ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS

Alao D. & Adeyemo A. B.
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
sesanadeyemo@gmail.com

ABSTRACT

Employee turnover is a serious concern in knowledge based organizations. When employees leave an organization, they carry with them invaluable tacit knowledge which is often the source of competitive advantage for the business. In order for an organization to continually have a higher competitive advantage over its competition, it should make it a duty to minimize employee attrition. This study identifies employee related attributes that contribute to the prediction of employees' attrition in organizations. Three hundred and nine (309) complete records of employees of one of the Higher Institutions in Nigeria who worked in and left the institution between 1978 and 2006 were used for the study. The demographic and job related records of the employee were the main data which were used to classify the employee into some predefined attrition classes. Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows were used to generate decision tree models and rule-sets. The results of the decision tree models and rule-sets generated were then used for developing a a predictive model that was used to predict new cases of employee attrition. A framework for a software tool that can implement the rules generated in this study was also proposed.

Keywords: Employee Attrition, Decision Tree Analysis, Data Mining

1. INTRODUCTION

The Barron's Business dictionary defined attrition as the normal and uncontrollable reduction of a work force because of retirement, death, sickness, and relocation. It is one method of reducing the size of a work force without the management taking any overt actions. The drawback to reduction by attrition is that reductions are often unpredictable and can leave gaps in an organization. Generally attrition is the reduction or loss of employees through different conditions. If organizations know why their employees are likely to leave, they can develop effective policies and strategies for employee retention. Most employees make a number of transitions between jobs during their working lives.

These may include job changes within a single employer and leaving one firm to take a job in another firm. In either case, there is usually the intention to grow and increase in skills, responsibility, and remuneration, and/or improve the "fit" between employee skills and desires and job requirements (Fisher, 2004). The loss of an organization's employee can be divided into three broad groups, induction crises, natural wastage, and retirement (Bennisonn and Casson, 1984). High turnover often means that employees are unhappy with the work or compensation, but it can also indicate unsafe or unhealthy conditions, or that too few employees give unsatisfactory performance (due to unrealistic expectations, inappropriate processes or tools, or poor candidate screening). The lack of career opportunities, challenges and dissatisfaction with the job-scope or conflict with the management have been cited as predictors of high job turnover (Dijkstra, 2008).

A high level of labour turnover can also be caused by many factors such as: inadequate wage levels leading to employees moving to competitors, poor morale and low levels of motivation within the workforce, recruiting and selecting the wrong employees in the first place, meaning they leave to seek more suitable employment, A buoyant local labour market offering more (and perhaps more attractive) opportunities to employee, Poor organization and lack of development. There are two types of employee turnover, namely; *voluntary* and *involuntary*. Voluntary turnover is initiated by the employee; for example, a worker quits and takes another job. Involuntary turnover is initiated by the organization; for instance, a company dismisses an employee due to poor performance or an organizational restructuring. Also in the case of involuntary turnover the employee turnover is not controllable; such as retirement, dismissal or death (Allen., 2008; Igbaria,1991).

There are also two basic types of involuntary termination, known often as being "fired" and "laid off." To be fired, as opposed to being laid off, is generally thought of to be the employee's fault, and therefore is considered in most cases to be dishonorable and a sign of failure. Typically, the characteristics of employees who engage in involuntary turnover are no different from job stayers. However, voluntary turnover can be predicted (and in turn, controlled) by the construct of turnover intent. Voluntary turnover is the most important issue that industries should think about.

Another important distinction is between *functional* and *dysfunctional* voluntary turnover. Dysfunctional turnover is harmful to the organization and can take numerous forms, including the exit of high performers and employees with hard-to-replace skills, departures of women or minority group members that erode the diversity of a company's workforce, and turnover rates that lead to high replacement costs. By contrast, functional turnover does not hurt an organization. Examples of this type of turnover include the exit of poor performers or employees whose talents are easy to replace. All organizations collect data about their employees. However, the actions taken with that data varies widely among organizations (Nagadevara, et al, 2008).

Within a company, much data is available to help develop an effective retention management plan. To create a sound plan, there is a need to determine the extent to which turnover is a problem in the firm, diagnose turnover drivers, and formulate retention strategies. Additionally, the Human Resource (HR) management systems have records of all job status changes, including voluntary terminations, employee's job-action history, the length of time in a position, and salary history. This large collection of employee data within an organization, especially that possessed by the Human Resource (HR) arm of the organization can be analyzed for the effective prediction of employee attrition. Data mining can be helpful to human-resources departments in identifying the characteristics of their most successful employees, most especially aid in figuring out employee with high attrition (turnover) potentials. Information obtained, such as universities attended by highly successful employees, can help Human Resource (HR) departments focus recruiting efforts accordingly.

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 behavior of the data object.

