Evaluating The Performance Of An Employee Using Decision Tree Algorithm

N. Magesh M.E.,
Assistant Professor, Dept of Computer Science and Engineering,
Institute of Road and Transport Technology, Erode-638 316.

Dr. P. Thangaraj Ph.D.,
Professor, Dept of Computer Science and Engineering,
Bannariamman Institute of Technology,Sathyamangalam – 638 401.

S. Sivagobika, S. Praba, R. Mohana Priya
Final Year CSE, Dept of Computer Science and Engineering,
Institute of Road and Transport Technology, Erode-638 316.

ABSTRACT

The main objective is to evaluate the performance of employee using Decision Tree algorithm. The data mining classification methods like decision tree, rule mining, clustering etc. can be applied for predicting the performance of an employee working in an organization. The employee data are evaluated for giving promotion, yearly increment and career advancement. In order to provide yearly increment for an employee, it should be evaluated by using past historical data of employees. The historical data stored in the table are subjected to learning by using the decision tree algorithm and the performance are found by testing the attributes of an employee against the rules generated by the decision tree classifier. This paper concentrates on collecting data about employees, generating a decision tree from the historical data, testing the decision tree with attributes of an employee and generating the output as whether to give the promotion or not. The information about an employee are collected by using the user interface. This information is compared with the trained data stored in the decision tree. The final goal node is to determine whether the employee will get yearly increment, promotion or not.

KEYWORDS- classification, learning, training, testing, prediction, Decision tree, J48 algorithm.

1 INTRODUCTION

A performance evaluation is a constructive process to acknowledge an employee's performance. Goals and objectives are the critical components of effective performance evaluation. The evaluation form needs to have a set of measurable goals and objectives spelled out for each area. In educational institutions, it is called as a self appraisal form. The main attributes present in the form are reading material, percentage of results, short term course attended and so on. These parameters are used to quantify the information about an employee. It acts as knowledge for taking conclusion. This paper concentrates on performance review of an employee in an educational institution. The performance evaluation is an important part in the principle of management which includes various tools for improving the performance of an employee.

1.1 PERFORMANCE EVALUATION

Most of the organizations will conduct the employee performance evaluation at least once a year, but it can occur more frequently when there is a clear rationale for doing the evaluation, at the end of a project, at the end of each month, and so on. McKinsey, a leading strategy consulting firm is an example which has managers to evaluate the employees at the end of every consulting engagement. So, in addition to the annual performance evaluation, consultants can receive up to 20 mini-evaluations in a year. Importantly, the timing should coincide with the needs of the organization and the development needs of the employee.

Performance evaluations are critical. Organizations are hard-pressed to find good reasons that they unable to dedicate an hour-long meeting at least once a year to ensure that the mutual needs of the employee and organization are being met. Performance reviews help the managers to feel more honest in their relationships with their subordinates and feel better about themselves in their supervisory roles. Subordinates are assured that a clearer understanding of what goals and objectives are expected from them, their own personal strength and areas for development, and a solid sense of the relationship with their supervisor. Avoiding performance issues ultimately decreases the morale, credibility of management, organization overall effectiveness, and waste most of the management time to do what isn't being done properly.

1.2 PERFORMANCE REVIEW PROCESS

At some point in the year, the supervisor should hold a formal discussion with each staff member to review individual activities to date and to modify the goals and objectives that employee is accountable for. This agreed-upon set of goals and objectives is sometimes called an employee performance plan. The supervisor should have been actively involved in continual assessment of the staff through regular contact and coaching. If major concerns arise, the performance plan can be modified or the employees can receive development in areas in which they may be weak. This is also a time for the employee to provide formal feedback to the supervisor on coaching, planning, and how the process seems to be working.

At the end of the year, a final review of the activities and plans for developing the next year objectives will begin. Again, this is a chance to provide constructive and positive feedback and to address any ongoing concerns about the activities of employee and competencies. Continuing education opportunities can be identified, and for those systems that are linked to the compensation, salary raises will be linked to the employee's performance during the year. Again, there should be no surprises to either employee or supervisor, as continual assessment and coaching should take place throughout the year. Supervisors and managers are involved in the same series of activities with their own supervisors to ensure that the entire organization is developing and focused on the same common objectives. Most of the organization collects the performance assessment by using feedback forms or self appraisal forms. The evaluation process is a critical issue in a managerial process which can be achieved by using a decision tree algorithm.

1.3 PERFORMANCE REVIEW OF EMPLOYEE

There are typically three areas to be considered during performance review: (1) preparation for the review (2) what to do if the review is negative and (3) what should ultimately take away from the review.

