factors contributed to capability factor of the project personnel.

Campell (1990) has specified few parameters for empirically analyzing the capability of a employee and listed the following factors that contribute to the performance.

- job-specific task proficiency
- non-job-specific task proficiency
- written and oral communication
- demonstrating effort
- maintaining personal discipline
- facilitating peer and team performance
- supervision/leadership
- management/administration

These factors are broadly listed for describing the performance in most jobs in any sphere (Arnold et al. 2004).

Taking into consideration the complexity of the topic, this study concentrates only on technical aspect of project personnel working on web based projects which are more of service than product and this domain is growing rapidly.

2.2.3. People- Capability Maturity Model

The central question is how to improve the human component and thereby the software development process. Software is a human intrinsic product and few practitioners have come up
with models for dealing with human aspects of software engineering. To date, improvement programs for software organizations have often emphasized process or technology, not people. Some practitioners did come up with models proposing the importance of human component in software development. The companies who followed it did have an advantage.

Many practitioners like Humphrey W.S. (1995) have stated about the crucial role of people aspect of software. People aspect may be individual capability, team cohesion and a good project leader. Coming to individual capability, Boehm has stated few personnel attributes like analytical ability, application experience and programming language experience in COCOMO model which affect the software development process considerably.

Curtis B. et al. (August 1990) have developed a People Capability Maturity Model stating few areas like people capability, job performance, team cohesion and good project coach for software success. The People Capability Maturity Model (P-CMM) discusses on people capability factors which should be benchmarks practices for improvement of software development process by increasing the capability of software engineer. It presents a documented roadmap for organizational improvement by concentrating on human aspects of software organization.

P-CMM i.e. the People Capability Maturity Model mainly focused on people component. The P-CMM is a maturity framework that focuses on continuously improving the management and development of the human assets of a software or information systems organization. It provides guidance on how to continuously improve the ability of software organizations to recruit, develop, motivate, organize, and retain the talent needed to steadily improve their software development capability. The main objectives listed in P-CMM are as follows

- improve the capability of software organizations by increasing the capability of their workforce

- ensure that software development capability is an attribute of the organization

- align the motivation of individuals with that of the organization

- retain human assets i.e. skilled and knowledgeable people within the organization
These practices have been selected from experience as those that have significant impact on individual, team, unit, and organizational performance. The P-CMM includes practices in such areas as work environment, communication, staffing, managing performance, training, compensation, competency development, career development and team building.

With the help of the People-Capability Maturity Model for Software (P-CMM), many software organizations like The Lockheed Martin Corporation, Computer Sciences Corporation, Intel Corporation, Novo Nordisk A/S, Tata Consultancy Services, Infosys, and Wipro technologies, the U.S. Army, Federal Emergency Management Agency and Boeing Company have made effective improvements in their software development (Curtis et al. 2003).

Through the P-CMM many of these organizations have discovered that significant changes are required in the way they manage, develop, and use their people involved in software development.

With the help of People-Capability Maturity Model for software many software companies have been able to make failure free software effectively. P-CMM emphasize upon people component, but there is need to find patterns and use good techniques like data mining for analyzing the people component.

Accordingly, software organizations must become centers of excellence that take talented individuals from universities and other sources and develop them into motivated and productive software engineering teams. Increasing the knowledge, skills, and performance of software developers should be the top priority of the company.

There is therefore need to investigate deeply and focus on attributes related to personnel capability which affect performance at work place. The performance thereby differentiates between good, average and poor programmers based on their capability factors. However human aspect is complicated since it has technical, emotional, psychological, social and economic factors associated with it.
2.2.4 Project personnel - Important Entity of Human Aspect of Software Engineering

It has been realized that human aspect is an equally important dimension in the software triad of project, people and process. It is utmost important to concentrate more on the human aspect of software engineering for software success. Researchers are looking into various dimensions and also various methods for inspecting the same.

This study focuses on the performance of project personnel based on his capability attributes. It is time to bring a change in yardsticks which most software companies are following based on certain assumptions or glancing data or following a conventional method for long time for selecting the project personnel. Conventional and static personnel selection approach is no longer contributing to high quality products. Companies need to bring a change in their managerial policies related to selection and retention of project team members.