The difference between Data Mining and statistics is that Data Mining automates the statistical process required in several tools. Statistical inference is assumption driven in the sense that a hypothesis is formed and tested against data. Data Mining, in contrast is discovery driven. That is, the hypothesis is automatically extracted from the given data. The other reason is Data Mining techniques tend to be more robust for real-world messy data and also used less by expert users (Berson et al., 1999). Data mining tools predict future trends and behaviors 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.

Some data mining techniques are (Bharati, 2010):

i. **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 analyzes 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.

ii. **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.

iii. **Memory Based Reasoning (MBR)/Case Based Reasoning.** This technique has results similar to a neural network's but goes about it differently. MBR looks for "neighbor" kind of data rather than patterns. It solves new problems based on the solutions of similar past problems. MBR is an empirical classification method and operates by comparing new unclassified records with known examples and patterns.

iv. **Regression analysis**. Regression models are the mainstay of predictive analytics. The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. That relationship is expressed as an equation that predicts the response variable as a linear function of the parameters.

v. **Rule induction**. Rule induction involves developing formal rules that are extracted from a set of observations. The rules extracted may represent a scientific model of the data or local patterns in the data.

vi. **Clustering** is the identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.

vii. **Association rule and correlation** is usually to find frequent item sets 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 dataset is generally very large and a high proportion of the rules are usually of little (if any) value**.** (Bharati 2010)

2. RELATED WORKS

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. They show how data mining discovers and extracts useful patterns from this large data set to find observable patterns in HR. 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.

Hamidah et al (2011), in their work described the background of data mining, data mining in human resource application and an overview of talent management. Their literature study reveals that most researchers have discussed HR applications from different type of application. However, there should be more HR applications and Data Mining techniques applied to different problem domains in HRM field research in order to broaden our horizon of academic and practice work on HR applications using Data Mining techniques.

Due to these reasons, they proposed the suitable Data Mining techniques for performance prediction based on initial experiment. They suggested for future work that the data in HR can be tested using other Data Mining techniques to find out the best accuracy of the techniques, especially for talent management data. Besides that, the relevance of attributes should be considered as a factor to the accuracy of the classifier. It was also suggested that in future experiment, attribute reduction experiment should take place in order to choose the relevant attributes for each of the factor. Once the relevant attributes are attained, the next modeling steps can be established to recommend. Finally, the ability to continuously change and obtain new understanding is the power of HR application, and this can be the HR applications of future work.

Nagadevara et al, (2008), explored the relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover in a rapidly growing sector like the Indian software industry. The unique aspect of this research was the use of five predictive data mining techniques (artificial neural networks, logistic regression, classification and regression trees, classification trees (C5.0), and discriminant analysis) on a sample data of 150 employees in a large software organization. The results of the study clearly show a relationship between withdrawal behaviors and employee turnover. This study raised several issues for future research. First, further research could explicitly collect data on demographic variables across a large sample of organizations to examine the relationship between demographic variables and turnover. Second, large scale data on variables in the past academic research which have a relationship with turnover can be collected longitudinally. Such a data set will allow for more rigorous analysis and also a refined prediction model. Third, the context specific variables of employee turnover which emerged from this study would warrant a deeper understanding of the phenomena. There is a need for more empirical research and in particular, longitudinal research using data within corporations to refine the model. Last, more research needs to be conducted on various samples to confirm the validation of the theoretical model and the prediction model proposed in the study.

Wei-Chiang and Ruey-Ming (2007), in their work explored the feasibility of applying the *Logit* and *Probit* models, which have been successfully applied to solve nonlinear classification and regression problems, to employee voluntary turnover predictions. A numerical example involving voluntary turnover data of 150 professional employees drawn from a motor marketing enterprise in central Taiwan was used with a usable sample size of 132. The data set was divided into two parts, the modeling data set and the testing data set. The modeling data set was used to test the *logit* and *probit* models.

The testing data set was not used for either model building or selection, and was used for estimating model performance when applied to future data. The empirical results of their investigation revealed that the proposed models have high prediction capabilities and that the two (*logit* and *probit) models* also provide a promising alternative for predicting employee turnover in human resource management. The authors suggested that turnover research should move in new directions based on new assumptions and methodologies, which would raise new issues and problems (such as the use of neural networks and support vector machines to conduct classification problem for detecting stayer or leaver).

In a dissertation by Marjorie Laura Kane-Sellers (2007), the researchers carried out a study to explore the variables impacting employee voluntary turnover in the North American professional sales force of a Fortune 500 industrial manufacturing firm. By studying VTO (Voluntary Turn Over), the intention was to gain a better understanding of HRD (Human Resource Development) interventions that could improve employee retention. The focal firm provided observations of the employee database for all members of the professional technical sales force over a 14-years longitudinal period. The original database conveyed 21,271 discrete observations identified by unique employee clock number.

