CHAPTER 3

Framework for mining project personnel data for performance prediction

3.1 Introduction

3.2 Research Methodology

3.2.1 Defining Objectives along with constraint
3.2.2 Data Collection
3.2.3 Data Analysis and Preprocessing
3.2.4 Mining Models
   3.2.4.1 Classification
      3.2.4.1.1 Bayesian Classification
      3.2.4.1.2 Decision trees

3.2.5 Implementation of methods

3.3 Interpretation and Summary
3.1 Introduction

Human aspect of software engineering should be considered as a significant organizational asset. An increasingly competitive and global business environment of software industry, along with ever-improving quality and customer service, has highlighted the importance of recruitment and selection of appropriate project personnel. In this context, the application of the appropriate methods for finding the factors that enhance the project personnel's performance can lead to the improvement of the software development process and company profit. In order to implement strategies related to human talent evaluation systems, there is a need to ascertain if existing personnel resource processes are supporting or causing hindrance to the growth of the organization (A TP Track Research Report 2005). It is utmost important to design and develop a methodology to manage the large data base related to project personnel and extract useful information from it. Based on the findings by appropriate methods, strategies and capability models can be developed.

Data mining is an approach to assist companies in developing more effective strategies. It is a computational and accurate technique for providing information and knowledge extraction. Data mining technology can be used to transform hidden knowledge into manifest knowledge. In this research, use of data mining for performance analysis of project personnel will effectively provide capability related knowledge which when integrated with company culture will provide a decision support system for enhancing the software development process significantly.

3.2 Research Methodology

This study deals with investigation of performance of the project personnel and carries out root cause analysis of performance using data mining techniques and further provides a solution to fill performance gaps by suggesting a capability work culture model for improving the quality of the software developed which in turn will provide a competitive edge for the company.
For achieving the above mentioned objective, this research built a research framework for exploring technical abilities or capability factors needed in a programmer for contributing towards good performance during software development.

To investigate such factors data mining techniques were considered.

Data mining is the process of discovering hidden knowledge from data. It uses a combination of a knowledge base, sophisticated analytical algorithms, domain knowledge to reveal hidden patterns.

Weka tool is an open source easily available tool which has a wide range of available algorithms for such purpose. Weka is a cluster of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a dataset. Weka contains all features and methods to able data pre-processing, visualization of data and applying classification, regression, clustering, and association rules.

Intelligence and prior knowledge is required in every phase of software development. Motivation for this study is to enhance the quality of software developed by investigating the human aspects and desired quality attributes required in project personnel in order to develop required quality of software products.

Large amount of dormant data related to project personnel is available in the repository of software industries. Hence, it is possible to discover the pattern in the relationship between employee’s attributes and their performance at work. This research study therefore orients to find through machine learning methods those hidden technical attributes which are needed for good performance in a software engineer. Thus, the hypothesis framed is that project personnel working on similar projects will tend to show similar performance characteristics. However, this research has many constraints. To consider few of them, each project personnel have a different background. Companies have different work culture due to which the performance of same personnel may vary. Also, performance is dependent on many non-technical aspects such as job satisfaction and other psychological issues. Therefore, study of human aspect of software engineering faces many restrictions due to variability of human behavior. Due to the above mentioned reasons, this study limits to investigate technical aspects of project personnel, which
will contribute to high performance for software products. Yet again one more challenge is wide variety of software project domains and types within each domain. Hence, this study is focused to analyze human components based on technical characteristics for projects dealing with web based applications.

This research work involved applying classification models to find patterns in work behavior of software project personnel based on few parameters related to project, project personnel and company culture. Through this study, relationship can be drawn between performance and personnel's attributes. The discovered knowledge from the data mining models can be deployed for finding the right people and thereby enhancing the software development. This research therefore aimed to construct a framework using data mining techniques to explore the relationships between personnel profiles and performance at work and its effect on software development. Through the proposed methodology, hidden information could be extracted from large volumes of personnel data and thus the project leaders are able to comprehend and focus on selection for the software project through the discovered knowledge. Fig 3.1 shows the framework and flow of research.
The research framework consists of the following steps - Problem Definition, data collection and preprocessing, applying classification technique and deriving results. Firstly the research objective was analyzed and the research problem was defined along with framing of hypothesis considering the constraints. Based on the problem definition, data was collected. The data was preprocessed accordingly for applying to various classifiers. The results from various classifiers were analyzed on the basis of results generated and also accuracy. Based on it management could frame rules related to selection and retention criteria of the project personnel. Through the proposed methodology, hidden information can be extracted from large volumes of personnel data and thus the project leaders are able to comprehend and focus on selection for the software project through the discovered knowledge.
3.2.1 Defining objectives along with constraints