Mining software engineering data has emerged as a research direction over the past decade (Rama Krishna et al. 2010). This research direction achieved substantial success in practice. Software repositories contain a wealth of valuable information about software projects. Using the information stored in these repositories, software companies depend more on historical and field data rather than static assumptions and intuitions. On processing this data, this data set becomes the input for a data mining method or a machine learning system. These methods are capable of learning from historical data set and predict the output of new data set. Classification is to find the class of an entity based on its attributes or characteristics. Therefore classification is both descriptive and predictive. It describes the present pattern and the same is the basis for prediction the future trends.

Technical attributes like programming skills, reasoning skills, domain knowledge skills, academic scores whether experience and college tier have a impact, need an insight which can be achieved by data mining approach. This study aims to develop a framework that uses data mining methods mainly classification to distinguish and extract those attributes of project personnel which contribute to high performance in the project and enhance the quality of software.
2.3 Overview of data mining

Data Mining is a methodology which is also called knowledge discovery and to extract patterns and association in the data. This involves using of combination of statistical methods, numerical analysis and machine learning techniques.(King et al. 1998).

Data mining has attracted a great deal of attention in the information industry in recent years due to databases growing immensely and the imminent need for turning this data into useful information or knowledge. It has wide applications from ERP systems, market analysis, fraud analysis, production industries or medical sciences too. With the tremendous progress in hardware technology which can store large volumes of data, simultaneously a software domain for dealing with this data came into existence in the last three decades. Many methodologies came into existence for transaction management, information retrieval and data analysis. Data mining is actually knowledge mining from the data. Alternatively, it is also called knowledge discovery from data or KDD.

Date mining methodologies can analyze large or small data set and discover hidden patterns and relationships. Data mining has a vast set of techniques incorporating statistics, numerical analysis, neural networks, genetic algorithm and decision trees.(Oded et al. 2005). The taxonomy of data mining techniques is shown in figure 2.2.
Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given dataset. Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transaction data like collections of items bought by customers, or details of a website frequentation. It gets frequent item sets and can also generate strong association rules from those sets. It works on the principle of min support and minimum confidence threshold. FP-Growth or Frequent Pattern Growth. This technique improves Apriority to a large extent since it first builds a compact data structure called FP-Tree and then extracts frequent item sets from the FP-Tree. (Silberschatz et al. 1998).
Clustering is an unsupervised learning that finds natural grouping of instances given unlabeled data. There are many instances when the data does not have pre-defined classes. In this situation, it is important to whether the data consists of distinct classes. Clustering divides a dataset into several clusters in which the intra-class similarity is maximized while the inter-class similarity is minimized. Clustering is a data mining technique of grouping features, abstract or physical objects into classes of similar objects. A cluster is a subset of objects that are similar. The distance between any two objects within the cluster is lesser than the distance between any object within the cluster and any object not situated inside it. (Gose et al. 1996).

SLINK or Single-Linkage clustering is one of several methods of hierarchical clustering. It is based on grouping clusters in bottom-up fashion or also called agglomerative clustering, at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other. BIRCH (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm used to perform hierarchical clustering over particularly large data-sets. An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points in an attempt to produce the best quality clustering for a given set of resources (memory and time constraints). In most cases, BIRCH only requires a single scan of the database. CURE (Clustering Using REpresentatives) is an efficient data clustering algorithm for large databases that is more robust to outliers and identifies clusters having non-spherical shapes and size variances. k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (Han J. et al. 2006).

CLARA (Clustering for Large Applications) is also a good way of clustering larger datasets is to try and extend existing methods so that they can cope with a larger number of objects. The focus is on clustering large numbers of objects rather than a small number of objects in high dimensions. Kaufman and Rousseeuw suggested the CLARA (Clustering for Large Applications) algorithm for tackling large applications. CLARA extends their k-means approach.
for a large number of objects. It works by clustering a sample from the dataset and then assigns all objects in the dataset to these clusters. (Gose et al. 1996).

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter, Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

Classification derives a function or model that identifies the categorical class of an entity based on its characteristics, features or attributes (Thair N. Phyu 2009). Classification is a supervised learning since the inputs and the possible outcomes are known. For example, in this study categories were made about performance based on input set of characteristic about the project personnel. A decision tree generated by classification technique can be converted to set rules. Some of the most well-known decision tree algorithms are ID3 (Iterative Dichotomise 3), C4.5, CHAID (Chi-square Automatic Interaction Detection) and CART (Construction and Regression Tree).

