# Data Mining: Evaluating Performance of Employee's using Classification Algorithm Based on Decision Tree

*Rohit Kotalwar*[1]
K J Somaiya College of Engineering,
University of Mumbai, India
Rohit.k@somaiya.edu

*Raj Chavan*[2]
K J Somaiya College of Engineering,
University of Mumbai, India
Raj.chavan@somaiya.edu

*Sunny Gandhi*[3]
K J Somaiya College of Engineering,
University of Mumbai, India
Sunny.gandhi@somaiya.edu

Vinayak Parmar[4]
K J Somaiya College of Engineering,
University of Mumbai, India
vinayak.p@somaiya.edu

***Abstract*---The ability to predict a employee's performance is very important in industrial Sector. Employee's performance is based upon diverse factors like personal, social, skills, punctuality, dependability, Interpersonal relation, task/responsibility, Work standard, actual performance, Psychological and other environmental variables. Data Mining is a tool which can be used to accomplish the objectives. Data mining techniques are used to discover hidden information patterns and relationships of large amount of data, which is very much helpful in decision making. A single data contains a lot of information. The type of information is produced by the data and it decides the processing method of data. A lot of data that can produce valuable information, in industrial sector contains this valuable information. The performance of employee is calculated, to know them well, the best way is by using valid management and processing of the employees' database .**

*KEYWORDS: **DATA MINING, DECISION TREE, EMPLOYEES EVALUATION.***

## I. INTRODUCTION

Data Mining sometimes is also called knowledge discovery in databases (KDD). We can also find the existing relationships and patterns. Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge. Employee's retention has become an indication of industrial performance and enrollment Management. Here, potential problem will be identified as earlier. The raw data was preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. One of the most useful data mining techniques is Decision Tree. Classification maps data into predefined groups of classes. It is often referred to as supervised learning because the classes are determined before examining the data. The prediction of employee's performance with high accuracy is more beneficial. To improve their performance the supervisor will monitor the employee's performance carefully.

(a) Generation of data source of predictive variable.

(b) Identification of different factors, which affects an employee's performance during industrial career
(c) Construction of a Decision Tree model using Classification data mining techniques on the basis of identified predictive variables and their values.

## II. DATA MINING TECHNIQUES

### A. *CLASSIFICATION*

Estimation and prediction are viewed as types of classification. The problem usually is evaluating the training data set and second applied the model developed.

CLASSIFICATION ALGORITHMS

| TYPE | NAME OF ALGORITHM |
|---|---|
| Statistical | Regression, Bayesian |
| Distance | Simple distance, K nearest<br>Neighbors |
| Decision Tree | ID3, C4.5, CART, SPRINT |
| Neural network | Propagation, NN Supervised learning |

### B. *CLUSTERING*

Clustering groups the data, this is not predefined. By using this technique we can identify dense and sparse regions in object space. The following table provides the different clustering techniques. Clustering algorithm is best for grouping the data.

CLUSTERING ALGORITHM

| TYPE | NAME OF ALGORITHM |
|---|---|
| Similarity and distance measure | Similarity and distance measure |
| Outlier | Outlier |
| Hierarchical | Agglomerative, divisive |
| Partitional | Minimum spanning tree, squared matrix, Kmeans,<br>nearest neighbor, PAM, Bond energy, clustering with neural networks |
| Clustering<br>large database | BIRCH, DB Scan, Cure<br>Categorical ROCK |

### C. *ASSOCIATION*

The main task of this association rule mining is to find set of binary variables that frequently occurs in the transaction database. The goal of feature selection problem is to identify groups, which is correlated with each one of the targe t variable. Apriori, CDA, DDA, investingness measure etc are the association rule mining algorithm.

## III. DATA MINING PROCESS

Data are analyzed using classification method to predict the Employee performance.

### A. DATA PREPERATION

The data set used here, which is obtained from various departments. Data stored in different tables was joined in a single table after joining process errors were removed.

### B. DATA SELECTION AND TRANSFORMATION

Fields are selected, that is required for data mining. All predictive variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table 1.