One of the challenges that any industry faces is to assign right type of people for the development process of software products. Data mining techniques enables one to analyze the skill set of people and their relevance to the type of projects. This work thus enables one to apply data mining techniques to analyze large volumes of data regarding personnel information pertaining to technical characteristics in order to reveal those attributes which contribute to good performance. Further, some major constraints during study was that this study is limited to web based applications which are more of service than product which includes applications of type, property portal, online shopping sites, search engines, email portals etc. Rationale for this limitation is due to constraint of research time and resource availability of empirical data from industries. However, with the data set obtained, a deep study of literature survey and industry based investigations; this research work progressed by formulating the hypothesis as below.

**Hypothesis** – Certain attributes pertaining to capability of project personnel working on similar project and under similar work conditions impact the project personnel's performance and thereby influence the quality of software that is developed and the growth rate of company.

Some of the attributes considered are programming skills, reasoning skills, college tier, academic scores, degree obtained, experience, job satisfaction etc. Thus the impact of such attributes on performance is estimated using data mining techniques.

Project personnel information was collected from those projects which are developed in similar domain using similar technology and programming language. The constraint of this study is that it has dealt with web based projects with windows operating system and C/C++/PHP as programming language. These projects are more of service type than products based applications.
3.2.2 Data Collection

After framing the objectives the research, the next stage was data collection. Software industry is vast ocean where there are many giant as well as small startup companies. Also companies are specialized in various types of projects. Some companies deal with ERP systems, some with operating systems and system programming, and some with networking solutions. Some companies look after legacy systems. Some companies deal with web based applications like mailing, social networking sites, property portals, shopping portals etc. Software has entered in our pockets and at tips of our fingers through mobile computing. The list is endless, since software has penetrated in every domain of life.

The data regarding software industry is itself a vast population since software industry is spread over vast domains. Every type of software requires a different expertise. It was extremely important to narrow down on and draw the constraints of the study. This study concentrated on few companies which developed web- based applications. Therefore the first constraint was that all empirical study was done on a particular domain of software i.e. web based application. Also the company locations were in Bangalore. The second constraint was that software companies in Bangalore were considered. The sampling was purposive. It was observed that a giant company which was first to introduce mailing service and messenger was at the verge of closing down. A similar company which started with just three college students had dominated the global market. Likewise a property portal which lasted for many years got subdued by emerging software. This motivated to look into human aspects angle of these companies and find out its role for growth of company.

Data was collected from reputed software firms of Bangalore dealing with web based applications. Data regarding project personnel is generally with human resource department as it is this department which is authorized for selection and recruiting of personnel to their organization. However, this study requires many attributes for consideration so data was collected partially from human resource department, majorly from project team lead, few from questionnaire to project personnel and some from internet sources. The main data sources were as follows
• Project personnel Profile: Human Resources Department has got information about the personnel. It contains many attributes like gender, college tier, degree obtained, academic scores, experience etc.

• Questionnaires: forms which are completed and returned by respondents. These questionnaires were given to the project team leaders and a different set to the project personnel too. Through these many variables like social, economic and family background could be derived. Questions regarding rating of project personnel in terms of team spirits, time efficiency and skills were derived from project manager. Questions regarding job satisfaction and contentment with the job in terms of work culture and pay were also asked from the project personnel.

• Interviews: forms which are completed through an interview with the respondent. More time consuming than questionnaires, but they are better for more complex questions.

• Direct observations: making direct measurements is the most accurate method for many variables

The data that was collected came from various sources. Also the data was very large. More than 25 companies were visited. More than 2500 employees were surveyed. More than 500 projects were considered and project managers were interviewed.