The study design combined descriptive, correlation, factor analysis, multiple linear regression, and logistic regression analysis techniques to examine relationships, as well as provide some predictive characteristics among the variables. Initially, descriptive statistical techniques were used to develop baseline turnover rates, retention rates, and years of tenure. The mean tenure for the population as well as for each ethnic, gender, assignment location, supervisor, educational level, and sales training participation group was calculated. Hierarchical descriptive techniques also provided the mean salary by job title, ethnicity, gender, educational level, and sales training participation. In this study, data-mining analysis commenced with descriptive analysis techniques.

This step facilitated an understanding of the scope of the problem as well as the characteristics of the dataset. The results of the descriptive analysis provided insight into missing data as well as cell size of the subgroups contained in the population. Exploratory factor analysis techniques were used in order to understand co-variance between variables and to develop valid constructs. With the groups determined (VTO versus non-VTO, Trained versus Untrained, Caucasian versus non-Caucasian), Analyses of Variance (ANOVA) were conducted to examine the difference between and within the various dichotomous groups. The final step involved binomial logit regression in order to test models used to predict an employee's likelihood to maintain organizational membership under different conditions. The Education sector is one of the vital sectors for any country, performing a number of roles in the economy. It was therefore chosen for this study. Like all other organizations the education sector in Nigeria is also facing the same employee turnover problem. It plays an important role in our economy. Such institutions play a pivotal role in stimulating the level of industrialization, poverty alleviation and human development. And a healthy academic system depends on the performance of sound personnel (employees). In this work predictive data mining models (decision tree algorithms) were used to generate rule-sets that can be used to help recognize employee's with high probability of attrition in the nearest future.

3. MATERIALS AND METHODS

3.1 Decision Trees

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 strategies 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 is summarized as follows:

```
create a node N;
if samples are all of the same class, C then
    return N as a leaf node labeled with the class C;
if attribute-list is empty then
    return N as a leaf node labeled with the most common class in samples;
select test-attribute, the attribute among attribute-list with the highest information gain;
label node N with test-attribute;
for each known value ai of test-attribute
    grow a branch from node N for the condition test-attribute= ai;
    let si be the set of samples for which test-attribute= ai;
    if si is empty then
        attach a leaf labeled with the most common class in samples;
    else attach the node returned by Generate_decision_tree(si,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, categorisation and generalisation of a given set of data. Data comes in records of the form:

$(x, Y) = (x_1, x_2, x_3, ……., x_k, Y)$ …………(1)

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalise. The vector **x** is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.

Decision trees used in data mining are of two main types:

- 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).

The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman *et al*. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split. There are many specific decision-tree algorithms. Notable ones include:

- ID3 algorithm
- C4.5 algorithm
- C5.0 algorithm
- CHi-squared Automatic Interaction Detector (CHAID). Performs multi-level splits when computing classification trees

The most common types of decision tree algorithm CHAID, CART and C4.5. CHAID (Chi-square automatic interaction detection) and CART (Classification and Regression Trees) were developed by statisticians. CHAID can produce tree with multiple sub-nodes for each split. CART requires less data preparation than CHAID, but produces only two-way splits. C4.5 comes from the world of Machine Learning, and is based on information theory. The most well-know algorithm in the literature for building decision trees is the C4.5 (Quinlan, 1993). C4.5 is an extension of Quinlan's earlier ID3 algorithm (Quinlan, 1979). One of the latest studies that compare decision trees and other learning algorithms has been done by (Tjen-Sien Lim et al. 2000).The study shows that C4.5 has a very good combination of error rate and speed.

C4.5 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 $S = s_1, s_2, \cdots$, of already classified samples. Each sample $s_i = x_1, x_2, \cdots$, is a vector where $x_1, x_2, \cdots$, represent attributes or features of the sample.

The training data is augmented with a vector $C = c_1, c_2, \ldots$,

Where $c_1, c_2, \cdots$, represent the class to which each sample belongs.

At each node of the tree, C4.5 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 C4.5 algorithm then recurses on the smaller sublists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

3.2 Pseudocode

In pseudocode, the general algorithm for building decision trees is (Kotsianti, 2007):

