

# Employee's Performance Predication System Using C4.5 Classification Algorithm

Thin Thin Swe<sup>1st</sup> Lecturer (*Author*)  
Faculty of Information Science  
University of Computer Studies  
Patheingyi, Myanmar

Phyu Phyu Swe<sup>2<sup>nd</sup> t</sup> Lecturer (*Author*)  
Faculty of Information Science  
University of Computer Studies  
Patheingyi, Myanmar

## Abstract

*The intelligent decision can be achieved by using classification and prediction methods. Two forms of data analysis are assumed that classification and prediction that can be used to extract the important data classes or to predict future data trends. The performance of the employee prediction is the task that deployed the right person with the right skill in the proper position at the appropriate point of time. This task has many managerial decisions that depend on various factors such as the duration of the service, the working time, the teaching skills and the outcome of the work. Taking managerial decisions on these issues can be a cause for misunderstanding, misunderstandings, discrimination and uncertain decisions. In this paper C4.5 classification algorithm is used to predict a target class.*

**Keywords**—C4.5 algorithm, employee performance, classification algorithms

---

## I. INTRODUCTION

Human Resource (HR) has become one of the major concerns of the management of almost all businesses, including the educational institutions and government bodies, as human resources (HR), whose capacity is the present and future of people's performance. Without this, the right person can be defined as a result of performance management, ensuring proper effectiveness, the process of ensuring leadership continuity in important positions and the development and decision making mechanism for the management of the supply.

In the information and knowledge discovery, Data mining is a beginning and promising field. There is a great deal of confidential knowledge in the information industry. The data mining strategy can be used to access and restore knowledge retrieval to retrieve and restore database jobs. Data mining is a collection of techniques for efficient automatic detection of tokens in large databases. This must be used to make a decision-making process for companies so that a person can execute. Data mining techniques provide a way to find solutions for using data mining capabilities such as classification, regression, prediction, clustering, time series analysis, summary, mapping rules, and sequence discovery.

In classification technique, from an input data set, classification models can be built. It involves finding rules that partition the data into separate groups. The classification input is the training data set whose class labels are previously known. It also explores the training data set and constructs a model based on the class label and intention to allocate a class label to the future unlabeled records. Since the class field is well-known, classification is known as supervised learning. There are many classification models such as decision tree, genetic algorithm, and statistical models and so on. Decision tree is a very popular technique because building a tree does not require any knowledge of the field expert or parameter preparation and is suitable for discovering exploratory knowledge. The classification model can be produced by using rules that are human for reading and interpretation. C4.5 techniques is the member of the decision tree families and it can produce both rulesets and decision tree, and tree building for improving the prediction accuracy.

Data mining is a young and promising field of information and knowledge discovery [1]. It started to be an interest target for information industry, because of the existence of huge data containing large amounts of hidden knowledge. With data mining techniques, such knowledge can be extracted and accessed transforming the databases tasks from storing and retrieval to learning and extracting knowledge.

The improvement of employee selection had been worked by the researchers like Chein and Chen [2] have, by building a model. They used the data mining techniques to predict the performance of newly employees or applicants by reviewing on attributes selected from their CVs, job applications and interviews. Their performance could be predicted to be a base for decision makers to take their decisions about either employing these applicants or not.

In their study [3], the influence of motivation on job performance for state government employees in Malaysia had been tested. A positive relationship between affiliation motivation and job performance is showed. According to the people with higher affiliation motivation and strong interpersonal relationships with colleagues and managers tend to perform much better in their jobs. The authors in [4] had discussed in their paper Human Recourses (HR) system architecture to forecast an applicant's talent based on information filled in the HR application and past experience, using Data Mining techniques. The goal of the paper was to find a way to talent prediction in Malaysian higher institutions. So, they have specified certain factors to be considered as attributes of their system, such as, professional qualification, training and social obligation. Then, several data mining techniques (hybrid) were applied to find the prediction rules. ANN, Decision Tree and Rough Set Theory are examples of the selected techniques.

Decision tree is one of the most used techniques, since it creates the decision tree from the data given using simple equations depending mainly on calculation of the gain ratio, which gives automatically some sort of weights to attributes used, and the researcher can implicitly recognize the most effective attributes on the predicted target. As a result of this technique, a decision tree would be built with classification rules generated from it [1]. Naïve Bayes classifier is another classification technique that is used to predict a target class. It depends in its calculations on probabilities, namely Bayesian theorem. Because of this use, results from this classifier are more accurate and effective, and more sensitive to new data added to the dataset. Karatepe et al. [5] defined the performance of a frontline employee, as his/her productivity comparing with his/her peers.