However, at this point it was necessary to determine the sample for the empirical study. The sample should be optimum as to represent the population of the data. If too large, the results may not be consistent. If too small, it may again not give proper and accurate results.

In order to collect the required data, a questionnaire was prepared and given to both project leaders and project personnel. Information regarding various attributes was asked in the questionnaire like degree obtained, college tier, age, gender, social background, family background, native city, time efficiency, satisfaction with the job in terms of pay and work
culture, languages familiar with, experience in similar project, certifications of various courses etc. considering the fact that many attributes might predict the performance class. Attributes involving personal, educational, skill assessments results and work related details were collected. Factors like gender, social and economic background were not taken to avoid any discrimination. This study took attributes related to educational background, internal assessment results based on skill factors and work culture of company for analysis and performance prediction.
## Table 3.1: Initial List of attributes

<table>
<thead>
<tr>
<th>Eligibility Checks</th>
<th>Factors contributing to recruitment of right talent for successful delivery of software</th>
<th>Fill Relevant Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Work Permit required to work and Validity</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Relocate/Acceptance to work at the required place</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Medical fitness to perform assigned work</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SC/ST criteria to be met</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Background Verification done</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>References checked</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identifying information</th>
<th>Fill</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
</tr>
<tr>
<td>3</td>
<td>Sex</td>
</tr>
<tr>
<td>4</td>
<td>Marital Status</td>
</tr>
<tr>
<td>5</td>
<td>Family</td>
</tr>
<tr>
<td>6</td>
<td>SSN/PAN/Passport Details</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Educational Competency</th>
<th>From personal file</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>School</td>
</tr>
<tr>
<td>2</td>
<td>Degree Obtained</td>
</tr>
<tr>
<td>3</td>
<td>Professional Certifications</td>
</tr>
<tr>
<td>4</td>
<td>Research Work</td>
</tr>
<tr>
<td>5</td>
<td>Publications</td>
</tr>
<tr>
<td>6</td>
<td>Patents obtained</td>
</tr>
<tr>
<td>7</td>
<td>Academic scores</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Past Experience</th>
<th>From HRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Position</td>
</tr>
<tr>
<td>2</td>
<td>Responsibilities</td>
</tr>
<tr>
<td>3</td>
<td>Salary</td>
</tr>
<tr>
<td>4</td>
<td>Past Experience</td>
</tr>
<tr>
<td>5</td>
<td>Trainings</td>
</tr>
<tr>
<td>6</td>
<td>extracurricular activities, hobbies etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Capability Tests</th>
<th>To measure the qualities, abilities, and performance of the candidate</th>
<th>rated by project manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reasoning skills</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Programming skills</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Domain skills</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Time efficiency</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>communication skills</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interview</th>
<th>Direct Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewing: After clearing the employment test, the candidates are now has to face the next step in selection procedure. For selection of deserved candidate, interview is very important. The main objective of interviewing is i) to measure the applicant against the specific requirement of the job, ii) to find out the suitability of the candidate, iii) to seek more information about the candidate</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Team Working Skills</td>
</tr>
<tr>
<td>2</td>
<td>Potential for growth</td>
</tr>
<tr>
<td>3</td>
<td>Methodology</td>
</tr>
<tr>
<td>4</td>
<td>Strengths</td>
</tr>
<tr>
<td>5</td>
<td>Weaknesses</td>
</tr>
<tr>
<td>6</td>
<td>Threats</td>
</tr>
<tr>
<td>7</td>
<td>Habits/Addictions</td>
</tr>
<tr>
<td>8</td>
<td>Personality, confidence, responsiveness- general efficiency</td>
</tr>
</tbody>
</table>
The above mentioned attributes were taken into consideration during data collection. Few attributes were taken from human resource department, few from project managers and some by interviewing the project personnel. However out of these many only few were important to the management. Therefore many were removed purposively and few were removed for unbiased conclusions.

3.2.3 Data Analysis and Preprocessing

The data has come from various sources as shown in data collection. After the sample was selected, it was very necessary to integrate and normalize the data as per the requirements of the classification techniques.

