Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 6 No. 4. – www.cisdijournal.net

# Job Opportunity Factors Analysis Using Decision Tree Algorithms

**I.A Ganiyu**
Department of Computer Science
Ramon Adedoyin College of Science & Technology
Oduduwa University
Ipetumodu, Osun State, Nigeria.
E-mail: eedrisooh@ymail.com
Phone: +2348037437877

**I.O. Awoyelu**
Department of Computer Science and Engineering
Obafemi Awolowo University
Ile-Ife, Osun State, Nigeria.
E-mail: olukemiawoyelu@yahoo.com

## ABSTRACT

The causes of highly unemployment rate in the labour market have for a long time been the focus of study among many stakeholders, higher education managers, parents, government and researchers. From studies investigating unemployed applicants and its related problems, it has been observed that job opportunity success is dependent on many factors such as: year of graduation, industrial experience, higher degree, age, grades and professional/relevant skills. This work uses data mining techniques to investigate the category of applicants that are likely to secure a better and promising future based on the criteria used by the employees to select the right applicants for the right jobs. The data set used comprised of two hundred (200) records of employees of one of the communication companies that does massive recruitment. The analysis was carried out using Decision Tree algorithms. Various Decision Tree algorithms were investigated and the algorithm which best models the data was used to generate rule sets are used to analyse the factors contributing to the job opportunity success in the industry. The rules generated can serve as a guide to educational administrators, stake holders, parent and the applicants in their career pursuing activities.

**Key words:** Data Mining, Decision Tree, Artificial Neural Network, Knowledge Discovery..

## 1. INTRODUCTION

Data mining is a process through which valuable knowledge can be extracted from a large database. The necessity for the development of data mining evolved due to the immense and quick growth of the volume of stored corporate data. Ordinary querying methods could no longer produce results showing hidden patterns in such vast amounts of data. Using advanced methods derived from artificial intelligence, pattern recognition and statistics, data mining can construct a comprehensively descriptive model on input data. The data model can be produced in various forms and serves the purpose of describing and predicting behaviour of the data object (Alao and Adeyemo, 2013). Some data mining techniques are artificial neural networks, decision trees, memory-based reasoning, regression analysis, clustering, rule induction and association rules (Bharati, 2010).

Artificial Neural Networks are non-linear, predictive models that learn through training. Although they are powerful predictive modelling techniques. Neural networks were designed to mimic how the brain learns and analyses information. Organizations develop and apply artificial neural networks to predictive analytics in order to create a single framework. Neural networks are ideal for deriving meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques. Decision Trees are tree-shaped structures that represent decision sets. It uses real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process. Decision trees are useful for helping you choose among several courses of action and enable you to explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action. These decisions generate rules, which then are used to classify data. Decision trees are the favoured technique for building understandable models.

Data mining tools predict future trends and behaviours by reading through databases for hidden patterns, they allow organizations to make proactive knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. The overall goal of the data mining process is to extract knowledge from an existing data and transform it into a human-understandable structure for further use.

## 2. RELATED WORKS

According to Singh (2008), at the heart of almost every human resources management program or activity is the need for accurate and thorough job information. Job analysis is thus a prerequisite activity for the effective management of human resources. However, many important assumptions that underlie such fundamental uses of job analysis in management are becoming questionable in today's business environment. Job analysis focuses on the collection of work-related information for the job as it currently exists and/or has existed in the past.

Jayanthi et al. (2008) presented the role of data mining in Human Resource Management Systems (HRMS). A deep understanding of the knowledge hidden in Human Resource (HR) data is vital to a firm's competitive position and organizational decision making. Analyzing the patterns and relationships in HR data is quite rare. The HR data is usually treated to answer queries. Because HR data primarily concerns transactional processing (getting data into the system, recording it for reporting purposes) it is necessary for HRMS to become more concerned with the quantifiable data. The paper demonstrates the ability of data mining in improving the quality of the decision-making process in HRMS and gives propositions regarding whether data-mining capabilities should lead to increased performance to sustain competitive advantage.