1.3.1 UPCOMING REVIEW PREPARATION

Document the achievements and list anything that to be discussed at the review. If the achievements haven't kept track, then sometimes have to be spent figuring out what is accomplished from the last review and, most importantly, how the employer has benefited, such as increased profits, grown the client roster, maintained older clients and so on. These are easier to document when the goals and objectives are clear.

1.3.2 POOR REVIEW PROCESS

If the outcome is an unfair review, a person who conducts that the review will be responsible. That person should initially try to discuss the review with those who prepared it. Wait until the reviewer can look at the review objectively. If a person eventually reaches the conclusion that the review was true, then set an appointment to meet the reviewer. Use clear examples that counteract the criticisms made. The process of presenting every written content will backup the presenter is known as paper trail which is always helpful. If the paper trail is not left, it is to be done in future.

1.3.3 REMEDY FOR POOR REVIEW PROCESS

The review should be regarded as a learning opportunity. For instance, if the goals and objectives are clear such that the performance was easy to document, then valuable information is taken away by the reviewer. This is used to evaluate the yearly increment, career advancement and promotion of an employee, based on the values of this attribute, the outcome is predicted using a Decision Tree algorithm. With the help of clustering and decision tree of data mining technique, it is possible to discover the key characteristics for future prediction.

In this classification, it is used to evaluate employee's performance and as there are many approaches that are used for data classification. Information like theory and laboratory pass percentage, paper presented, experience, attendance percentage was collected from employee management system, to predict the performance at the end of the year. This paper investigates the accuracy of decision tree techniques for predicting employee performance.

2 DATA MINING

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or “mining" knowledge from large amounts of data [2]. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

2.1 DATA MINING PROCESS - OVERVIEW

Various algorithms and techniques such as Classification, Clustering, Regression, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. They are explained briefly as follows.

2.1.1 CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning, the training data are analyzed by classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable, the rules can be applied to the new data tuples [3]. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

2.1.2 CLUSTERING

Clustering is the process of identification of similar classes of objects. The clustering techniques are used to identify dense and sparse regions in object space and can discover the overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as a preprocessing approach for attribute subset selection and classification.

2.1.3 REGRESSION

Regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown. Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. The process of training a regression model involves finding the best parameter values for the function that minimize a measure of the error. Linear and nonlinear regression are two different types of regression. The linear regression model is an approach to modeling the relationship between a scalar dependent variable and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

2.1.4 PREDICATION

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are the variables used in the prediction. Unfortunately, many real-world problems are not simply predictive. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification [2]. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks are used in creating both classification and regression models.

2.1.5 ASSOCIATION RULE

Association and correlation have been used to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However, the number of possible Association Rules for a given data set is generally very large and a high proportion of the rules are usually of little (if any) value.

2.1.6 NEURAL NETWORKS

The neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, the network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples [13]. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are better at identifying patterns or trends in data and well suited for prediction or forecasting needs. A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique.

2.1.7 DECISION TREES

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision [3] [13]. The concept of decision trees was developed and refined over many years by J. Ross Quinlan starting with ID3 (Interactive Dichotomizer 3) [8] [9]. A method based on this approach use an information theoretic measure, like entropy, for assessing the discriminatory power of each attribute. It is a tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Various tools are used in constructing a decision tree. The WEKA is an important tool used for constructing a decision tree. This paper uses WEKA for constructing a decision tree.

There are two operations in decision tree as follows:

Training : The records of employee with known result is trained as attributes and values which is used for generating the decision tree based on the information gain of the attributes.

Testing: The unknown records of employee are tested with the decision tree developed from the trained data for determining the result.

3 DECISION TREE LEARNING ALGORITHM

Decision tree learning is one of the most widely used and practical methods for inductive inference. The decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed on these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. The three widely used decision tree learning algorithms are: J48, ASSISTANT and C4.5.

Decision trees classify instances by traverse from root node to leaf node [1]. It starts from the root node of the decision tree, testing the attribute specified by this node, then moving down the tree branch according to the attribute value in the given set. The final output is based on values stored in the table. The result is purely based on the number of dependent variables. Hence the independent variables are not considered. Hence among many techniques the decision tree method is more suitable.

3.1 REASONS FOR USING DECISION TREE

The decision tree is commonly used for gaining information for the purpose of decision - making. Decision tree starts with a root node on which it is for users to take actions. From this mode, users split each node recursively according to a decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of the decision and its outcome.

3.2 J48 DECISION TREE

A decision tree depicts the rules for dividing data into groups. J48 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s1, s2,...$ of already classified samples. Each sample $si = x1, x2, ...$ is a vector where $x1,x2,...$ represent attributes or features of the sample. The training data are augmented with a vector $C = c1, c2, ...$ where $c1, c2, ...$ represent the class to which each sample belongs.