Data preprocessing is the most crucial aspect in data mining. More than 35 attributes were taken initially. Out of which only few technical were taken into consideration without age, gender, social and economic discrimination. Data preprocessing involves reduction of unnecessary attributes and secondly redundant tuples. Management of companies was interested in knowing the importance of few important attributes. Therefore some attributes were removed as per instructions. Some were removed to avoid any age, gender, social and economic discrimination. Later some attributes which ranked very low in hierarchy after applying data preprocessing methods were removed so that concentration is there on most important factors. Data preprocessing can be carried out by various methods such as Genetic Algorithm, Neural Networks Rough Set Theory and Entropy Based Discretization. Basically, all methods give a hierarchical order of attributes so that unimportant attributes can be easily trimmed. This study has used Entropy Based Discretization method since it is capable of extracting knowledge from large and unclear dataset. Entropy Based Discretization reduces the attributes and also helps in improving the accuracy. Sampling of data set should be such as to represent the entire data properly and removing only the redundant tuples. After removing the unimportant attributes, the redundant tuples were removed by using ID3.
The approach for data preprocessing has been shown in Figure 3.2. As shown in figure 3.2 it has been done in two phases. Firstly unimportant and insignificant attributes have been removed. This step is also called column reduction. The second phase was reducing unimportant and redundant tuples. This phase is also called row reduction.

Identifying and eligibility attributes were not considered for the research since the research focused in finding those technical factors which influenced the performance. Some attributes were integrated, aggregated and normalized. Some were not considered since management wanted to know the impact of few attributes only. Some were further removed since they were ranked very low in hierarchy of importance by data preprocessing method i.e. when column reduction was done. The final list of attributes that were taken for data analysis is as shown in Table 3.2. The column reduction was done through two methods i.e. Information Gain and Gain ratio in Weka tool which is based on Entropy.

The attribute selection methods for column reduction are explained below.

- **Attribute Selection Measures (Column Reduction)**

Attribute selection measure is a heuristic approach for splitting the data set. Therefore, they are also called splitting rules in forming decision trees. Attribute selection method also provides ranking of attributes in the data set. The most important attribute comes as root and subsequently
other attributes come in the tree depending on their importance. Attribute selection measures can be used for data pre processing too for column reduction i.e. reducing unimportant attributes.

Two methods were taken for achieving the same. They were information gain and gain ratio, both based on entropy. Entropy is considered to be a very good discretization measure. It was introduced by Claude Shannon in his premium work on information theory (Weiss et al., 1998).

Information gain used in ID3 classification is also based on the same concept. Also Gain Ratio which is used in CART decision tree is based on Entropy. Entropy based discretization reduces data size effectively. It can be used for row reduction as well as column reduction. Column reduction methods are also called attribute selection methods.

Let D be the data set. Let there be m class labels or output classes denoted by C_i (i = 1..m). Let |D| and |C_i,D| denote the number of tuples in D and number of tuples in class C_i. Further is explained the concept of information gain and gain ratio.

a) Information Gain

Information gain is used in ID3 for splitting criteria. Information gain is based on entropy. The attribute with highest information gain is chosen as the splitting attribute for node N. The attribute which minimizes the information needed to split the tuples and reflects the least randomness is chosen.

Information needed to classify a tuple D is given by

\[ \text{Info} (D) = - \sum_{i=1}^{m} P_i \log_2 (P_i) \] ........................equ(3.1)

where class labels have m distinct class values. C_i,D is the set of tuples of class C_i in D. let |D| and |C_i,D| denote the number of tuples in D and C_i,D. In equation (3.1), p_i is the probability that a tuple in D belongs to C_i and is estimated by |C_i,D|/|D|
The data set consists of attribute A having v distinct values. When test is carried out on attribute A, there will be v outcomes. Or attribute A can be used to split D into v partitions (D_1..D_v), where Dj consists of those tuples in D whose outcome is a_j. These partitions are also the branches from node N. However, when splitting takes place it contains tuples from other classes too. Therefore, the partitions are generally impure. At this point, there is the need of information or also called entropy about the classification. Info_A(D) is the expected information required to classify the tuple from D based on the partitioning by A.

\[
\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info}(D_j) \tag{equ 3.2}
\]

where \( \text{Info}(D_j) = -\sum_{i=1}^{m} p_i \log_2(p_i) \) \tag{equ 3.3}

where p_i is the probability that a tuple in D belongs to C_i having attribute value j.