Nowadays, the applicant's performance have to be forecast in the human resource (HR) system based on the information filled in HR application and previous experience using data mining techniques. In any organization, performance prediction is becoming progressively critical method of approaching HR functions. In fact, performance prediction involves human resource planning that regards process for managing people in organization. Therefore, this paper has been made by applying decision tree C4.5 classification algorithm to predict employees' performance in human resource management (HRM) and it generated the classification rules for the historical HR records and test them on unseen data to calculate accuracy.

## II. C4.5ALGORITHM

C4.5 is an algorithm developed by Ross Quinlan and it can be used to generate a decision tree. Typically, C4.5 appeared by extending the Quinlan's earlier ID3 algorithm. The decision trees produced by C4.5 can be used for classification. It can also be assumed as a statistical classifier. C4.5 is one of the most well-known inductive learning algorithms and it can be used to build decision trees as well as prediction rules. Unlike other classification approach, it can also deal with continuous attributes and null attribute values. For the continuous attributes, they should be discretized first.

TABLE 1 C4.5 ALGORITHM

<b>ALGORITHM I : C4.5 CLASSIFICATION ALGORITHM</b>
Input: Training dataset T; attributes S.  Output: Decision tree Tree. <ol style="list-style-type: none"> <li>1. If T is NULL then</li> <li>2. Return failure</li> <li>3. End if</li> <li>4. If S is Null then</li> <li>5. Return Tree as a single node with most frequent class label in T</li> </ol>



```

6. End if
7. if |S| = 1 then
8. return Tree as a single node S
9. end if
10. set Tree = {}
11. for a ∈ S do
12. set Info(a,T) = 0 and SplitInfo (a,T) = 0
13. compute Entropy (a)
14. for v ∈ E values (a,T) do
15. set Ta,v as the subset of T with attribute a = v
16. Info (a,T) +=  $\frac{|T_{av}|}{|T_a|}$  Entropy (av)
17. SplitInfo(a,T) +=  $\frac{|T_{av}|}{|T_a|} \log \frac{|T_{av}|}{|T_a|}$ 
18. End for
19. Gain(a,T) = Entropy(a) – Info (a,T)
20. GainRatio (a,T) =  $\frac{Gain(a,T)}{SplitInfo(a,T)}$ 
21. End for
22. Set abest = argmax{GainRatio(a,T)}
23. Attach abest into Tree
24. For v ∈ E values (abest, T) do
25. Call C4.5 (Ta,v)
26. End for
27. Return Tree

```

C4.5 tree uses a greedy approach that uses an information theoretic measure, gain ratio to build a decision tree from the training data. The training instances are divided into subsets related to the values of the attribute by choosing an attribute for the root of the tree. If the class labels entropy in the subsets is less than the class labels entropy in the full training set, then information has been gained through splitting on the attribute. C4.5 chooses the attribute that gains the most information to be at the root of the tree. The algorithm is applied recursively to form sub trees, terminating when a given subset contains instances of only one class. The algorithm for C4.5 algorithms is described in Table 1.

C4.5 algorithm uses a pruning procedure which takes out branches that do not contribute to the accuracy and replace them with leaf nodes. It allows attributes' value range into two subsets. Definitely, the best threshold is searched by this algorithm for the maximization the gain ratio criterion. All values above the threshold create the first subset and all other values create the second subset.

### III. EMPLOYEE DATASET

The system is intended to presume the employees' performance in the universities by means of a classification model. In order to collect the required data, the prepared questionnaire including attributes concerned with employees' information is distributed to the employees in the universities. The list of the collected attributes is presented in Table 1.1. This system uses 270 employees' data (90 employees from the Computer University (Yangon) and 95 employees from the Technological University (Yangon) and the rest are from other universities).

The overall system design for training data and testing data are described in Figure 1 is the system flow diagram for the training employees explains all the specified functionalities in a graphic view. Firstly, the user chooses range of testing data from database and the system uses the test data for training. Secondly, the data are taken out of the database. Thirdly, the decision tree and classification rules are generated through by using Algorithm I in Table 1. Finally, the classifier model is made up by the system.