ID3 was developed by Quinlan Ross (1986) which was based on concepts of learning systems described by Hunt. ID3 adopts a greedy approach in which decision trees are constructed recursively by divide and conquer method. Attribute selection measure is information gain measure for splitting criteria. ID3 is limited to categorical values in building a tree model. ID3 gives a full tree without pruning since it preprocesses noise.

C4.5 builds decision trees from a set of training data in the same way as ID3 and is also based on Hunt's Algorithm(Hunt et al. 2001). However this is a further extension of ID3 and works for continuous data too. In this study, sample set consists of a d-dimensional vector (x1,x2,......xd), where d is the number of attributes and xi represents attributes. The strategy is very straightforward. Heuristic procedure is used for attribute selection. The attribute that best discriminates the tuple to class according to selection measure i.e. information gain is taken to
make decision for splitting the tree into split point or split tree. The root node will be the attribute whose gain ratio is maximum. The C4.5 algorithm further works recursively on sub lists. C4.5 uses pre pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

CART is Classification and Regression Trees which was developed by Breiman. It is also based on Hunt’s algorithm. CART has the ability to deal with both categorical and continuous attributes to build a decision tree. It is able to handle missing values too. CART uses Gini Index as an attribute selection measure to build a decision tree. CART produces binary splits unlike ID3 and C4.5. Therefore, it produces trees which are binary in nature. CART uses cost complexity i.e. Post pruning to produce simplified tree. Generally, the smallest decision tree that minimizes the cost complexity is preferred.

CHAID or Chi-square Automatic Interaction Detection is a Classification Tree technique that evaluates complex interactions among predictors and displays the modeling results in an understandable and easy to comprehend tree structure. The trunk of the tree represents the total modeling database. CHAID then creates a first layer of branches by displaying values of the strongest predictor of the dependent variable. CHAID automatically determines how to group the values of this predictor into a manageable number of categories.

Artificial Neural Networks are similar to biological neural networks in the performing of functions collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which individual units are assigned. Neural Networks possess two important characteristic i.e. it contains sets of adaptive weights or numerical parameters that are tuned by a learning algorithm and the capability of approximating non-linear functions of their input. Kohonen is a very popular neural network. The self-organizing map (SOM) invented by Kohonen performs a form of unsupervised learning. A set of artificial neurons learn to map points in an input space to coordinates in an output space. The input space can have different dimensions and topology from the output space, and the SOM will attempt to preserve these.(Earl et al. 1996).
Regression is an important data mining (machine learning) method that can be used to fitting an equation to a dataset. The easiest form of regression is linear regression, uses the equation of the straight line, \( y = mx + b \) and establishes the suitable values for \( m \) and \( b \) in order to predict the value of \( y \) based on a given value of \( x \). The multiple regression technique allows more than one input variable and allows for the appropriate of more complex models. Regression analysis is extensively used for predicting and forecasting. The appearance of regression to visualize is linear regression through a single predictor. A linear regression method can be used if the association between \( x \) and \( y \) can be approximated through a single predictor. A nonlinear regression refers by means of two or more predictors like \( x_1, x_2, \ldots, x_n \). While multiple predictors are used, the nonlinear association cannot be visualized in two dimensional space. Regression modeling has many applications in financial forecasting, trend analysis, biomedical and drug response modeling, business planning, environmental modeling, time series prediction, and marketing (Fu 2004).

2.3.1 Exploring Data Mining

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or pattern analysis. The process of knowledge discovery is explained below.

2.3.1.1 Knowledge Discovery in Database

The current information age has high end hardware technology to store huge amounts of data. More and more information is stored in databases and turning these data into knowledge creates a demand for new, powerful tools. Data analysis techniques used before were primarily oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. These interpretations and insights are the required knowledge. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts. The current situation however needed a new way to deal with these never ending databases and new methods to analyze this huge amount of data. A new area came into being known as Knowledge Discovery in Databases, also known as KDD. The process of KDD is depicted as Figure 2.3 and
consists of an iterative sequence of the following steps as stated by Fayyad et al., (1996). The data mining process has been depicted in Fig 2.3 and the iterative steps are listed below.