Information gain is given by

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \tag{equ 3.4}
\]

The attribute with highest gain is chosen as the splitting criteria at node N.

Information also gives the importance of attributes and therefore is a good way for attribute selection and data reduction.

b) Gain Ratio

To overcome the shortcomings of information, Quilan used gain ratio in C4.5 which is successor of ID3. It uses the extension of information gain by using split information. Split Information normalizes the information gain and is denoted in equ( 3.5).

\[
\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|}) \tag{equ 3.5}
\]
It differs from information gain since it considers the number of tuples having that particular outcome or class label with respect to total number of tuples in D. The gain ratio is defined in equation 3.6.

\[
\text{GainRatio (A)} = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \quad \text{(equ 3.6)}
\]

The attribute with maximum gain ratio is selected for splitting.

Applying attribute selection methods i.e. Information Gain and Gain Ratio, unimportant attributes were removed. Weka tool kit has the facility for the same. Table 3.2 shows the results of attributes ranking in Weka tool kit.

Other method for discretization is Histogram Analysis. In this method values are partitioned such that each partition contains same number of tuples. It is then recursively applied until the pre specified number of levels has been reached.

Information gain in Weka tool gave an attribute ranking. Thereafter the attributes with low rankings or irrelevant attributes were eliminated. Thereby column reduction was achieved. This trims the data considerably for analysis. Table gives the results obtained from using the two methods which use Entropy Based Discretization in Weka Tool i.e. Information Gain and Gain ratio. The table gives the attribute listed in hierarchical order according to their importance. Based on this, the attributes with lower ranking can be removed column wise for application of mining models.

- **Reduces unnecessary and redundant instances (rows reduction)**

The second phase is to employ ID3 for reducing noisy instances. Then, a new data set from the results of entropy based discretization will be selected as the inputs for decision trees for further analysis to reduce the instances of redundancies or impurities of the original data.

ID3 gives the tree with the probability associated with each branch. The branches which have null value associated indicate those tuples which are noisy and do not have relevance. Those tuples are removed from the data set. Some sample rules with their associated probability are
shown below. Such tuples with very low probability were removed from the data set and this is how row reduction was achieved.

The algorithm for data reduction is given as below

Input data set with reduced attributes

Perform n-fold cross validation by ID3 based on information gain

For each fold i, i = 1,......n

Use i th fold as testing data set

Use the union of the other n -1 fold as training data set

FOR each instance j in testing set

{
    IF instance j has high probability
        Put instance j into final_data_set

}

The data set is now ready for analysis. Without data pre processing the trees and results are difficult to interpret since it amounts to huge unclear trees.

As can be seen some vectors are having very low occurrences and therefore low probability. These tuples are removed in MSEXCEL easily by writing a pseudo code. These tuples are necessary to remove since they create ambiguity in the results by creating large trees which are not easy to interpret.

3.2.4 Mining Models

Data mining has got a big collection of methods for mining knowledge. It can be done by association, clustering, classification or regression methods. All these methods have been briefed in chapter 2 section 2.3 in overview of data mining. Considering the data type for this study, it was most appropriate to apply supervised learning method since the output class labels were
available as seen in the data. Therefore, this study the main focus was on classification. Various classifiers which were used are detailed below.

### 3.2.4.1 Classification

There are many classification techniques such as neural networks, K-nearest methods and support vector machines. In this study, however bayesian classification and decision trees are used since decision trees are easy to interpret and understand and it also gives rules which acts as the final information needed for management policies. Decision tree is a flowchart consisting of internal nodes, leaf node or terminal node and branch. Branch represents an outcome of a test, internal node represents a test on an attribute and each leaf holds the class label. The most popular decision trees are ID3, CART, and Random Forest.(Witten I et al. 2005).