At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The J48 algorithm then recurs on the smaller sub lists [11] [12].

3.3 DEFINITIONS USED IN THE DECISION TREE

3.3.1 ENTROPY

Putting together a decision tree is all a matter of choosing which attribute to test at each node in the tree. A measure called information gain which will be used to decide which attribute to test at each node is defined. It is noticed that entropy is a measure of the impurity in a collection of training sets. Information gain is itself calculated using a measure called entropy, which is first defined in the case of a binary decision problem and then defined for the general case. Given a binary categorization, C, and a set of examples, S, for which the proportion of examples categorized as positive by C is p_+ and the proportion of examples categorized as negative by C is p_-, then the entropy of S is:

$$Entropy(s) = -p + log_2(p_+) - p - \log_2(p_-) \qquad \longrightarrow \quad (1)$$

3.3.2 INFORMATION GAIN

There is a problem of trying to determine the best attribute to choose for a particular node in a tree. The following measure calculates a numerical value for a given attribute, A, with respect to a set of examples, S. Note that the values of attribute A will range over a set of possibilities known as the Values (A), and that, for a particular value from that set, v, it is written as Sv for the set of examples which have value v for attribute A. The information gain of attribute A, relative to a collection of examples, S, is calculated as:

$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \longrightarrow \quad (2)$$

3.3.3 SPLITTING CRITERIA

$$\text{Split Information}(S,A) = -\sum_{i=1}^{n} \frac{|s_i|}{|s|} \log_2 \frac{|S_i|}{|s|} \qquad \longrightarrow \quad (3)$$

and

$$\text{Gain Ratio}(S,A) = \frac{\text{Gain (S,A)}}{\text{Split Information (S,A)}} \qquad \longrightarrow \quad (4)$$

The process of selecting a new attribute and partitioning the training examples is now repeated for each non terminal descendant node [1] [14]. Attributes that have been incorporated higher in the tree are excluded. So that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either of two conditions is met:

- Every attribute has already been included along this path through the tree, or
- The training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

4 WEKA

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [15] [5]. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alphanumeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from databases and file systems with simple to use options and visual interfaces [15].

4.1 ISSUES IN DECISION LEARNING ALGORITHM

Practical issues in learning decision tree include determining how deeply to grow the decision tree, handling continuous attributes, choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs and extensions to the basic J48 algorithm that address them. J48 has itself been extended to address most of these issues, with the resulting system renamed C4.5.

4.1.1 OVERFITTING

Overfitting is a significant practical difficulty for decision tree learning and many other learning methods. For example, in one experimental study of J48 involving five different learning tasks with noisy, nondeterministic data, overfitting was found to decrease the accuracy of learned decision trees by 10-25% on most problems.

4.1.2 AVOIDING OVERFITTING

A hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances (i.e., including instances beyond the training set).

4.2 GENERATING HAPPY GRAPH USING TRAINING DATA

Given a hypothesis space H, a hypothesis h∊ H, is said to **overfit** the training data if there exists some alternative hypothesis h’ ∊ H, such that h has smaller than h’ over the training examples, but h’ has a smaller error than h over the entire distribution of instances.

The Figure 1 illustrates the impact of overfitting in a typical application of decision tree learning. In this case, the J48 algorithm is applied to the task of learning. The horizontal axis of this plot indicates the total number of nodes in the decision tree, as the tree is being constructed. The vertical axis indicates the accuracy of the predictions made by the tree. The solid line shows the accuracy of the decision tree over an independent set of test example (not included in the training set). Predictably, the accuracy of the tree ever, the accuracy measured over the independent test examples first increases, then decreases. As can seen, once the tree size exceeds approximately 25 nodes, further elaboration of the tree decreases its accuracy over the test examples despite increasing its accuracy on the training examples.