```
1. Check for base cases
2. For each attribute a
       Find the normalized information gain from splitting on a
3. Let a_best be the attribute with the highest normalized information gain
4. Create a decision node that splits on a_best
5. Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node
```

Amongst other data mining methods, decision trees have various advantages:

- **Simple to understand and interpret.** People are able to understand decision tree models after a brief explanation.
- **Requires little data preparation.** Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- **Able to handle both numerical and categorical data.** Other techniques are usually specialised in analysing datasets that have only one type of variable. Ex: relation rules can be used only with nominal variables while Neural Networks can be used only with numerical variables.
- **Uses a white box model.** If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. An example of a black box model is an artificial neural network since the explanation for the results is difficult to understand.
- **Possible to validate a model using statistical tests.** That makes it possible to account for the reliability of the model.
- **Robust.** Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- **Performs well with large data in a short time.** Large amounts of data can be analysed using standard computing resources.

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.

3.3 Data collection

The data used for this research work was collected from the personnel records of employees in one of the Higher Institutions in South-West Nigeria. The data set was composed of employee records of 309 staff members of the Institution extracted from the total number of staff records which was 4326. The records extracted covers a period of 28 years from 1st of September1978 to 30th June 2006. The following variables were selected from the employee records for building the required and targeted features:

a. Unique serial number of each staff member
b. Sex
c. Date of Birth
d. State of Origin
e. Grade Level/ Step
f. Date of 1st Appointment
g. Date Employee Left
h. Salary per Annum
i. Reason for Leaving

3.4 Data Preprocessing

There were some missing values discovered in the working data, these were deleted from the dataset. Also, while trying to load the record from the original Microsoft Excel format it was collected, the date format of (dd/mm/yyyy) and the amount format of (0,000,000.00) had to be converted to (yyyy) and (0000000) formats respectively to suit use by the data mining softwares. The entire record was then converted into CSV (Comma Separated Values) for easy loading unto the Data mining softwares.

The following set of 6 employee attrition related variables were derived in order to segment the employees based on their demographic and job related variables:

1. **Sex:** The gender of the employee.
2. **State of Origin:** The state of origin of the employee.
3. **Length of Service:** the duration of time in years that the employee has worked for in the Institution. This was

calculated by subtracting the employee's year of 1st appointment from the year employee left the institution.

4. **Rank:** The official title of the category the employee belongs to in the Institution. This was derived from the grade level and step of the employee.
5. **Salary (Per Annum):** Income of employee per annum.
6. **Reasons for Leaving:** The reason why an employee left the organization.

3.5 Development of Employee Prediction Model

The tools used for the data mining stage were WEKA 3.6.7 and See5 for windows. WEKA (Waikato Environment for Knowledge Analysis) 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.

See5 for Windows is a sophisticated data mining tool for the discovery of patterns that delineate categories, assemble them into classifiers and use them to make predictions. It is an improvement on Quinlan's ID3 and C4.5 algorithms. See5 classifiers are expressed as decision trees/Seetrees or set of **if-then** rules forms that are generally easy to understand. See5 for Windows can only generate trees based on C5 algorithm while WEKA allows many algorithms giving room for comparison to determine the better classifier among those used for the study.

3.6 Data Preparation

The data was prepared, pre-processed and cleansed using the *preprocess* tab of the Explorer window of the WEKA GUI Chooser. WEKA's preprocessing capability is encapsulated in an extensive set of routines known as *filters.* Filters allow for data to be preprocesses based on either the instance or attribute values. The missing values in the data were replaced by WEKA using the ReplaceMissingValue filter to substitute the global mean or mode of the training dataset for the missing values before the model is built. Table 1 shows the employee details consisting of the Variable type, Unique, Missing, Distribution and Statistics (minimum, maximum, mean and standard deviation) values of the attributes. Figure 1 shows the Employee Dataset Distribution Chart.