Employee Related Variables

| Variables | Possible values |
|---|---|
| Quality | Good, Bad. |
| Productivity | Meets Expectations, Improvement Needed. |
| Independence | Meets Expectations, Improvement Needed. |
| Reliability | Meets Expectations, Improvement Needed. |
| Job Skills | Serious, Common. |
| Interpersonal Relationships | Meets Expectations, Improvement Needed. |
| Cooperation | Meets Expectations, Improvement Needed. |
| Commitment | Meets Expectations, Improvement Needed. |
| Attendance | Meets Expectations, Improvement Needed. |
| Initiative | Perfect, Ordinary. |
| Creativity | Meets Expectations, Improvement Needed. |
| Adherence to Policy | Meets Expectations, Improvement Needed. |
| Overall Performance | Meets Expectations, Improvement Needed. |

Table 1. Predictive Variables

A. *Quality* – The extent to which an employee's work is completed thoroughly and correctly following established process & procedures. Required paperwork is thorough and neat.

B. *Productivity / Independence / Reliability* - The extent to which an employee produces a significant volume of work efficiently in a specified period of time. Ability to work independently with little or no direction/ follow-up to complete tasks / job assignment.

C. *Job Knowledge* - The extent to which an employee possesses and demonstrates an understating of the work instructions, processes, equipment and materials required to perform the job. Employee possesses the practical and technical knowledge required of the job.

*D. Interpersonal Relationships / Cooperation / Commitment* – The extent to which employee is willing and demonstrates the ability to cooperate, work and communicate with coworkers, supervisors, subordinates and/or outside contacts. Employee accepts and responds to change in a positive manner. Accepts job assignments and additional duties willingly, takes responsibility for own performance and job assignments.

*E. Attendance* – The extent, to which an employee is punctual, observes prescribed work break/meal periods and has an acceptable overall attendance record. Employee's willingness to work overtime as required.

*F. Initiative/ Creativity* – The extent to which an employee seeks out new assignments, proposes improved work methods, suggests ideas to eliminate waste, finds new and better ways of doing things.

*G. Adherence to Policy* – The extent to which the employees follows company policies, procedures and work conduct rules. Complies with and follows all safety rules and regulations, wears required safety equipment.

*H. Lead (if applicable)* – The extent to which the employee demonstrates proper judgment and decision-making skills when directing others. Directs work flow in assigned areas effectively to meet production / area goals.

*I. Overall Performance* – Rate employee's overall performance in comparison to position duties and responsibilities.

## *IV. MINING MODELS*

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). There are different classification methods. In the present study we use the Decision Tree algorithm.

## *V. APPLICATION OF DECISION TREE ON PERFORMANCE OF EMPLOYEE*

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

To Construct the Decision Tree, we use following method

1. Select a variable of training samples as nodes, create a branch to every possible value of the variables. Accordingly, the training sample set is divided into several sub-set.

2. Do the same method to each branch, Training sample is the subset corresponding to the branches and one of the subsets which its parent node is divided into. When the node of all the training samples belongs to the same classification, or no remaining attributes can be used to further divide, Or the branch does not have samples, stop splitting the node branching and make it a leaf node.

Decision tree is a classifier in the form of a tree structure where each node is either:

- A leaf node- indicates the value of the target attribute
- A decision node- specifies some test to be carried out on a single attribute- value, with one branch and sub-tree for each possible outcome of the test.

Selection of the attribute is done by using method Information gain. To create tree, we need to find information gain value for each attribute.

Select the attribute with the highest gain.

Consider two classes, P and N.

- Let the set of examples S contain p elements of class P and n elements of class N
- The amount of information, needed to decide if an arbitrary example in S belongs to P or N is defined as

$$I\ (p, n) = -\ (p\ /\ p + n)*\log_2\ (p\ /\ p + n)\ -\ (n/\ p + n)*\log_2\ (n\ /\ p + n)$$

Assume that using attribute A, a set S will be partitioned into sets $\{S_1, S_2, ..., S_v\}$

If $S_i$ contains pi examples of P and $n_i$ examples of N, the entropy, or the expected information needed to classify objects in all sub-trees $S_i$ is

$$E\ (A) = \sum_{i=1}^{v} ((p_i + n_i)/(\ p + n))*\ I(p_i\ , n_i) \quad (1)$$

Entropy:

Expected amount of information needed to assign a class to randomly drawn objects in S under the optimal, shortest-length code.