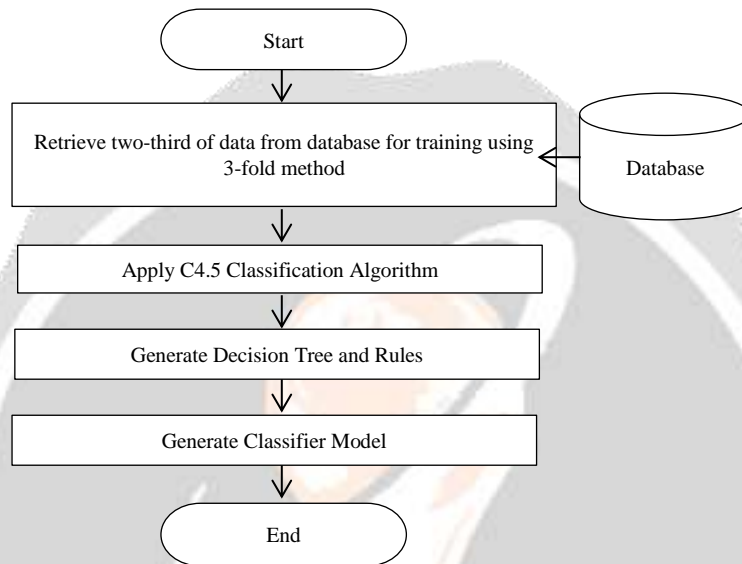


Fig.1.1 System flow diagram for training employee data

In the testing section, the system allows the user to test with two ways. The first way is to test with new data and the second way is to test with existing data. If the user chooses to test with new data, the system will arrange to input attributes values. Otherwise, the user chooses range number of testing data and the system retrieves data from database. Then, the generated classifier model is used to predict the employees' performance is good, fair or poor. Finally, the system determines and displays the performance results to user. The process on the data of testing employees is described in Figure 2.

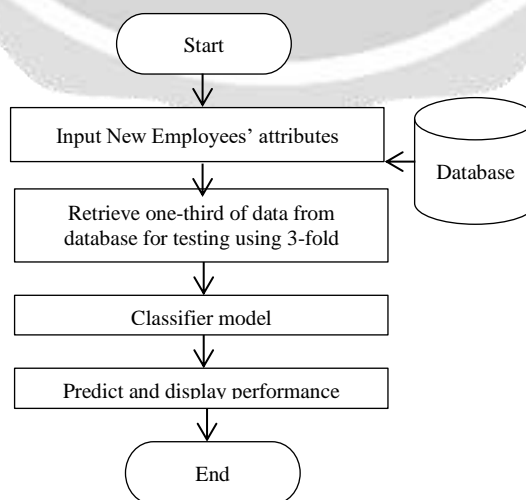


Fig.2. System flow diagram for tesing employee's data

TABLE 2 DESCRIPTION OF ATTRIBUTES USED FOR PREDICTING THE PERFORMANCE CLASS

No	Attribute	Description	Possible Values
1	Age	age	Above 23 years
2	Gender	gender	Male, Female
3	MStatue	Marital status	Single, Married without kids, Married with kids and other
4	Degree	degree	Master, Ph.D.
5	Rank	rank in the current university	Tutor/ Demonstrator. Assistant Lecturer, Lecturer, Associate Professor, Professor
6	Service Period	service period in the university teaches	Above 1 year
7	Subject	Number of subjects that teacher teaches	1,2,3, etc
8	Sociable	social ability, and teamwork with partners	A,B, C (grades of social ability good or fair or poor)
9	Teaching skill	teaching skill	A,B, C (grades of social ability good or fair or poor)
10	Working hours	number of working hours per week	4 hours, 5 hours, etc
11	Activity	activities and contribution	A, B, C (grades of employees' activity good or fair or poor.
12	Work outcome	work outcome	A, B, C (grades of employees' work outcome good or fair or poor.
13	Performance	performance	Good, Fair, Poor

#### IV. IMPLEMENTATION OF THE SYSTEM

The system implementation in this paper includes a two-step processes that are model construction and model usage. Moreover, estimating the accuracy of the model is also included.

The model construction part includes presenting a set of predetermined classes. Each sample is assumed as a predefined class as determined by the class label attribute. Training set is the set of tuple used for model construction. This model is characterized as classification rules, decision tree and mathematical formulae. In the model usage part, the unknown objects are classified and the accuracy of the model is estimated. For the clarity

purpose; implementation details are divided into three parts: (1) Training, (2) Testing and (3) Calculating accuracy.

*A. Training Employees data of the System*

Firstly, two-third part of employees' data for training is regained from database. Later on getting that information, the system starts its functions such as extracting data from database, showing Training employees' database, applying C4.5 algorithm, generating decision tree and rules. Eventually, the system produces the classifier model.