Basic steps are data cleaning, data integration, data selection, data transformation, data mining, Pattern evaluation and finally knowledge presentation (Daniel et al., 2005). All these steps are elaborated below.

a. Data cleaning

Real-world data is incomplete and inconsistent or simply raw. It may contain redundant and duplicate information. It may not be normalized i.e. scaled accordingly. It may contain data which needs to be sorted and arranged for processing. This is due to the fact data comes from multiple sources and each source may have a different format or structure. Noisy data gives unclear results and causes ambiguity while applying mining methods. It is therefore important to pre process the data and also reduces the data without jeopardizing the data mining results. Most mining tools have procedures for pre processing data effectively. The Weka toolkit used in this study too has a data pre processing section, in which data can be pre processed before putting it into experiments. Other popular tools are R, SPSS which too can preprocess the data before applying data mining techniques.

b. Data integration

The task of data integration combines data from various sources in a single data store.

c. Data selection

The task of data selection includes sampling and reducing data such that we have relevant data for analysis.
d. Data transformation

Data are transformed or consolidated into forms appropriate for mining by performing the following:

- **Smoothing**: It works to remove the noise from data. Such techniques include binning, clustering, and regression.

- **Aggregation and generalization**: It means applying aggregation operations to the data. In this study few attribute values have been aggregated. For example the skills had a wide range of values. Those values were aggregated and generalized into good, average and poor.

- **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.
Data mining is the process of analyzing data from using computational methods which are basically integration of various mathematical, statistical and numerical analysis methods. It aims at extracting hidden patterns and results in knowledge discovery. It gives a different perspective of the data and the results obtained can be summarized into useful information. This information can be used to increase revenue and growth rate of the company.

Data mining is a logical process that is used to search through large or small amount of data to find interesting patterns and correlations. Data mining techniques find patterns that were previously unknown.
f. Pattern evaluation

This step is required to identify the accuracy of the methods used and also the results obtained. It also involves finding the consistency of the results obtained by different data samples and thereby the methods used on them. Sometimes a very accurate method is difficult to evaluate and interpret. Evaluation includes understanding the various results, checking whether the discovered knowledge is accurate and understandable by domain experts. It also involves checking the impact of the discovered knowledge.

g. Knowledge presentation

This final step includes the use of visualization and knowledge representation techniques to present the mined knowledge to the user. The discovered knowledge must be organized and presented in a way that the user can use. The results of data mining techniques can further be converted in graphs, trees and rules which can be easily followed by users.

It is clear from figure 2.3 that an integral part of the process of KDD is data mining. Data mining overcame limitations of data analyzing techniques used before. Data mining is an interdisciplinary field of various disciplines like database systems, statistics, machine learning, pattern Recognition and information sciences. Each of the domains taken separately would use techniques in their domain. However data mining techniques is confluence of all these sciences are therefore very powerful and useful technique. The technique used in this study is explained further.

Chapter 3 and 4 will deal with details about data taken into consideration for the research and application of Bayesian model and other classification techniques on the research data.
2.4 Related Work in Data mining

Researchers have used data mining techniques in last few decades to analyze large or small data sets in engineering, science, marketing and management to extract hidden knowledge using data mining wide and diverse applications. For example in marketing, it is used for identifying customer buying behavior. This can further be utilized for customer satisfaction. It is used to banks to see the patterns related to loan payments, client credit payments trends etc. Data mining is also used widely in biological sciences. Humans have around 3000 genes. Data mining has been used to reveal interesting information related to it using classification. Also it has been used to find interesting and important information from surgery databases to predict the duration of different surgeries (Combes et al. 2008).

Data mining has been used in software engineering too. Data mining techniques help to improve the decision making during software development by using data stored related to the software project (Chang et al. 2006). Data mining techniques can be applied in solving Software Engineering problems like detecting bugs, to aid in pattern discovery, and to help developers deal with the complexity of existing software, in order to create more failure-free software. Challenges in software engineering are mainly during requirement gathering, systems integration and evolution, maintainability, pattern discovery, fault detection, reliability, and complexity of software development. Also software engineering data can be broadly categorized into three types: sequences, graphs, and text. All these types of data have patterns which are difficult to extract by glancing. Data mining techniques like classification, association and clustering is able to deal with such data and produce the required patterns and discover knowledge from it.