Calculate Information Gain i.e. Gain (A): Measures Reduction in Entropy achieved because of split. Choose the split that achieves most reduction (maximizes Gain)

$$\text{Gain}\ (A) = I\ (p, n) - E\ (A) \qquad (2)$$

As following, there is an example of employees working table (as Table2). Choose 10 employees doing the same job, we will show the whole process of application of decision tree algorithm through "factors affecting employee performance"

Table 2. Employees Database

| Working Task | Job Skills | Initiative | Working Quality | Performance Result |
|---|---|---|---|---|
| Heavy | Serious | Perfect | Good | Meets Expectation |
| Heavy | Common | Perfect | Good | Meets Expectation |
| Heavy | Serious | Ordinary | Good | Meets Expectation |
| Heavy | Serious | Ordinary | Good | Meets Expectation |
| Heavy | Serious | Ordinary | Bad | Needs Improvement |
| Easy | Serious | Perfect | Bad | Needs Expectation |
| Easy | Serious | Perfect | Bad | Meets Expectation |
| Easy | Common | Perfect | Good | Meets Expectation |
| Easy | Serious | Perfect | Bad | Meets Expectation |
| Easy | Common | Ordinary | Bad | Needs |

| | | | | |
|---|---|---|---|---|
| | | | | Improvement |

A. *Attribute Selection*

The performance result is “Meets Expectation” is 7 and performance result is "Needs Improvement" is 3, just means that target class property "performance results" has two different values. In these two value, Class p=8,n=2, calculate the expectations information and information entropy to a given sample. From formula (1):

I (p, n) = - 7/10*log2 (7/10) - 3/10*log2 (3/10) = 0.8812 bits

By Formula (2)

E (working task) =5/10*I (4,1) + 5/10*I (3, 2) =0.8464 bits

E (working quality) = (5/10)*I (2,3) = 0.4855 bits

E (Job Skills) = (7/10)*I (5, 2) + (3/10)*I (2, 1) =0.7810 bits

E (Initiative) = (6/10)*I (5, 1) + (4/20)*I (2, 2) =0.87074 bits

Now, Calculating information gain:

Gain (working task) = 0.8812 - 0.8464 = 0.0348 bits

Gain (working quality) = 0.8812 - 0.4855 = 0.3957 bits

Gain (Job Skills) = 0.8812 - 0.7812 = 0.1 bits

Gain (Initiative) = 0.8812 – 0.87074 = 0.0104 bits

The maximum information gain is Working quality.

So we select working quality as root node.

Now expand root node.