*B. Training Employees data of the system*

User input may consist of (1) personal information such as age, gender and marital status (2) educational information such as degree, (3) professional information such as rank, service period, subject, social able, teaching skill, working hours, activity and work outcome. These attributes are used to predict the employees' performance to be good, fair or poor. After getting that information from the user, the system is to determine the unknown data by means of generated classifier model.

The accuracy section, calculates upon existing employees dataset. So, the system requests the training range and testing range from the user. After getting the training and testing ranges, the system uploads employees data from the database and calculates accuracy. Finally, the system shows accuracy percentage result to user. Accuracy rate is the percentage of correctly classified test set sample by the classifier mode. Typically, training set and testing set are independent each other. If the accuracy is acceptable, the classifier model is used to classify new data.

**V. EXPERIMENTAL RESULTS**

The main purpose of the system is to classify whether the tested employees' data are good, fair or poor. The input of the system is the range of the number of employees' data (one-third for Testing and two-third for training) and the information of user while the output of the system is the classes (good or fair or poor). To calculate the accuracy, this system uses the classifier accuracy equation.

Accuracy is better measured on a test set consisting of class labeled tuples that were not used to train the model. In a performance measure, it measures the number of correct predictions and ignores the number of incorrect predictions of the classifier. To calculate the accuracy, the following Equation 1 is used:

$$\text{Accuracy} = \frac{t}{n * 100}$$

where t is the number of correctly classification and n is the total number of samples.

*C. Analysis of Experimental Result*

The system accuracy is tested the first one-third (1 to 90). The first one-third of the data (1 to 90) for testing sets and the remaining are training sets. The accuracy values are calculated with accuracy formula.

TABLE III. ACCURACY RESULT OF FIRST TESTING

Accuracy Rate	
Training Data	180
Testing Data	50
Number of Rule	32
Number of correct count	85
Accuracy	94%

TABLE IV. CLASSIFICATION RULES FOR FIRST TESTING

Rule	Rule Antecedent	Performance Decision
1	If Activity = A and WorkOutcome = A and Sociable = A →	Good
2	If Activity = A and WorkOutcome = B and Degree = PhD and WokingHours < 17 →	Poor

3	If Activity = B and TeachingSkill = A and WorkOutcome = B →	Fair
4	If Activity = B and TeachingSkill = B and WorkOutcome = B →	Fair
.	.	.
.	.	.
.	.	.
30	If Activity = A and WorkOutcome = C and WorkingHours < 17 →	Poor
31	If Activity = B and TeachingSkill = C and WorkOutcome = C →	Poor
32	If Activity = C →	Poor

According to the Table, it can be noticed that the classification accuracy is much more in first one-third testing (1-90). And, the start node is Activity. The attributes with the highest gain ratio are WorkOutcome, TeachingSkill, and Socialable. Table shows the generated classification rules and decision tree can be seen in the following Figure.

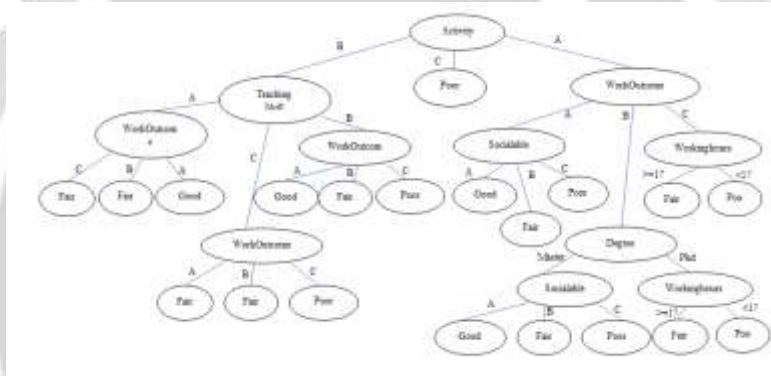


Fig.3 Decision tree for first testing

The system is also tested using the second testing of the one-third (91-180) as the testing set. The built decision tree is to start with Activity. And, the more value attributes are WorkOutcome, Social able and Degree. The accuracy is more slightly increase to 92% in second testing of the one-third (91-180). Accuracy results for second testing are shown in Table 4.4.