Table 1: Employee Dataset Pre-processing Details

EMPLOYEE DETAILS								
Attributes	Variable Type	Unique	Missing	Distribution	Statistic			
					Min	Max	Mean	StdDev
Sex	Nominal	0%	0%	2	85	224		
State Of Origin	Nominal	1%	0%	13	1	214		
Length Of Service	Numeric	2%	0%	36	0	42	11.725	9.227
Rank	Nominal	0%	0%	4	5	192		
Salary (Per Annum)	Numeric	15%	0%	88	61421	603675	141767.162	124046.195
Reasons	Nominal	0%	0%	9	2	140		

Building the Predictive Model

Classification techniques were used to develop the prediction models used in the study. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

WEKA supports many classifiers from which the following classifiers were selected for the study namely: the C4.5 (J48), REPTree and CART (SimpleCart) decision tree algorithms. Attributes were ranked by significance using information gain by carrying out an Attribute Importance analysis. The widely used F-measure and the AUC probability tree learning measures were used as evaluation metrics for the classifier models generated.

With default values, See5 constructs just a single decision tree, while the rule set option generates rulesets made up of unordered collection of relatively simple if-then rules. The rule sets are a lot easier to interpret than the generated decision tree since a rule set generated from a tree usually has fewer rules than the tree has leaves. The boost option constructs classifiers corresponding to a user specified number of trials. For the purpose of this study, the Single Seetree, Rulesets and Boost Seetree were generated.

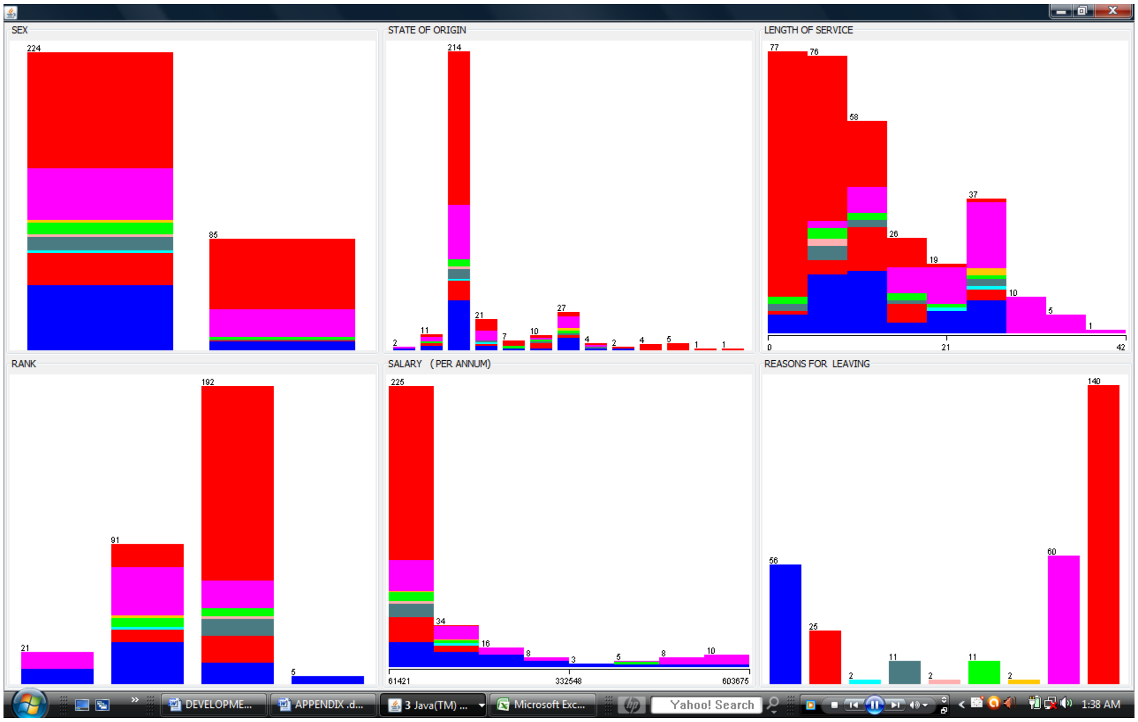


Figure 1: Employee Dataset Distribution Chart

4. RESULT AND DISCUSSION

In carrying out the analyzes the variables were code as follows:

A = DISCIPLINARY CASES

B = JUNIOR NON-ACADEMIC STAFF ABOVE 50 YRS

C = SENIOR NON-ACADEMIC STAFF ABOVE 50 YRS

D = RELOCATION

E = ILL-HEALTH

F = UNSATISFACTORY PERFORMANCE

G = NO REQUIRED QUALIFICATION FOR ADVANCEMEN

H = VOLUNTARY RETIREMENT

I = OUT - SOURCING

The performance metrics used in accessing the performance of the classifier models are:

- **TP Rate (True Positive Rate)**: 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** which is the diagonal element divided by the sum over the relevant row in the confusion matrix. For the C4.5 (J48) algorithm TP Rate will be 27/(27+4+2+1+14+8)= 0.482 for Class A and 129/(129+1+6+1+1+2)= 0.921 for Class I.
- **FP Rate (False Positive Rate)**: the proportion of cases which were classified as one class but belong to a different class. FP Rate is calculated as the column sum of the class, minus the diagonal element, divided by the rows sum of all other classes in the confusion matrix. That is for C4.5 (J48) FP Rate is (44-27)/(253)= 0.067 for Class A
- **Precision:** the proportion of the cases which truly have the actual class among all the instances which were classified as the class. To calculate precision, the diagonal element is divided by the sum over the relevant column in the confusion Matrix. For C4.5 (J48) the precision is 27/ (27+4+1+1+2+1+7+1) = 0.614 for Class A.
- **F-Measure:** is a combined measure for Precision and Recall and is simply calculated as 2*Precision*Recall/(Precision + Recall).
- **Receiving Operating Characteristic (ROC):** the graphical display of TPR against FPR while **AUC** represents the area under ROC curve.

Model Comparison

Table 1: WEKA Classifier Comparison Using F-Measure and AUC

CLASS	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC AREA
C4.5 (J48)	0.67	0.143	0.613	0.67	0.636	0.784
REPTree	0.618	0.156	0.553	0.618	0.579	0.749
CART (SimpleCart)	0.641	0.137	0.579	0.641	0.608	0.777

From the results shown in table 1, the C4.5 (J48) classifier performed better than the other two classifiers implemented in WEKA that were used for this study.

Table 2: Comparing classifier performance using classification accuracy rate

CLASSIFIER	C4.5 (J48)	REPTree	CART (SimpleCart)	JRip	SeeTree	SeeRule	Boost SeeTree
ACCURACY	0.67	0.62	0.64	0.61	0.74	0.73	0.74

A comparison of the results obtained using the various classifiers used for this study is presented in table 2. The results show that the Boosted SeeTree and the Boost SeeRule both performed best with a 0.74 accuracy, followed by the C4.5 (J48) and the CART (SimpleCart) classifier. The REPTree gave the least accuracy of 0.62. Considering the JRip and SeeRule rule sets generated, the SeeRule performed better than JRip with an accuracy of 0.73.

JRIP (Repeated Incremental Pruning to Produce Error Reduction (RIPPER) Rule Learner) Rule Set

The JRip rule leaner generated 4 rules, with the longest rule consisting of 3 attributes and the shortest rule has 1 attribute.

The Default rule is `REASONS FOR LEAVING = OUT - SOURCING.`

The other rules generated by JRip using WEKA are:

```
(STATE OF ORIGIN = DELTA) and (RANK = JUNIOR NON-ACADEMIC STAFF) and (LENGTH OF SERVICE >= 11) => REASONS FOR LEAVING=JUNIOR NON-ACADEMIC STAFF ABOVE 50 YRS (4.0/0.0)

(SALARY ( PER ANNUM) >= 91847) and (LENGTH OF SERVICE <= 11) => REASONS FOR LEAVING=DISCIPLINARY CASES (35.0/9.0)

(LENGTH OF SERVICE >= 20) => REASONS FOR LEAVING=VOLUNTARY RETIREMENT (69.0/26.0)
```

See5 (C5) Rule Set

Each rule generated by See5 has the following characteristics:

- Rule Number: for unique identification of each rule
- Statistics (*n*,lift *x*) or (*n/m*, lift *x*): this summarizes the performance of the rule, where *n* is the number of training cases covered by the rule and *m* is the number of cases that do not belong to the class predicted by the rule. The lift *x* is the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set.
- One or more conditions that must all be satisfied for the rule to be applicable.
- A class predicted by the rule.
- A value between 0 and 1 that indicates the confidence with which the prediction is made.
- The default class used when none of the rules apply.

The rules generated by the See5 (C5) algorithm are presented:

```
Rule 1: (12/3, lift 8.8)
       SEX = M
       LENGTH OF SERVICE > 10
       LENGTH OF SERVICE <= 18
       RANK = JUNIOR NON-ACADEMIC STAFF
       SALARY > 89887
       ->  class JUNIOR NON-ACADEMIC
STAFF ABOVE FIFTY YRS  [0.714]

Rule 2: (24/3, lift 4.7)
       LENGTH OF SERVICE <= 10
       SALARY > 89887
       ->  class DISCIPLINARY CASES
[0.846]

Rule 3: (22/3, lift 4.6)
       LENGTH OF SERVICE <= 18
       SALARY > 159614
       ->  class DISCIPLINARY CASES
[0.833]

Rule 4: (2, lift 4.1)
       SALARY > 82857
       SALARY <= 83072
       ->  class DISCIPLINARY CASES
[0.750]

Rule 5: (4/1, lift 3.7)
       SEX = M
       SALARY > 103547
       SALARY <= 108549
       ->  class DISCIPLINARY CASES
[0.667]

Rule 6: (144/24, lift 1.8)
       SALARY <= 82857
       ->  class OUT - SOURCING  [0.829]

Rule 7: (215/77, lift 1.4)
       LENGTH OF SERVICE <= 18
       SALARY <= 159614
       ->  class OUT - SOURCING  [0.641]

Rule 8: (2, lift 3.9)
       SEX = M
       LENGTH OF SERVICE > 12
       SALARY > 108549
       SALARY <= 109397

       ->  class VOLUNTARY RETIREMENT
[0.750]

Rule 9: (25/8, lift 3.4)
       SEX = F
       LENGTH OF SERVICE > 10
       SALARY > 89887
       ->  class VOLUNTARY RETIREMENT
[0.667]

Rule 10: (69/26, lift 3.2)
       LENGTH OF SERVICE > 18
       SALARY > 89887
       ->  class VOLUNTARY RETIREMENT
[0.620]