Thus generating decision Tree, as figure 1

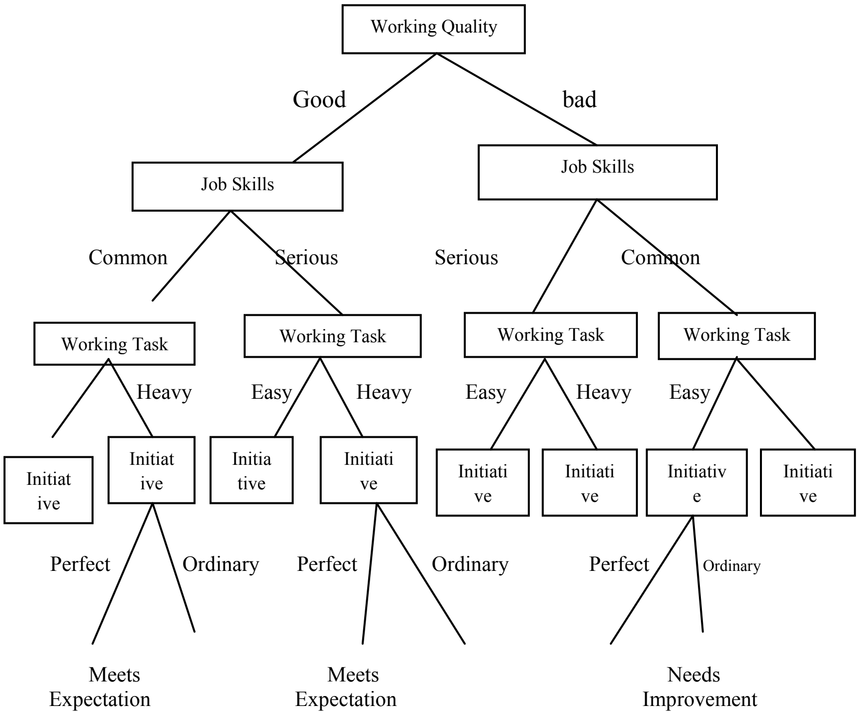

Figure 1. Decision Tree

## B. *Conclusion on application of decision tree.*

- Staffs who have Heavy tasks and good quality, their performance is good, and most of them are actively cultivate their skills, Initiativeness is perfect.
- Staffs who have Easy tasks and bad quality, surely their performance is poor, and they

mainly don't pay attention to cultivate their working skills, Initiativeness is ordinary.

- Comparing with the SQL query this algorithm performs efficiently.
- Efficiency is considered on the basis of Time Complexity. Because Redundancy factor is removed due to this scheme.

## VI. CONCLUSION

Using this research, the Industry superiors will have the ability to predict the employee's performance. A current performance evaluation is required to support recommendations for merit salary adjustments and in-grade or grade change salary increases. This also helps the supervisor to find the employee's performance and those employees needed special attention for reducing falling ratio for taking action at right time. Decision Tree method is used on Employee's database to predict the Employee's performance on the basis of previous year database. Data mining is a powerful analytical tool that enables industrial institutions to better allocate resources and staff, and proactively manage employee's outcomes. The management system can improve their policy, enhance their strategies and thereby improve the quality of that management system.

## VII. Acknowledgement

We Want to Thank the PROF. UDAY JOSHI who helped us for preparing this Journal Paper.

## VIII. REFERENCES

[1] Chen Xiaofan, Wang Fengbin," Application of Data Mining on Enterprise Human Resource Performance Management" [J]. Nanchang Hangkong University,2010(3)

[2] Dr Lakshmi Rajamani, Mohd Mahmood Ali[J],"Automation of decision making process for selection of talented manpower considering risk factor: A Data Mining Approach". University College of Engineering, Osmania University, Hyderabad

[3] Han Jing, "Application of Fuzzy Data Mining Algorithm in Performance Evaluation of Human Resource". Economic and Management Institute, YanTai University, yantai, Shandong, P. R. China, 264005

[4] E.Gothai, Dr.P.Balasubramanie,"Performance Evaluation of Hierarchical Clustering Algorithms". Department of CT-PG, Kongu Engineering College, Perudurai, Tamilnadu, India.

[5] Ujjwal Maulik, "Performance Evaluation of Some Clustering Algorithms and Validity Indices". Member, IEEE, and Sanghamitra Bandyopadhyay, Member, IEEE

[6] Suo Qi. Database Marketing Research Based on Data Mining[J]. Technology and Market,2007(2):53-54

[7] Li Heshan. The Investigation of the Establishment of Enterprise Performance Management System[J].Zhongzhou Coal,2007(1):101-102

[8] Dong Yongfeng, Hou Xiangdan, Gu Junhua, Liu Hongpu. Application of Fuzzy Data Mining in Staff Performance Appraisal[J]. Hebei University of Technology,2005(34)

[9] Ding Zhibin, Yuan Fang, Dong Hewei. Application of Data Mining in Higher Student Achievement Analysis[J]. Computer Engineering and Design,2006(4):590-592

[10] Wen Guiqing, Jia Jinlian. Sticking Point and Mining of Human Resources in State-owned Enterprises[J],Shangxi Coal Technology,2007(1)

[11]DataMing[EB/OL].http://www.stcsm.gov.cn/learning/lesson/xinxi/20021125/lesson.asp