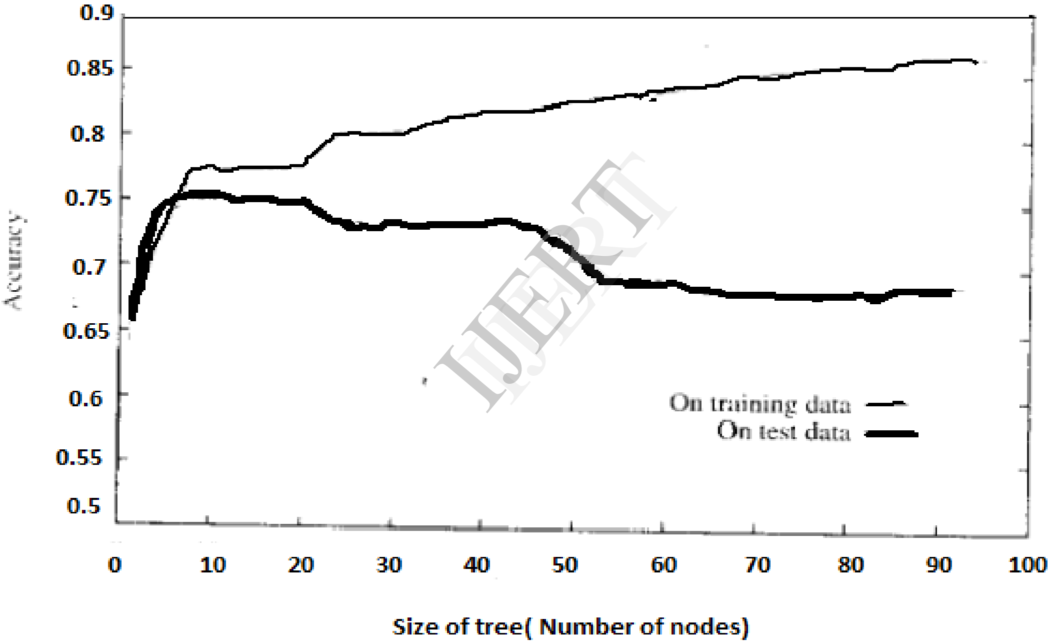


Figure 1 – Happy Graph for training the decision tree

5 EXPERIMENTAL SETUP

An employee performance is evaluated by his records and increment or promotion is given using data produced. These records can include theory and laboratory pass percentage, paper presented, national conference participated etc. It includes the entering the data in self appraisal form and feeding the form values into the table in the qualitative and quantitative format. The J48 decision tree algorithm is able to process both typesof data.

5.1 Data Preparation

Initially the size of training data sets is 15. The past data about employee are collected and stored in a table. It acts as training data for the decision tree. If the data size is increased to 50 or 60, then the happy graph is generated as shown in Figure 1.

5.2 Data Selection and Transformation

In this step only those fields were selected which were required for data mining. All the predictor and response variables which were derived from the data are given in TABLE 1 for reference.

TABLE 1 – DATA ABOUT EMPLOYEE

S.NO	ATTRIBUTES	DESCRIPTION	POSSIBLE VALUES
1	UEM	University Examination Average Marks	{Yes when UEM>=50% &No when UEM<50%}
2	TP	Teaching Plan	{Yes, No}
3	SOL	Synopsis of Lectures	{Yes , No}
4	RM	Reading Material	{Yes, No}
5	UE	University Evaluation	{Yes when UE>=50%& No when UE<50%}
6	IE	Internal Evaluation	{ Yes when UE>=50%& No when UE<50%}
7	PS	Paper Setting	{Yes, No}
8	AHA	Assessment of Home Assignments	{Yes when AHA is 2,3 &No when AHA is 0,1}
9	AQ	Assessment of Quizes	{Yes, No}
10	COE	Conduct of Examinations	{Yes ,No}
11	EOP	Evaluation of Project	{Yes, No}
12	DOI	Details of Innovations	{Yes, No}
13	RC	Research Contribution	{Yes, No}
14	IPC	Any other Improvement of professional competence	{Yes when IPC is 1& No when IPC is 0}
15	COI	Contribution in other institutional work	{Yes, No}
16	DW	Departmental work	{Yes, No}
17	CEA	Have you involved in any consultancy and extension activities	{Yes, No}
18	HAR	Honors/Awards received	{Yes, No}
19	AS	Assessment and Suggestions	{Yes, No}
20	IS	Increment in Salary	{Yes, No}

The meanings of attributes are given as follows:

- **UEM -** University Examination Average Marks. The class values are taken as real values, yes when UEM>=50% and no when UEM<50%.
- **TP -** Teaching Plan. When there is a teaching plan the value is yes otherwise no.