Classification is a major aspect in data mining. It predicts the y given the values of a vector of predictor variables x. In this study Y, a finite set of unordered values i.e. Performance outputs are classified based on input attributes X which are technical capabilities of project personnel. The following classification methods are used in this research.

#### 3.2.4.1.1 Bayesian Classification

Classification by bayes method is the process by which a model is created or chosen to try to best predict the probability of an outcome. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data. Bayes Classification is a predictive data mining technique which makes prediction using historical data. Predictive models have the specific aim of allowing one to predict the unknown value of a variable of interest given known values of other variables. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined by examining the data by expert or many experts of that domain.

Bayes classification in pattern recognition and data mining methods are developed based on Bayes rule of conditional probability. Bayes theorem offers a way to unfold experimental distributions in order to get the best estimates of the true ones. Bayes rule is a technique to estimate the likelihood of a property given the set of data as input also called evidence. The
approach is called “naïve” because it assumes the independence between the various attribute values. Naïve Bayes classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and are then used to predict the class membership for a target tuple with certain values of the attributes. Therefore it is predictive too. The naïve Bayes approach is simpler to use because it requires a small training data set and over-simplified assumptions.

Naive bayes has been used in many real time experiments for prediction. Bayesian Networks provides a probabilistic method of reasoning under uncertainty. It reveals the patterns in the data which illustrate the high probability factors or also called reasons. It has been used for predicting software defects where the probability of defects assisted in removing the bugs (Fenton et al., 2008). It has also been used by Chien et al. (2002) for finding the location of fault in a electric power delivery system based on the database provided. The bayesian network could classify the fault and non fault depending on the probability associated.

The Bayesian decision making refers to choosing the most likely class, given the value of the features or attributes. The probabilities of class membership are calculated from the bayes' theorem. Bayes theorem is explained below:

If the tuple X is denoted by vector(x1-----xd) and class of Ci , given the probability p(Ci) and P (X|Ci) which denotes the prior probability that the random sample is a member of class Ci and P(X/Ci) is the conditional probability of obtaining attribute values X given the sample is from Ci. Our goal is to estimate the probability that a sample belongs to class Ci , given that it has attribute values X which is denoted by P(Ci|X) which can be calculated according to 3.1 as stated by bayes theorem.(Gose et al.1996).

The Derivation of bayes’ classification can be thus written as below:

D : Set of tuples
Each Tuple is an ‘d’ dimensional attribute vector
X : (x1,x2,x3,…. xd)
Let there be ‘k’ Classes : C1,C2,C3…Ck
If there are \( d \) attributes or features and \( k \) classes, then probability of the attribute vector is denoted by equation 3.7.

\[
P(X_1, \ldots, X_d) = \sum_{j=1}^{k} P(C_j)P(x_1, \ldots, x_d|C_j)...............(\text{equ 3.7})
\]

which can be computed assuming that each attribute is independent within each class by equation 3.8

\[
P(x_1, \ldots, x_d|C_j) = P(x_1|C_j)*P(x_2|C_j)*\ldots*P(x_d|C_j)........(\text{equ 3.8})
\]

Bayes theorem of conditional probability states that a tuple with attributes values \( x_1, x_2, \ldots, x_d \) belonging to class \( C_i \) is denoted by equation 3.9. (Hogg et al.,1983).

\[
P(C_i|x_1, \ldots, x_d) = \frac{P(C_i)P(x_1|C_i)\ldots P(x_d|C_i)}{\sum_{j=1}^{k} P(C_j)P(x_1|C_j)\ldots P(x_d|C_j)} ...... (\text{equ 3.9})
\]

Naïve Bayes classifier predicts \( X \) belongs to Class \( C_i \) if

\[
P(C_i/X) > P(C_j/X) \text{ for } 1 \leq j \leq k, j \neq i
\]

Maximum Posteriori Hypothesis is given by equation 3.10

\[
P(C_i /X) = P(X/C_i) * P(C_i) / P(X) \text{ ..........(3.10)}
\]

Bayes classification aims to Maximize \( P(X/C_i) * P(C_i) \) as \( P(X) \) is constant.