Default class: OUT - SOURCING
```

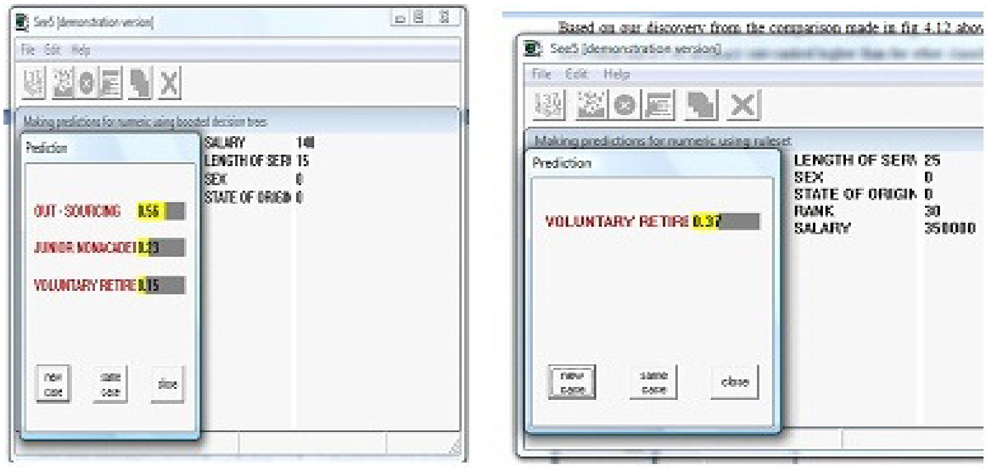
A total of 10 rules were generated by See5 from the 309 training cases. All of the rules generated have confidences above 0.500. Most of the rules show that employees who were on lower ranks, had stayed for less than 20 years and received lesser salaries or employees who had spent longer years in service, but were still on lower ranks and still received lesser salaries left due to OUT-SOURCING and DISCIPLINARY CASES respectively.

5. RESULT ANALYSIS

The See5 single decision tree generated a decision tree of size 15 with 3 subtrees with a misclassification of 25.2%. The attribute usage result shows that Salary attribute had a 100% usage, the Length of Service attribute had 49% usage, 16% usage for the Sex attributes and 16% for Rank. This shows that the Salary earned by an employee and the Length of service rendered by an employee were the prime factors that contributed to an employee either staying or leaving an organization. The boosted decision tree generated 4 trials of the single decision tree and came up with trees with sizes ranging between 5 and 15. This resulted in reduced error rate. The attribute usage of the boosted decision tree shows that all the attributes except the State of origin attribute are of importance in predicting employee attrition with the Length of Service and Salary attributes having a 100% usage and the rank attribute having the least usage of 19%. The results therefore indicate that employee demographic and job related attributes as important factors that affect employee turnover within an organization. The most important attributes were the Salary and Length of service of employee.

5.1 Employee Attrition Prediction System Architecture

The See5 Decision Tree (and rules) with an accuracy level of 0.74 gave the better result compared with the other decision tree algorithms and rule learner. Hence it can be used for the prediction of previously unseen cases. Using the see5 software interactive interface (figure 2) the rules were used to predict the likelihood of employees leaving the institution based on the rules generated from the historical data used.



Making predictions for numeric using boosted decision trees

OUT - SOURCING 0.56
JUNIOR NONACADE 0.23
VOLUNTARY RETIRE 0.15

SALARY 140
LENGTH OF SER 15
SEX 0
STATE OF ORIGIN 0

Based on our discovery from the comparison made in fig 4.12 abov

Making predictions for numeric using ruleset

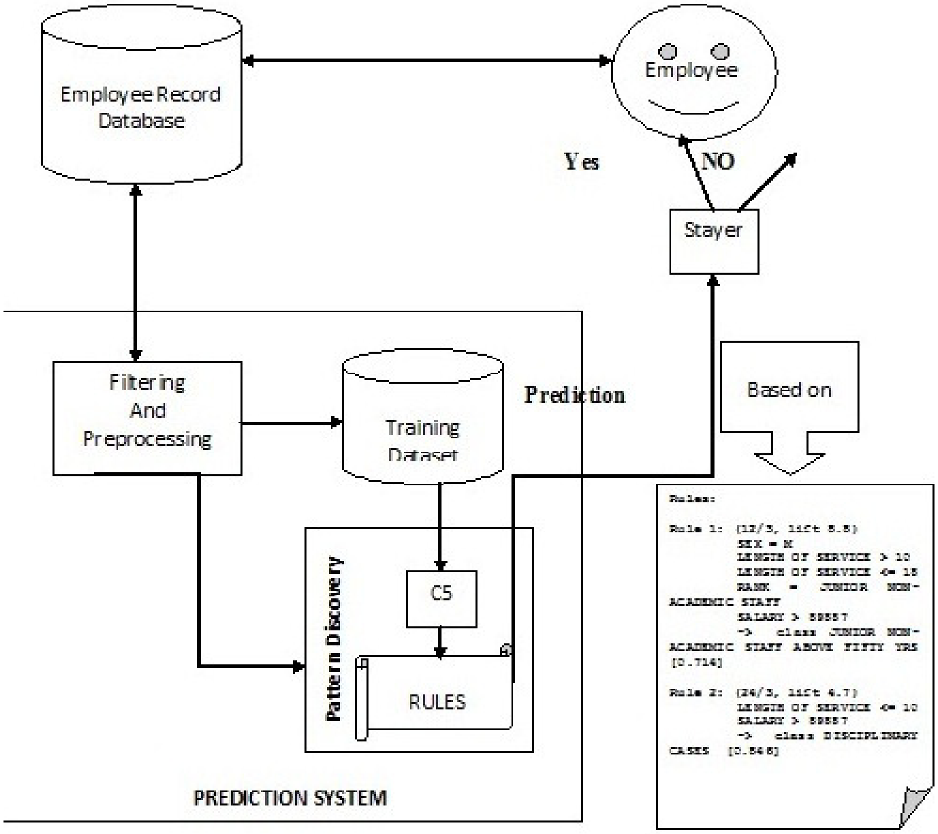
VOLUNTARY RETIRE 0.37

LENGTH OF SER 25
SEX 0
STATE OF ORIGIN 0
RANK 30
SALARY 350000

Figure 2: Predicting New Cases Using the See5 Boost Decision Tree Rules

Rules such as those generated from the C5 algorithm could be used to develop an employee attrition prediction system for organizations. The objective of the system is to identify employees with high attrition potentials based on the rules generated in this study from historical employee data with similar attributes. The proposed framework for this application is shown in figure 3.

The dataset contained in the historical employee records database will be used to develop the employee records database. The pattern discovery module will train the preprocessed data in the database using the C5 algorithm to generate prediction rules which can be used to predict the likely future behavior of employees. The prediction system can be incorporated into HR software to be used by HR departments during annual performance reviews and when analyzing job prospects of new employees.