Data mining techniques have been applied in many application domains such as banking, fraud detection, network intrusion detection and telecommunications (Hans and Kamber, 2003). Also it has been applied successfully in areas like business, educational, health care management, business, sports etc. In modern world, a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Education, Business, Agriculture and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. (Varun and Anupama , 2011). Data mining methodologies have also being used to enhance and evaluate the higher education tasks. Many researchers have proposed some methods and architectures for using data mining for higher education (Adeyemo and Kuyoro, 2013).

Decision trees are graphical representations of alternative choices that can be made by a business, which enable the decision maker to identify the most suitable option in a particular circumstance. Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top down recursive divide-and-conquer manner. A greedy strategy is usually used because they are efficient and easy to implement, but they usually lead to sub-optimal models.

A bottom-up approach could also be used. The algorithm (Alao and Adeyemo, 2013) is summarized as follows:

1. Create a node N;
2. If samples are all of the same class, C then
3. Return N as a leaf node labeled with the class C;
4. If attribute-list is empty then
5. Return N as a leaf node labeled with the most common class in samples;
6. Select test-attribute, the attribute among attribute-list with the highest information gain;
7. Label node N with test-attribute;
8. For each known value $a_i$ of test-attribute
9. Grow a branch from node N for the condition test attribute = $a_i$;
10. let $s_i$ be the set of samples for which test-attribute= $a_i$;
11. if $s_i$ is empty then
12. attach a leaf labeled with the most common class in samples;
13. else attach the node returned by Generate_decision_tree($s_i$,attribute-list_test-attribute)

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. In data mining, trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Decision trees used in data mining are of two main types, namely: Classification tree analysis and Regression tree analysis. Classification tree analysis is when the predicted outcome is the class to which the data belongs. Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital). One of the most useful characteristics of decision trees is their comprehensibility. People can easily understand why a decision tree classifies an instance as belonging to a specific class. Since a decision tree constitutes a hierarchy of tests, an unknown feature value during classification is usually dealt with by passing the example down all branches of the node where the unknown feature value was detected, and each branch outputs a class distribution. The output is a combination of the different class distributions that sum to 1.

The assumption made in the decision trees is that instances belonging to different classes have different values in at least one of their features. Decision trees tend to perform better when dealing with discrete/categorical features.

## 3. MATERIALS AND METHODOLOGY

This section explains the methodology adopted in this work.

### 3.1 Data Collection

In this study, the data used was collected from Business Process Outsourcing (BPO) - a subset of outsourcing that involves contracting of the operations and responsibilities of a specific business process to a third-party service provider. The Company resides in Ibadan. The data set collected was 300 staff members out of which 200 records were used due to the missing of relevant data from the data set. The variables age, sex, marital status, qualification and professional qualification were selected from the employee records for building the required and targeted features. These variables with their descriptions are specified in Table 1.

**Table 1: Variables with their Descriptions**

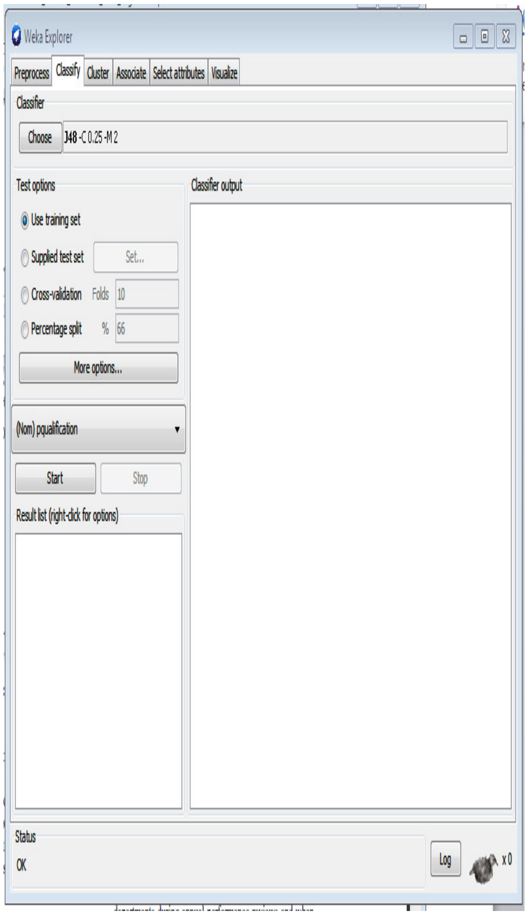
| Field | Description |
|---|---|
| Age | Age of the employee |
| Sex | Gender of the employee. |
| Marital Status | Marital status of the employee i.e. Male or Female. |
| Qualification | Highest qualification of the employee |
| Professional Qualification | This is the number of professional qualification of the employee. |

This work was conducted using WEKA (Waikato Environment for Knowledge Analysis) version 3.7.9 - an open source software developed by a team of researchers of Waikato University and certified by IEEE. WEKA is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The WEKA workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It is written in Java and runs cross platform.

WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. WEKA provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. WEKA allows many algorithms giving room for comparison to determine the better classifier among those used for the study. For the course of this research work, J48 which is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool was considered because C4.5 made a number of improvements to ID3.

## 5. DISCUSSION OF RESULTS

Clicking on Classify tab on the panel open a menu which allow the user to choose the corresponding J48 algorithm.



Classifier: J48 -C 0.25 -M 2

◉ Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) pqualification

Status: OK

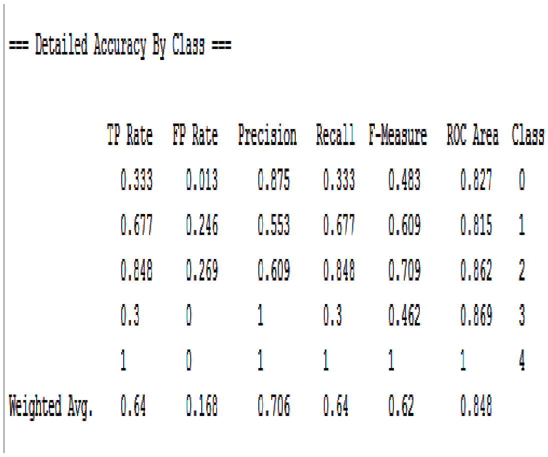
**Figure 1: Classify Tab**

J48 algorithm is chosen because of its best result analysis as compared to other decision tree algorithm present in Weka, The performance metrics used in assessing the performance of the classifier models are: True Positive (TP) Rate and False Positive (FP) rate.

True Positive Rate is the proportion of cases which were classified as the actual class, indicating how much part of the class was correctly captured. It is equivalent to Recall. False Positive Rate is the proportion of cases which were classified as one class but belong to a different class.

Precision is the proportion of the cases which truly have the actual class among all the instances which were classified as the class. F-Measure: is a combined measure for Precision and Recall and is simply calculated as: 2*Precision*Recall/(Precision + Recall). Receiving Operating Characteristic (ROC) is the graphical display of TPR against FPR while AUC represents the area under ROC curve.

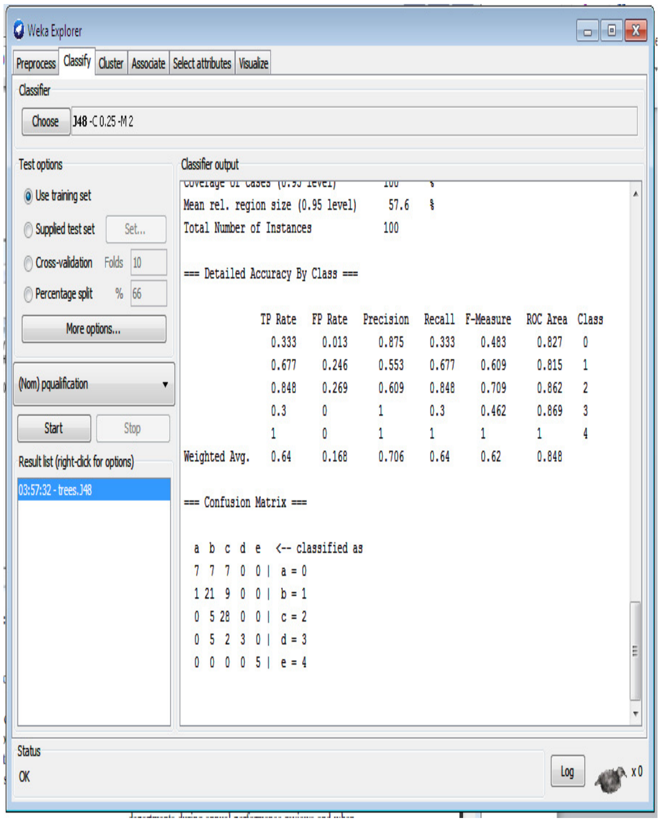
The result of its performance can be seen below