Ajay Prakash et al. (2012) have used data mining techniques for code reuse in software. Nowadays’ most of the software products are developed by using existing versions or features in order to reduce the delivery time of software product, to improve the productivity and quality and to reduce the development effort. Software reuse has been a solution factor to acquire the existing knowledge from software repository. To extract existing knowledge from software repository authors have used data mining. Their study gives the description of software reuse process and knowledge discovery process. Data mining techniques are used effectively to extract useful knowledge from software repository using different software metrics. Finally, this knowledge was used by project managers for better management of the software projects.
The development of machine learning algorithms in recent years is applied in various domains of software engineering like prediction of bugs etc. (Gabriela Czibula et al. 2014). Authors have applied clustering algorithms for defect prediction in software. They have stated that software defect pattern analysis is among the most challenging domains for structured machine learning over the next few years. Data mining algorithms, especially the Random Forest ensemble methods, are appropriate for adapting to the behavior of the observed data. They have used data mining methods for achieving software defects, primarily because these methods can model the inherent stochasticity of software testing and other real world.

Pushpavathi T.P et al. (2012) have applied data mining to detect bugs and use the extracted knowledge to create failure-free software. Data mining techniques like clustering and association were used for finding patterns of defects.

Researchers like Mohanty et al. (1997) have applied data mining in human talent prediction too in various industries.

Data mining has been used in other industries for discovering knowledge about human capital for enhancing the product quality and quantity. Chien et al. (2008) have done an empirical study for enhancing the manufacturing of semiconductor industry by analyzing human capital using data mining techniques. They have considered experience, college tier, gender, age, job satisfaction for finding the impact on performance. They developed a data mining framework for personnel selection to explore the association rules between personnel characteristics and work behaviors, including work performance and retention. They have analyzed which personnel are fit for which job category in the industry. A highly educated person with a post graduate or doctorate degree will perform differently than a graduate. A diploma holder fits well and performs well in a different job category. They have carried out a detailed study for enhancing the productivity of semiconductor industry by investigating the human capital using data mining techniques. An empirical study for indirect labor (IDL) including engineers with different job functions in one of the world largest semiconductor foundry company located in the Hsinchu Science Park in Taiwan is studied to demonstrate the validity of this approach. In particular, they used decision tree, a very good classification technique to discover latent knowledge and form rules to assist in personnel selection decisions.
Ranjan (2008) too have used data mining technique like classification methods to bring out factors required for human capital enhancement in an institution. Managing an organization talent has become one of the challenges to the HR professionals. This task involves a lot of managerial decisions in order to decide the right person for the right job at the right time. Sometimes, these types of decisions are very uncertain and difficult; and it depends on various factors such as human experience, knowledge, academic performance etc. The process to identify an existing talent in organization is among the top talent management challenges and becomes a never ending issue. They have used decision trees to capture the patterns and convert the patterns into rules for management.

Jantan (2010) have concentrates on identifying the patterns that relate to the human talent in a organization. They have generated patterns by using some of the major data mining techniques, such as the clustering technique which is used to list the employees with similar characteristics, to group the performances and etc. From the association technique, patterns that are discovered can be used to associate the employee’s profile for the most appropriate job, associated with employee’s attitude to performance. The attributes for training dataset are selected based on the related factors for employee performance like work outcome; knowledge and skill; individual quality; and activities and contribution. Their study thereafter generated a forecasting model that contains classification rules for human talent prediction. The classification rules also show us about the interesting or important attributes for the dataset. The forecasting model will be used to determine whether the employee is recommended for promotion or not based on his/her performance. In particular among all available decision trees methods they have used C4.5 in Weka tool kit. The C4.5 classification algorithm is easy to understand as the derived rules have a very straightforward interpretation the study is based on the performance factors as per the performance appraisal standard used by the Malaysian public sector. Their research was conducted as a part of the science fund project funded by MOSTI (Ministry of Science, Technology and Innovation), Malaysia.