- **SOL -** Synopsis of Lectures. When there is a synopsis of the lectures the value is yes otherwise no.
- **RM -** Reading Material. When the employee consists of reading material then the class value is yes otherwise no.
- **UE -** University Evaluation. The class values are taken as real values. Yes when UE >=50% and no when UE<50%.
- **IE -** Internal Evaluation. The class values are taken as real values. Yes when IE >=50% and no when IE<50%.
- **PS -** Paper Setting. When the paper setting is available for the employee then the value is yes otherwise no.
- **AHA -** Assessment of home assignments. It is split into three class values: 0 for no , 1 for 1 or 2,2 for 3 or 4 and 3 for more than 4 home assignments.
- **AQ -** Assessment of Quizes. When the assessment of quizes for the employee is not available then the value is no otherwise yes.
- **COE -** Conduct of Examinations. When the conduct of examination is available for the employee then the value is yes otherwise no.
- **EOP -** Evaluation of dissertation/Project. When there is an evaluation of the project then the value is yes otherwise no.
- **DOI -** Details of innovations/contributions. When the details of innovation are available for the employee then the value is yes otherwise no.
- **RC -** Research Contribution. When there is a contribution of the employee towards the research then the value is yes otherwise no.
- **IPC -** Any other Improvements of professional competence. It is split into two values. When there are improvements then the value is 1 otherwise 0.
- **COI -** Contribution in other institutional work. When the other institutional work is available then the value is yes otherwise no.
- **DW -** Departmental work. When the departmental work is available then it is yes.
- **CEA -** Have you involved in any consultancy and extension activities. When the activities are present then the value is yes or otherwise no.
- **HAR -** honors/Awards received. When the employee receives award then the value is yes otherwise no.

TABLE 2 – TRAINING DATA

S.NO	UEM	TP	SOL	RM	UE	IE	PS	AHA	AQ	COE	EOP	DOI	RC	IPC	COI	DW	CEA	HAR	AS	IS
1	90	YES	YES	YES	85	90	YES	2	YES	YES	NO	NO	YES	1	YES	NO	YES	NO	YES	YES
2	48	YES	NO	YES	49	45	NO	0	YES	YES	NO	YES	NO	1	YES	YES	NO	YES	YES	NO
3	80	YES	YES	YES	90	90	YES	3	YES	YES	YES	YES	NO	1	NO	YES	YES	NO	YES	YES
4	45	YES	YES	NO	39	48	NO	0	NO	NO	NO	NO	YES	0	NO	NO	NO	NO	NO	NO

5	75	YES	YES	YES	70	80	YES	3	YES	YES	YES	YES	YES	1	YES	NO	YES	YES	NO	YES
6	40	YES	NO	NO	45	65	NO	1	NO	YES	YES	NO	NO	0	NO	YES	NO	YES	NO	NO
7	95	YES	YES	YES	85	90	YES	2	YES	YES	YES	NO	YES	0	NO	YES	YES	NO	YES	YES
8	35	YES	NO	YES	47	49	NO	3	NO	YES	NO	YES	NO	0	YES	YES	NO	YES	YES	NO
9	60	YES	YES	YES	70	79	YES	0	YES	YES	YES	NO	NO	1	YES	NO	YES	NO	YES	YES
10	75	YES	NO	NO	80	44	NO	1	NO	NO	NO	YES	YES	0	NO	NO	NO	YES	NO	NO
11	60	YES	YES	YES	80	87	YES	3	YES	YES	YES	YES	NO	1	YES	NO	YES	YES	NO	YES
12	50	YES	YES	YES	65	74	NO	3	NO	YES	YES	NO	YES	1	YES	YES	NO	NO	NO	YES
13	69	YES	YES	YES	70	79	NO	2	YES	YES	YES	NO	YES	1	NO	YES	NO	YES	YES	YES
14	89	YES	NO	YES	40	87	YES	3	YES	YES	YES	YES	YES	1	YES	NO	YES	YES	YES	YES
15	65	YES	YES	YES	75	45	YES	2	YES	YES	YES	YES	NO	1	NO	NO	YES	NO	YES	YES

- **AS -** Assessment and Suggestions. When there is an assessment and suggestions then the value is yes otherwise no.
- **IS -** Increment salary. If it satisfies the associative rule then it is yes otherwise no.

6 RESULT AND DISCUSSION

There are 15 datasets of the employees are taken for training. To determine the best attributes for a particular node in the tree we use the measure called Information Gain. The information gain, Gain (S, A) of an attribute A, relative to a collection of examples S. The value of information gain for each value is listed in TABLE 3 as follows

TABLE 3 – INFORMATION GAIN VALUES

S.No	Attributes	Values	S.No	Attributes	Values
1	IE	0.65829	11	COE	0.24286
2	UEM	0.596	12	HAR	0.10885
3	PS	0.51551	13	AS	0.05977
4	CEA	0.51551	14	DW	0.02584
5	RM	0.39828	15	RC	0.02584
6	AHA	0.36499	16	COI	0.02584
7	IPC	0.36499	17	DOI	0.00647
8	AQ	0.36499	18	UE	0
9	SOL	0.36499	19	TP	0
10	EOP	0.36499			

Selected attributes: 6,1,7,17,4,8,14,9,3,11,10,18,19,16,13,15,12,5,2 : 19, The final decision tree created for training data is shown in Figure 2.