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                 0.333    0.013     0.875    0.333     0.483     0.827    0
                 0.677    0.246     0.553    0.677     0.609     0.815    1
                 0.848    0.269     0.609    0.848     0.709     0.862    2
                 0.3      0         1        0.3       0.462     0.869    3
                 1        0         1        1         1         1        4
Weighted Avg.    0.64     0.168     0.706    0.64      0.62      0.848
```

**Figure 2 Results showing the performance of C4.5 (J48) classifier**



J48 -C 0.25 -M 2

● Use training set

Folds 10

% 66

(Nom) pqualification

03:57:32 - trees.J48

```
Coverage of cases (0.95 level)     100    %
Mean rel. region size (0.95 level)      57.6    %
Total Number of Instances            100


=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                 0.333    0.013     0.875    0.333     0.483     0.827    0
                 0.677    0.246     0.553    0.677     0.609     0.815    1
                 0.848    0.269     0.609    0.848     0.709     0.862    2
                 0.3      0         1        0.3       0.462     0.869    3
                 1        0         1        1         1         1        4
Weighted Avg.    0.64     0.168     0.706    0.64      0.62      0.848


=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  7  7  7  0  0 |  a = 0
  1 21  9  0  0 |  b = 1
  0  5 28  0  0 |  c = 2
  0  5  2  3  0 |  d = 3
  0  0  0  0  5 |  e = 4
```

Status: OK

**Figure 3: Results showing the performance of C4.5 (J48) classifier with Confusion Matrix.**

The confusion matrix is commonly named contingency table. The number of correctly classified instances is the sum of the diagonals in the matrix; all others are incorrectly classified.

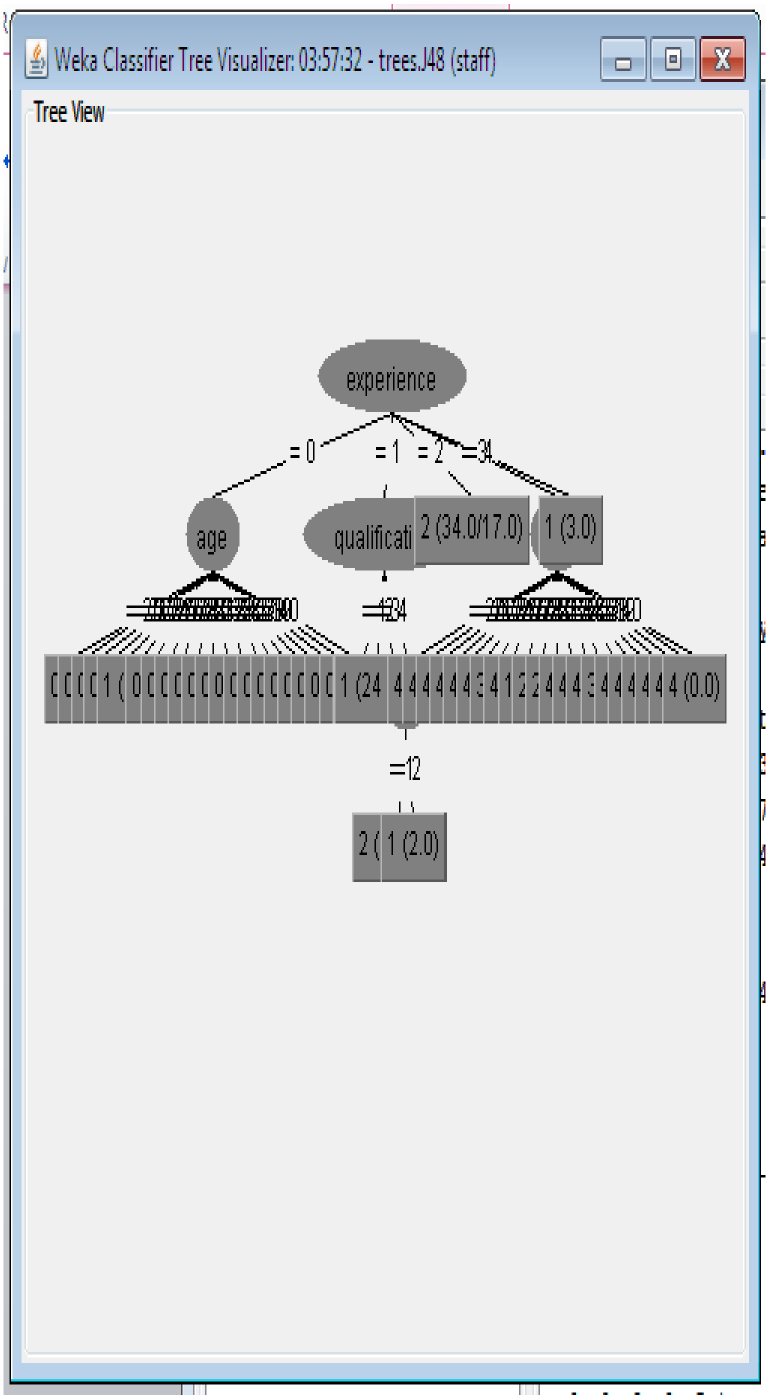


**Figure 4 : Decision tree rules**

## 5. RESULT ANALYSIS

Figure 4 is the decision tree constructed by the J48 classifier. This indicates how the classifier uses the attributes to make a decision. The leaf nodes indicate the outcome of a test, and each leaf (terminal) node holds a class label and the topmost node is the root node (Experience). Many Rules were generated from the decision tree and it can be expressed in English so that we humans can understand them.

**Decision rules**

Since the decision tree can be linearized into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules are of the form:

if condition1 and condition2 and condition3 then outcome.