With many attributes, it is computationally expensive to evaluate \( P(X/C_i) \). Therefore Naïve Assumption of “class conditional independence”. The final derived equation assuming class independence is given by equation 3.11 and 3.12

\[
P(X/C_i) = \prod_{k=1}^{d} P(x_k/C_i) \text{ ..........(equ 3.11)}
\]

\[
P(X/C_i) = P(x_1/C_i) * P(x_2/C_i) * \ldots P(x_d/C_i)........(\text{equ 3.12})
\]
Here, X is the vector related to the project personnel having d attributes with values x1, x2,...,xd. Also output classes are: C1- good, C2- average and C3-poor for performance. This research aims at finding through Bayesian classification, attribute values which gives high probability in the respective classes as per equation 3.12.(Butler et al.,1992).

3.2.4.1.2 Decision Trees

A decision tree is called a tree since it is very similar to a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Decision trees are used commonly in classification and prediction since it is simple and powerful way of knowledge discovery. The construction of decision tree does not require any parameter setting or domain knowledge and therefore is very popular technique. Decision tree is a structure where each internal node is denoted by rectangle and represents a test on an attribute. Branch represents an outcome of the test and each terminal node holds the class label. Leaf nodes are denoted by ovals. The node at the highest level is the root node. The decision tree has two phases i.e. growth and pruning. The first phase the tree is constructed based on attribute selection measures. Attribute selection measures are used to select attribute that best splits the tuples into distinct classes. Tree pruning indentifies and removes outliers in the training set to improve the accuracy of the classification method. Tree pruning can be done pre or post building the tree.
The main elements decision tree algorithms are basically:

- Rules for splitting data at a node based on the value of any one variable; by Gini Index, Information Gain or gain ratio
- Stopping rules for deciding when a branch is terminal and can be split no more and some threshold on improvement on accuracy, same output for all cases, very low Information Gain.
- Finally, a prediction for the target variable in each terminal node.

Decision trees are a hierarchical structure like a tree with leaves and branches. The structure of decision trees represents different levels of attributes hierarchically. Every leaf reveals the classification of an attribute based on split criteria, while the branch indicates the conditions of the attributes. A decision tree can be constructed by various methods to provide valuable information about the attributes and reveals patterns and rules thereby associated among the attributes (Maimon et al. 2005).

ID3 which is based on Hunts algorithm was developed by Quilan (1986). The tree is constructed in two phases simultaneously. Firstly it builds the tree along with pruning. Quilan further developed C4.5 which is a successor to ID3 and is based on Hunt’s algorithm. CART is another popularly available algorithm. CART was introduced by Breiman. It handles both continuous and categorical attributes to build a decision tree in addition to handle missing values. CART works in such a fashion as to build only binary trees. ID3 and C4.5 algorithms have multiple branches (Daniel et al. 2006).

The various decision trees have been elaborated in chapter 2 too. Table 3.2 gives the brief description of the various classification methods used in this research in tabular form. It gives the name by which it was introduced, the type of data it can handle and the level of knowledge it exhibits.

Weka contains tools for regression, classification, clustering, association rules and visualization. In this study, classification panel is used. The result not only gives the tree but gives other
accuracy parameters of the classifier too. Further chapters in this thesis will discuss more about results obtained i.e. the trees, rules and accuracy parameters from various mining methods.

Table 3.2: Description of Classification techniques

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Proposed by</th>
<th>Dealing with data types</th>
<th>Speed of classification</th>
<th>Knowledge extraction of data from classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id3</td>
<td>Quilan</td>
<td>discrete and continuous</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
<tr>
<td>J48 pruned tree</td>
<td>Quilan</td>
<td>discrete and continuous</td>
<td>Excellent</td>
<td>Good</td>
</tr>
<tr>
<td>RandomTree</td>
<td></td>
<td>discrete and continuous</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
<tr>
<td>CART Decision Tree</td>
<td>Breiman</td>
<td>discrete and continuous</td>
<td>Excellent</td>
<td>Good</td>
</tr>
<tr>
<td>Naivebayessimple</td>
<td>Baye</td>
<td>only discrete</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