TABLE V. ACCURACY RESULTS FOR SECOND TESTING

Accuracy Rate	
Training Data	180
Testing Data	90
Number of rules	36
Number of correct count	83
Accuracy	92%

The system is also tested using the third testing of the one-third (181-270) as the testing set. In this experiment, the prediction accuracy becomes more decreasing in 88%. And, the root node is TeachingSkill. The attributes with the highest gain ratio are WorkingHours, ServicePeriod, Socialable and WorkOutcome. Accuracy results for third testing are shown in Table 4.5 and accuracy results for overall system can be seen in Table 4.6.

TABLE VI. ACCURACY RESULTS FOR THIRD TESTING

Accuracy Rate	
Training Data	180
Testing Data	90

Number of rules	32
Number of correct count	85
Accuracy	94%

TABLE VII ACCURACY RESULT FOR OVERALL SYSTEM

	Number of rules	Number of testing	Number of correct count	Accuracy
First Testing	32	90 (one-third 1 to 90)	85	94%
Second Testing	34	90 (one-third 91 to 180)	83	92%
Third Testing	42	90 (one-third 181 to 270)	80	88%
Average Accuracy				91%

This system attempts to predict the performance of the tested employees' data into the three group: good, fair or poor using C4.5 classification algorithm. According to these experimental results, it can be noticed that how the system can predict correctly the tested employee's data. This system can predict accuracy percentages nearly 91%.

According to the experiments, there are several factors that effect on the performance of employees. TeachingSkill is one of the most effective factors.

Other educational factor degree has considerably affected the performance. Some personal information like age, marital status and gender also touches the performance. Nevertheless, the age has not clear effect on the performance, since sometimes the performance increases with age which adds the experience factor, other times it decreases showing the highest motivation with the younger employees. Marital status, on the other hand, is clear in its effect, since single employees have shown better performance from married employees and even much better than married with kids employees.

To affect the performance, several professional factors may also appear. One of the most positive factors on performance is Activity. Unexpectedly, a strange trend appeared regarding number of working hours which indicated that a higher performance can achieved by the greater number of working hour. The service period attribute has achieved an interesting influence on performance. For the employees, who has the more experience, he may achieve the high performance.

Generally, the senior employees' performance is more than juniors' performance. Finally, work outcome and sociable also have a great effect on performance.

According to the final observation on the accuracy of the classification model built for the three experiments, it can be noticed that the classification accuracy is also changed on the training dataset. The accuracy values and the root node value of the attributes which are varied according to the training data. When the system is tested the first two-third of the data (91-270) for testing sets, the system accuracy is 94% and its root node is Activity. Then, the accuracy will change to 92% if the system executes the second one-third of the data (1-90 and 181 to 270) for testing sets. Moreover, when the system runs on the third one-third of the data (1 to 180) for testing sets, the system accuracy decreases to 88% and the root node of the tree is varied TeachingSkill. So, the accuracy values and the root node values are also depended on the training dataset. According to the experimental result, the generated rules can predict quickly and accurately the talent of employees. Therefore, this system can be applied in every HRM department.

## VI. CONCLUSION

The prediction of the employee data by using C4.5 classification algorithm is presented in this paper. This paper has described the significance of the study on the use of data mining classification technique for employees' performance. In this paper, the greatest potential of C4.5 is observed for to classification. The generated classification rules can help the decision makers to determine their potential employees for promotion or may be for other tasks such as selection new employees, matching people to jobs, planning career paths, planning training needs for new and senior employ, predicting employee performance, predicting future employee.

## REFERENCES

- [1] [1] Han, J., Kamber, M., Jian P. (2011). Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers.



- [2] Chein, C., Chen, L. (2006) "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry", Expert Systems with Applications, In Press.
- [3] Salleh, F., Dzulkifli, Z., Abdullah, W.A. and Yaakob, N. (2011). "The Effect of Motivation on Job Performance of State Government Employees in Malaysia", International Journal of Humanities and Social Science, 1(4), pp. 147-154.
- [4] Jantan, H., Hamdan, A.R. and Othman, Z.A. (2010a) "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application", International Journal of Humanities and Social Science, 5(11), pp. 694-702.
- [5] Karatepe, O.M., Uludag, O., Menevis, I., Hadzimehmedagic, L., Baddar, L. (2006). "The Effects of Selected Individual Characteristics on Frontline Employee Performance and Job Satisfaction", Tourism Management, 27 (2006), pp. 547-560.
- [6] Fan, J., & Wen, P., "Application of C4.5 algorithm in web-based learning assessment system", Sixth International Conference on Machine Learning and Cybernetics. Hong Kong.