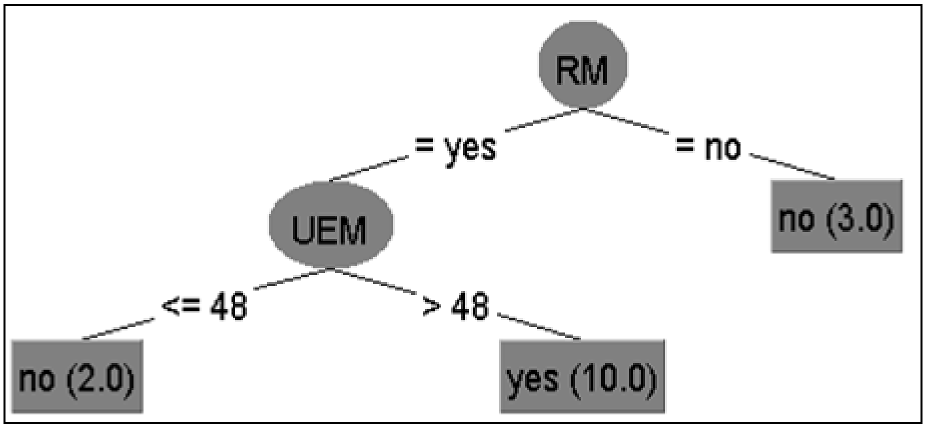


Figure 2 - DECISION TREE FOR TRAINING DATA

5.5 GIVING INPUT FOR THE DECISION TREE

There are two methods for giving input to the decision tree. They are

1. Training data as test data
2. Test data given by the user (unknown data)

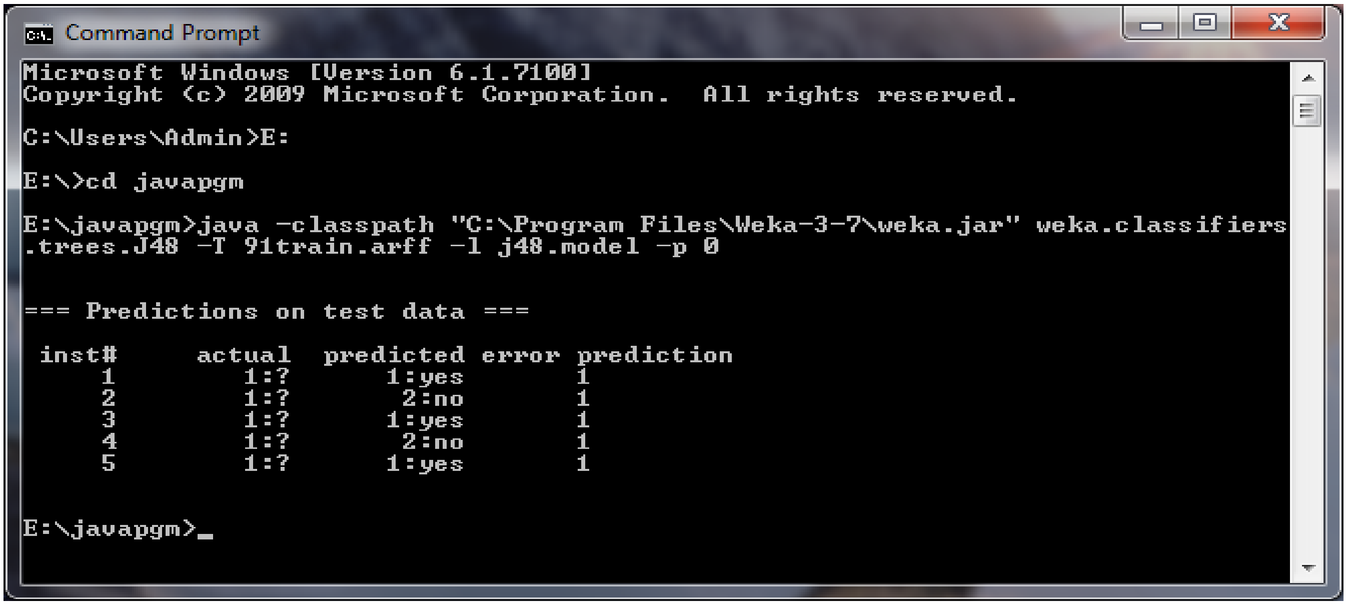
5.5.1 TRAINING DATA AS TEST DATA

There are 5 datasets of the employees are taken for testing. In this type the training data are given as test data to the decision tree and the prediction to be taken. The data to be taken to testing are given in the following table 4.