Further, short note on all the classifiers is provided for understanding the origin, differences and methodology of the techniques.

a) ID3 (Iterative Dichotomise 3)

ID3 was developed by Quinlan Ross which was based on concepts of learning systems described by Hunt(Mannila et al.,2001). 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.

b) C4.5 (also called J48 in Weka tool Kit)

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(Quinlan,1993). 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. 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.

c) CART

CART was developed by Breiman Weiss et al., (1998). 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.

d) Random Forest

The training algorithm for random forests applies the general technique of bootstrap aggregating or bagging. Given a training set $X = x_1, \ldots, x_d$ with responses $Y = y_1, \ldots, y_d$, bagging repeatedly selects a bootstrap sample of the training set and fits trees to these samples. A bagging classifier has greater accuracy typically; a few hundred to thousand trees are used, depending on the size and of the training set. Random forest gives accurate results and is more capable of handling noisy data. This is because the composite model reduces the variance of individual classifiers (Kohavi, 2008).

3.2.5 Implementation of classification methods

Firstly to experiment was conducted using a small data set and using Bayes Classification. On getting a positive pattern and important revelation, Weka tool kit was used to continue the experiments and validate the result of bayes classification.
The algorithm used for classification in this study is ID3, C4.5, random forest and CART. Under the "Test options", the 10-fold cross-validation is selected for the evaluation approach. Since, there is no separate evaluation data set, this option was necessary to get a reasonable idea of accuracy of the generated model. The model is generated in the form of decision tree as shown in the following chapter. These predictive models provide analytical way for performance analysis.

3.3 Interpretation and Summary

Data is subject to various classification techniques like Bayes classification, ID3, J48, CART and Random Forests discussed in this chapter. Though there are numerous data mining techniques, few were taken into consideration. The results and interpretations along with inferences are discussed further. The results are depicted in chapter 4. Chapter 5 deals with derivation of a model based on rules derived from results of chapter 3. Chapter 4 and 5 interprets the results in details and validates the results by empirical study across few companies.

The main objective of the study is to find those capability factors of the project personnel that impact performance. This investigation focused upon using data mining methodology for selection and recruitment of right project personnel that yield effective results for better software quality. The objective was to deploy data mining technique to focus on the capability factors of individual project personnel.

Human aspect of software engineering has become one of the main concerns in software companies to achieve quality objectives. Software industries are now paying attention to select the right talent who can perform consistently throughout all generic framework activities and execute the process properly. Software quality depends on people and process quality during development(Godfrey et al.,2009). Hence, this study proposes the use of classification algorithms that can exploit the patterns in the historical data and predict the performance based on project personnel attributes and thereby enhance the process and quality of software.

Data mining is the process of extracting hidden information from data. It has sophisticated analytical methods to uncover hidden trends and patterns(Daniel et al.,2005). These trends and
patterns can be extracted on by using various data mining algorithms. The important steps involved in data mining are data preprocessing and transformation of data for applying mining model. The data undergoes column reduction and row reduction for getting a good sample of data so that the results derived are interpretable. In this study, classification has been used since it gives clear and understandable outputs. Classification a supervised training method since the output classes is determined in this data. Bayesian classification and decision trees are used as mining models. Bayesian classification will give the high probability attributes which contribute to good performance. Decision trees will reveal in a form of tree with the most important attribute at root and subsequently other attributes in hierarchical order. Subsequently, patterns and detailed statistics can be obtained by the mining models which are illustrated in next chapter.

Through classification, one can derive rules. Categorization uses rule induction algorithms to handle categorical outcomes. In this study therefore classification was used effectively since data consisted of discrete values.

This research therefore aimed to construct a framework for human resource data mining to explore the relationships between personnel profiles and its effect on performance and thereby impact on software development. Through the proposed methodology, hidden information could be extracted from large volumes of personnel data and thus the project leaders are able to comprehend and focus on selection and recruitment of right project personnel for the software project through the discovered knowledge.

This study enables the management of software industry to refocus on human capability criteria and thereby enhance the development process of software project. A lot of importance is given to process within generic framework of software development and now it is time for deep investigation of human aspect for effective software development.