Saron et al.(2012) have used data mining for human talent prediction in institutions and took into account attributes like qualification, training and social obligation. Data mining techniques such as ANN, Decision Tree and Rough Set Theory were used by the authors for predictive analysis.
The study aimed to develop an academic talent model using data mining based on several related human resource systems. In the case study, they used 7 human resource systems in one of Government Universities in Malaysia. This study shows how automated human talent data mart is developed to get the most important attributes of academic talent from various tables like demographic data, publications, supervision, conferences, research, and others. Apart from the talent attribute collected, the forecasting talent academician model developed using the classification technique involving 14 classification algorithm in the experiment for example J48, Random Forest, BayesNet, Multilayer perceptron, JRip and others. They have also done comparative study on the methods. The paper contributes on how to develop latest and accurate academic talent management using data mining.

Ngai et al. (2009) have studied on various data mining techniques to improve sales of company by extracting knowledge by looking into various aspects of customers. Data mining methods gives the buying trends of the customer. Data mining techniques have given description of the items which are mostly purchased. Based on the knowledge derived, the company can make changes in their manufacturing and retailing policies. Thereby, the company can improve their sales and turnover too.

Azar et al.(2013) have derived rules using classification for finding reasons for employee for selection and recruitment. The authors have also stated that selection and recruitment should be done in two steps. Recruitment refers to the first half of the hiring process and selection to the next half. Recruitment aims at identifying suitable candidates and selection focuses on choosing the best of them.(Cooper et al. 2002). Selection presents the final step of decision-making in a hiring process. Recruitment and selection should be given utmost importance in the current competitive and global business environment to achieve high quality customer service. These factors when taken into consideration enhance the performance of bank employee and thereby increase the turnover of bank. A decision-making tool is provided for managers to use during the recruitment process. The effective factors in employees’ performance will be identified by discovering covert patterns of the relationship between employees’ test scores and their performance at work. The various data mining technique like Chi-squared Automatic Interaction Detector (CHAID), C4.5 and Classification And Regression Tree (CART) algorithm were used by the authors. The objective and the appropriate algorithm were determined based on few
irrelevant components, which the Commerce Bank Human Resources management’s experts describe. They had taken 26 effective variables and after data pre processing only five variables, such as province of employment, education level, exam score, interview score and work experience, were found to have effect on performance.

Researchers have used data mining techniques in last few decades to analyze large or small data sets in most fields of engineering, marketing and science to extract hidden knowledge.

Shiue et al. (2003) have used decision trees for Production Company. They developed an intelligent scheduling controller to support a shop floor control system to make real-time decisions for various production requirements. Based on various production requirements they formed a set of attributes or features they constructed a knowledge base for intelligent scheduling controller. The methodology was integration of genetic algorithm and decision trees. The rules generated assisted for controlling scheduling in the shop floor in the manufacturing company. This project was funded by National Chiao Tung University, Hsinchu, Taiwan.

Though data mining has been applied in many fields, it has not been used yet in human aspect of software industries. Practitioners of software industry have pointed out about the ability factors of project personnel and also other managerial issues in a software company, but no further literature specifies how to arrive at the right capable factors and select and retain the right talent.

This research is thus focused on a classification problem in which the algorithms for classification are used to predict the performance of a new project personnel based on the rules derived by the historical data of similar projects (Sangita et al. 2015a).

Data mining results in decision through accurate parametric methods and not assumptions. As part of this research an empirical study was conducted for selection criteria for software industry by introducing a knowledge based decision tree algorithm. It was found that though academic performance was given importance by most of Software Company, talents such as programming and reasoning skills contributed largely to performance in software companies. Depending on few selected attributes like reasoning skills, programming skills, experience skills, domain skills related to employee, the model could predict their performance (Sangita et al. 2013).
This research also included using bayesian classification for finding performance related attributes. Some of these attributes which were considered were personal characteristics, educational and professional attributes. As a result for their study, they found that employee performance is highly affected by education degree, skills and experience. Bayesian classification showed that other talent factors contributed largely to work performance rather than academic performance (Sangita et al. 2014). Further chapters of this thesis will deal with more in-depth explanation and application of bayes classification and how the conclusions were derived using this technique.