TABLE 4 – TRAINING DATA AS TEST DATA

SNO	UEM	TP	SOL	RM	UE	IE	PS	AHA	AQ	COE	EOP	DOI	RC	IPC	COI	DW	CEA	HAR	AS	IS
1	90	YES	YES	YES	85	90	YES	2	YES	YES	NO	NO	YES	1	YES	NO	YES	NO	YES	?
2	48	YES	NO	YES	49	45	NO	0	YES	YES	NO	YES	NO	1	YES	YES	NO	YES	YES	?
3	80	YES	YES	YES	90	90	YES	3	YES	YES	YES	YES	NO	1	NO	YES	YES	NO	YES	?
4	45	YES	YES	NO	39	48	NO	0	NO	NO	NO	NO	YES	0	NO	NO	NO	NO	NO	?
5	75	YES	YES	YES	70	80	YES	3	YES	YES	YES	YES	YES	1	YES	NO	YES	YES	NO	?

After applying the above data the output is as follows



```
Microsoft Windows [Version 6.1.7100]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\Admin>E:

E:\>cd javapgm

E:\javapgm>java -classpath "C:\Program Files\Weka-3-7\weka.jar" weka.classifiers
.trees.J48 -T 91train.arff -l j48.model -p 0


=== Predictions on test data ===

 inst#     actual  predicted error prediction
     1        1:?      1:yes       1
     2        1:?       2:no       1
     3        1:?      1:yes       1
     4        1:?       2:no       1
     5        1:?      1:yes       1


E:\javapgm>_
```

Figure 3 - OUTCOME OF TRAINING DATA AS TEST DATA

From the displayed output the “predicted” term is used to decide whether to give increment for the employee or not [5]. For the 5 instances there are 3 yes and 2 no are produced which are similar to the training data. Since the attributes RM is yes and UEM is above 48 for the instances 1,3 and 5 in the Table 4 the final decision is produced as yes by comparing with the decision tree in Figure 3.

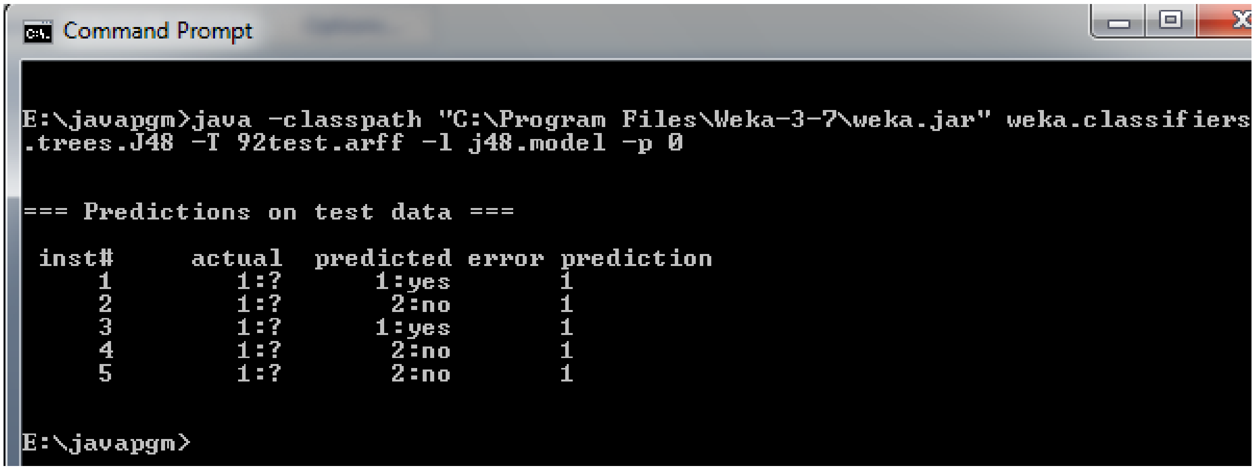
5.5.2 TEST DATA GIVEN BY THE USER (UNKNOWN DATA)

The data to be taken to testing are given in the following table 5

TABLE 5 – UNKNOWN DATA

S.NO	UEM	TP	SOL	RM	UE	IE	PS	AHA	AQ	COE	EOP	DOI	RC	IPC	COI	DW	CEA	HAR	AS	IS
1	90	YES	YES	YES	80	85	YES	3	YES	YES	NO	YES	YES	1	YES	YES	NO	NO	YES	?
2	45	YES	NO	NO	60	65	NO	1	YES	NO	NO	YES	NO	0	YES	NO	YES	YES	NO	?
3	80	YES	NO	YES	70	75	YES	2	NO	YES	YES	YES	YES	1	NO	YES	YES	NO	NO	?
4	60	YES	YES	NO	75	80	NO	3	YES	NO	YES	NO	NO	0	NO	NO	NO	YES	YES	?
5	50	YES	YES	NO	55	60	YES	0	NO	YES	YES	NO	YES	1	YES	YES	NO	YES	NO	?