Comparing the trends in the Market Industry with the rules generated by the classifier, one can be sure that all the rules generated have confidences above 0.500. Most of the rules show that the attribute that is considered most for the job opportunities is ‘Experience’.

The rules show that:

1. Employees with no or zero Experience were considered only by their age, with the view chance as it is shown in the diagram above

2. Employees with Experience ranges from 1-2 were considered only by their Qualification. In the result diagram above, applicant in this group have the most higher chance of getting the job

3. Employees with Experience ranges from 3-4 were also considered only by their Qualification.

## 6. CONCLUSION

We applied data mining techniques to discover knowledge in Education domain and the result of the analysis has shown that student with more Experience and more qualification are considered to be given a job**,** the future work is to deal with more records of students in order to obtain better generalization.

## REFERENCES

1. Alao, D, and Adeyemo, A. B. (2013). ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS. Computing, Information Systems & Development Informatics Vol. 4 No. 1 March, 2013
2. Parbudyal Singh (2008). Job analysis for a changing workplace. Human Resource Management Review 18 (2008) 87–99.
3. Jayanthi, R., Goyal, D.P., Ahson, S.I. (2008). Data Mining Techniques for Better Decisions in Human Resource Management Systems. International Journal of Business Information Systems, 3(5) 464 – 481.
4. Hamidah J., AbdulRazak H., and Zulaiha A. O. (2011). Towards Applying Data Mining
5. Techniques for Talent Managements, 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press.
6. Wikipedia (2014).
7. Adeyemo, A.B. and Kuyoro, S. O. (2013). Investigating the Effect of Students Socio-Economic/Family Background on Students Academic Performance in Tertiary Institutions using Decision Tree Algorithm. Journal of Physical and Life Science actaSATECH 4(2): 61 - 78 (2013).
8. Varun, K, and Anupama, C, (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. International Journal of Advanced Computer Science and Applications, (IJACSAVol. 2, No.3, March 2011) http://ijacsa.thesai.org/
9. Veeramuthu, P, and Periasamy, R, (2014). Application of Higher Education System for Predicting Student Using Data mining Techniques. International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 5 (June 2014) http://ijirae.com
10. Brijesh K. B & Saurabh P, (2011) Mining Educational Data to Analyze Students‟ Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
11. Exforsys Inc., (2006). How Data Mining is Evolving.