Further the research work used Weka Tool for finding factors that affect the performance at work place. They studied the impact of work conditions along with position of the employee on their performance. They used data mining methods in Weka environment. Weka tool supports most data mining methods. By applying various data mining techniques it was found that factors such as skills, college tier, working conditions and job satisfaction affected the performance of the employee largely. (Sangita et al. 2015b). After the application of data mining methods on data, the result is evaluated and interpreted for decision making. The hidden knowledge derived from the results is the basis for further decisions by the management. Thus, the intelligent decision support system can be used to predict in future about the performance of employee in the project. This study will provide the management with useful knowledge for enhancing the quality of software by deploying the right team at very start of project. With a good team and training provided to them the organization can improve the performance of the project personnel. The extracted patterns about performance will be the guidelines for important decisions by the management for quality software. This study has shown that companies need to follow a realistic performance assessment method from time to time for project success.
2.5 Data Mining Tools - Weka

Weka stands for Waikato Environment for Knowledge Analysis. The Waikato Environment for Knowledge Analysis (Weka) system was developed at the University of Waikato, New Zealand. Weka is an extensive library of data mining and machine learning algorithms.

Weka is a freely available machine learning software written in Java. It supports all functions of data mining like preprocessing, clustering, classification and so on. The learning methods are applied on dataset using Weka and the output is analyzed to extract information about the data. Weka also gives statistics regarding the performance of the several learners so that comparative study with respect to the performance can be done in order to choose the best classifier. The Weka can be run from the command line or from the interface.

2.5.1 Introduction

Weka toolkit is a widely used open source software tool which contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. Weka contains tools for regression, classification, clustering, association rules and visualization. The classify panel enables the user to apply classification algorithms to the resulting dataset, estimate the accuracy of the resulting predictive model, visualize erroneous predictions and the model itself. The Weka machine learning workbench offers a general-purpose environment for automatic preprocessing of data, classification, clustering, and feature selection for varied data mining problems. Weka not only includes implementations of algorithms for clustering, classification, and association rule mining, it also has a graphical user interfaces for visualization of data and results.

It is free and open source software used in educational institutes for teaching and application of machine learning and data mining topics. It is a very good research tool for developing and empirically comparing data mining techniques. It is applied widely in other academic fields, and in commercial settings. New algorithms can easily be incorporated and compared to existing ones on a collection of data sets.
Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

2.5.2 Basic Functionality of Weka

The functionality of Weka can be accessed through different graphical user interfaces, Experimenter interfaces, mainly the Explorer and Knowledge Flow interface shown in Figure 2.4. The most important interface, the Explorer, supports all the main functions data preprocessing, visualization, classification, clustering, data loading and filtering, and attribute selection methods. The Experimenter is a tool to set up the machine learning experiments that can evaluate the classification and regression methods. It permits the easy comparison of performance, and can tabulate summaries in ways that are easy to comprehend. Experiments are allowed to setting up and to run in parallel over different computers in a network so that multiple repetitions of the cross validation (default method for performance analysis) and can be distributed over multiple sample data.

The Knowledge Flow interface is a Java Beans application that facilitates the similar kind of data exploration, processing and visualization as the Explorer (along with additional features). The user can allow to define a workflow specifying how data is to be loaded, preprocessed, visualized and evaluated, which can be repeated multiple number of times.

This can makes it easy to optimize the workflow by changing the parameters of algorithms, or to apply it to other data sources.
2.5.3 Weka’s main features

• Data pre-processing

The file format of Weka is ARFF, Weka supports a variety of other formats (for example Matlab ASCII files, CSV) and connectivity of database through JDBC. The data can be filtered by a number of methods, ranging from removing specific attributes to advanced process such as principal component analysis.

• Classification

The Classify panel enables applying classification algorithms also termed classifiers to the dataset. It gives very good statistics regarding accuracy of the classifier and also visualization of the model. Weka contains around 100 classifiers. Classifiers are categorized into rule-based methods (decision tables, OneR, RIPPER), function-based learners (linear regression, SVMs, Gaussian processes), Bayesian methods (Naive Bayes, Bayesian nets, etc.), lazy methods (nearest neighbor and variants), tree learners (C4.5, Naive Bayes trees, M5) and miscellaneous methods. In addition, WEKA consists of multiple instance classifiers; meta-classifiers like boosting, bagging, stacking and interface for classifiers performed in Groovy and Jython.
• Clustering

Unsupervised learning is carried out by several clustering schemes, including k-means, EM based mixture models and various hierarchical clustering algorithms. However, not as numerous methods are obtainable as for classification, the majority of the classical algorithms are incorporated.