```
Rules:

Rule 1:  (12/3, lift 5.8)
         SEX = M
         LENGTH OF SERVICE > 10
         LENGTH OF SERVICE <= 15
         RANK   =   JUNIOR   NON-
ACADEMIC STAFF
         SALARY > 59557
         ->   class  JUNIOR  NON-
ACADEMIC  STAFF  ABOVE  FIFTY  YRS
[0.714]

Rule 2:  (24/3, lift 4.7)
         LENGTH OF SERVICE <= 10
         SALARY > 59557
         ->   class DISCIPLINARY
CASES   [0.846]
```

Figure 3: Framework for Employee Attrition Prediction System

6. CONCLUSION

In this study an employee data set which contained three hundred and nine (309) complete records of employees of one of the Higher Institutions in Nigeria who worked in and left the institution between 1978 and 2006 was used. The demographic and job related records of the employee were the major interest which were used to classify the employee into some predefined employment attrition classes. The Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows data analysis software were used to generate five decision tree models and two rule-sets. The predictive model with the best performance bases on accuracy rate was used to predict new cases of employee attrition and a framework for software that can implement an employee attrition prediction system was proposed.

Results obtained from the study shows that employee Salary and Length of service were determining factors for predicting employee attrition in the institution whose data was used for the case study. Employees who have worked longer in the organisation with no reasonable increase in income are likely to be more discouraged which influences their attrition. Also, low ranking employees with very few years of service put in are likely to turnover when they realize the income may not improve given their low ranks, they therefore leave in search of better paying jobs.

The findings of this study further supports studies that have been carried out in the area of predicting employee attrition. It also supports the findings of Nagadevara, et al (2008) that it is possible to predict employee turnover intentions even before they had make their final decision to leave. A limitation to this study was the reluctance of organizations to give out their data for research purposes, therefore, organizations that would like to exploit the benefits of studies such as this, should be ready to provide data that will be needed to efficiently implement the proposed model.

REFERENCES

1. Allen D. G., (2008), Retaining Talent: A Guide to Analyzing and Managing Employee Turnover, *SHRM Foundation's Effective Practice Guidelines Series,* SHRM Foundation.
2. Berson A., Smith, S. and Thearling, K. (1999). Building Data Mining Applications for CRM.. McGraw-Hill
3. Bharati, M. Ramageri, (2010), Data Mining Techniques And Applications, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
4. Dijkstra , E. (2008). What Drives Logistics Professionals? *http://europhia.com/docs/europhia-research-what-drives-logistics-professionals.pdf. Retrieved 2009-01-21.*
5. Hamidah J., AbdulRazak H., and Zulaiha A. O. (2011). Towards Applying Data Mining Techniques for Talent Managements, *2009 International Conference on Computer Engineering and Applications, IPCSIT* vol.2, Singapore, *IACSIT Press.*
6. Igbaria, M. (1991). Career Orientations of MIS: An Empirical Analysis. *MIS Quarterly*, 151-169.
7. 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
8. Kotsiantis, S.B. (2007) Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 31,249-268
9. Marjorie, L. K.. (2007). Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis. Texas, A&M University College of Education.
10. Nagadevara, V., Srinivasan, V. & Valk, R. (2008). Establishing a Link between Employee Turnover and Withdrawal Behaviours: Application of Data Mining Techniques, *Research and Practice in Human Resource Management*, 16(2), 81-99.
11. Wie-Chiang H., Ruey-Ming C. (2007); A Comparative Test of Two Employee Turnover Prediction Models. International Journal of Management; June 2007; Vol.24 No.2; pp. 216 – 229.