After applying the above data the output is as follows



```
E:\javapgm>java -classpath "C:\Program Files\Weka-3-7\weka.jar" weka.classifiers
.trees.J48 -T 92test.arff -l j48.model -p 0

=== Predictions on test data ===

 inst#     actual  predicted error prediction
     1        1:?      1:yes       1
     2        1:?       2:no       1
     3        1:?      1:yes       1
     4        1:?       2:no       1
     5        1:?       2:no       1

E:\javapgm>
```

Figure 4 – OUTCOME OF DATA GIVEN BY THE USER

Here the test data are an unknown data. For the given five instances there are 2 yes and 3 no are produced from the term “predicted”. Since the attribute RM is no for the instances 2,4 and 5 in the table 5 the final decision produced is no by comparing the decision tree in Figure 4. This implies that among the five employees only 2 people will be given the increment. This process goes on until all data classified perfectly or run out of attributes. The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules. One classification rules can be generated for each path from each terminal node to root node.

6 CONCLUSION

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce human readable classification rules and easier to interpret than other classification methods. In this paper, the classification task is used in employee database to predict the employee performance on the basis of trained dataset. As there are many approaches that are used for data classification, the decision tree method is used for measuring the performance of an employee in an organization. Information like attendance, paper presented, seminars attended were collected from the employee’s previous record, to predict the performance at the end of the year. This study helps to predict whether to give yearly increment, promotion and career advancement for an employee in an educational institution.

7 FUTURE ENHANCEMENT

Streaming parallel decision tree is designed for large data sets and streaming data, and is executed in a distributed environment. It provides a way to analytically compare the error rate of trees constructed with serial and parallel algorithms without comparing similarities between the trees themselves [16]. The decision tree is used in medical areas and its operations such as identifying and medicaments of a patient [17]. The decision tree architecture is used to provide a platform for a reliable, robust robot navigation system that will fulfill the requirements of navigating in unmodified environments [18].

8 REFERENCES

1. Anand Bahety Department of Computer Science University of Maryland, College Park "Extension and Evaluation of ID3 – Decision Tree Algorithm"
2. Brijesh Kumar, Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students Performance " ," IJACSA International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
3. Ying Liu, et all , "Region-based image retrieval with high-level semantics using decision tree learning," Journal of Pattern Recognition, Vol. 41, No. 8, pp. 2554 – 2570, Aug 2008.
4. Breiman, Friedman, Olshen, and Stone. ― Classification and Regression Trees, Wadsworth, Mezzovico, Switzerland. 1984,
5. Daniel Rodríguez "Making predictions on new data using Weka" University of Alcala.
6. Matthew N.Anyanwu, Sajjan G.Shiva, ―Comparative Analysis of Serial Decision Tree Classification Algorithms, International Journal of Computer Science and Security, volume 3.
7. Mitra S, Acharya T. Data Mining.Multimedia, Soft Computing, and Bioinformatics. John Wiley & Sons, Inc., Hoboken, New Jersey; 2003.
8. Parr Rud, O. Data Mining Cookbook.Modeling Data for Marketing, Risk, and Customer Relationship Management. John Wiley & Sons, Inc.; 2001.
9. Quinlan, J.R. Induction of decision trees. Machine Learning, volume 1. Morgan Kaufmann; 1876. p. 71-96.
10. Quinlan, J.R., (1883), C4.5:Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann.
11. S.Anupama Kumar and Dr. M.N.Vijayalakshmi "Efficiency of Decision Trees in Predicting Student's Academic Performance".
12. S.Anupama Kumar, Dr.M.N.Vijayalakshmi,"A Novel Approach in Data Mining Techniques for Educational Data" , Proc 2011 3rd International Conference on Machine Learning and Computing" (ICMLC 2011) , Singapore, 26th-27th Feb 2011,pp V4-152-154.
13. Samrat Singh, Dr. Vikesh Kumar "Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education".
14. Stuart Russell, Peter Norvig "Artificial Intelligence A Modern Approach" and Amos Storkey "Learning from Data: Decision Trees", School of Informatics ,University of Edinburgh Semester 1, 2004.
15. Weka, University of Waikato, New Zealand,http://www.cs.waikato.ac.nz/ml/weka/
16. Yael Ben-Haim Elad Tom-Tov "A Streaming Parallel Decision Tree Algorithm", IBM Haifa Research Lab, Haifa University Campus.
17. Farhad Soleimanian Gharehchopogh, Peyman Mohammadi, Parvin Hakimi "Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study".
18. Erick Swere and David J Mulvaney, "Robot Navigation Using Decision Trees".

19. Hang Yang, Simon Fong, Guangmin Sun, and Raymond Wong "A Very Fast Decision Tree Algorithm for Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network".