• Select Attribute

The set of attributes utilized is very important for classification performance. Different search methods and selection criteria are available. To mention a few methods, information gain and gain ratio gives the ranking of attributes. Based on the results, the experiments can be conducted with attributes with higher ranking since lower ranking attributes are insignificant. Weka gives the option of selecting the attributes needed for classification in the pre processing section.

• Data visualization

The data can be demonstrated visually by plotting the attribute values against the class or against the other attribute values. Classifier outcome can be compared to the training data in order to identify outliers and examine the classifier characteristics and decision boundaries. There are specific methods like focused visualization tools, such as tree viewer for any method that generates classification trees. Additionally, WEKA also gives a complete summary of accuracy parameters of algorithms for further evaluation of methods deployed.

2.5.4 Graphical User Interfaces

WEKA’s functionality can be applied through different graphical user interfaces, mainly Experimenter interfaces, the Explorer shown in Figure 2.5 and the Knowledge Flow interface. The Explorer interface allows rapid detection of data and supports all the main items stated previously - data loading and filtering, clustering, classification, selection of attribute and a variety of forms of visualization - in an interactive fashion.
2.5.5 The Weka Library

The programs in library intend to build a state of the art capability for developing techniques of machine learning and exploring the application of the learning algorithms. Particularly it facilitate to create a work bench for machine learning, finding the factors that contribute towards its application in various industries and extend novel techniques of machine learning and ways of evaluating their effectiveness.

Data mining and classification were a main objective to our research and a tool was needed to arrive at accurate results. The tool used for this research is Weka.

![Figure 2.5: Menu bar of experimenter with open file option](www.cs.waikato.ac.nz/ml/weka)

![Figure 2.6: Weka GUI Chooser](www.cs.waikato.ac.nz/ml/weka)
Once the Weka graphical user interface is started, as shown in Figure 2.6, one can use the Explorer button to enter the system and then load the data file. Subsequently, there appear options for preprocessing and classification. Hence, it is systematic software with all facilities for data mining where one can select from a menu of library options any particular classifier.

2.5.6 Weka Portability

The earlier versions of Weka i.e. Weka 1 and Weka 2 were limited to UNIX operating systems and distributions were made available for Linux and Solaris too. However, with the onset of Java version, it is available on all platforms, like Windows too. The lack of dependency on externally-maintained libraries makes maintenance of the code base much easier. Therefore, the present Weka tool Version 3 set has become extremely popular owing to its portability on any platform.
2.6 Summary

IT industries aim for fulfilling the target of completing the project within cost and time constraints along with delivering high quality, good service and innovation in their product for their subsistence in the global market. Due to the swift transformation of technology, software industries owe to manage a large set of data having precious information hidden. In recent years data mining technique enables one to effectively cope with this hidden information where it can be applied to code optimization, fault prediction and other domains which modulates and enhances the success nature of software projects.

It is the project personnel which transforms knowledge and gives shape to software product. Therefore, the quality of the product developed depends largely upon the capability of the project personnel along with process quality. The position of this research work therefore is to explore potentials of project personnel in terms of their competency and skill set and its influence on quality of project. The above mentioned objective is accomplished using a classification methods to capture the pattern of human performance. By this means, the hidden and valuable knowledge discovered in the related databases will be summarized in structured and interpretable form.

Application of data mining in human aspect of software industry has not progressed much. This study aims to develop a data mining framework for analyzing human talent in Software Company. In this way a dynamic and parametric approach is developed for recruiting and developing the right human talent.

Data mining has many algorithms for knowledge discovery in databases. Due to this it is also called KDD (Knowledge Discovery in Databases). The research involves applying various data mining techniques and tools for analyzing capability attributes of people involved in software development. Classification has been used and Weka tool which is an open source machine learning which has many data mining algorithms has been extensively explored for getting the patterns.

The further chapter deals with specific data mining techniques used in this study, results obtained and also inferences and conclusions.