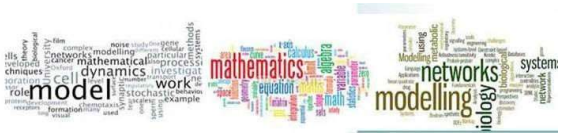

Article Citation Format
Baale, A.A. & Adelodun, F.O (2018): Analyses of Students' Vocational Data Using Some Selected Classification Algorithms. Advances in Mathematical & Computational Sciences . Vol. 6 No. 2. Pp 51-62.

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 27rd May, 2018
Review Type: Blind Final
Acceptance:: 11th June, 2018
DOI Prefix: 10.22624/AIMS/MATHS/V6N1P5

Analyses of Students' Vocational Data Using Some Selected Classification Algorithms

[1]Baale, A.A. & [2]Adelodun, F.O
[1]Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria.
[2]The Polytechnic, Ibadan, Oyo State, Nigeria.
E-mails: [1]aaadigun@lautech.edu.ng; [2]adelodunfelicia@gmail.com

ABSTRACT

Unemployment is a major issue battling the Nigerian economy. Today's trend is shifting towards skills acquisition rather than on certificate qualification which are capable of making the youths self-dependent, job creators and not job seekers (Okolocha,2012). The management of The Polytechnic, Ibadan, Nigeria is working tirelessly to ensure that each student learns a vocational skill to prepare them for entry into the labour market (Adebayo, 2016). In this study, data mining techniques were used to analyze students' vocational data of The Polytechnic, Ibadan in Nigeria to discover patterns and relationships that will help the school management to make important decisions and plan better for more productive execution of their vocational skills programme. Experiments were performed using ID3, C4.5 and Naïve Bayes classification algorithms to build models under WEKA 3.8.2 environment. Hold- out method and 10 folds cross-validation method were used to test the models. Confusion matrix was used to evaluate the performance of the models on the basis of accuracy, sensitivity and time taken to build the models and it was discovered the C4.5 algorithm gave the best performance with both validation techniques.

Keywords: Students, Academic, Performance, Evaluation, Vocation, Classification, and Algorithms.

1. INTRODUCTION

Data mining is a step in knowledge Discovery in Databases (KDD) and aims to discover useful information from huge amount of data (Jasser, Sidi, Mustapha and Binhamid, 2013). Even though it is an exploratory process, it can be used for confirmation investigations (Berson, Smith and Thearling, 2011). Data mining involves extracting hidden analytical information from large databases using multiple algorithms and techniques. Classification is the most commonly applied data mining technique (Ramageri, 2017). Data mining is an on-coming research field that received a great interest from researchers due to its significant benefits in educational institutions but it is of note that not much work has been done for evaluating performance of students in technical/vocational trade programmes (Ukwueze and Okezie, 2016). Researchers are more interested in analyzing students' academic subjects, attendance, demographic data and web usage than analyzing their vocational skills (Saleh, 2015).

Vocational Education (VE) is a means of acquiring new skills and competencies that extend professional opportunities. European Centre for the Development of vocational training (CEDEFOP) in 2011 refers to vocational education's achievement as a medium level qualification that can protect people from becoming unemployed. CEDEFOP further classified vocational education benefits under two main categories: economic benefits and social benefits. The economic benefits include economic growth, labour-market outcomes, firms' performance, employees' productivity, employment opportunities, earnings and professional status/career development. The social benefits include crime reduction, social cohesion, inclusion of disadvantaged groups, life satisfaction and individual motivation.

The primary objectives of Vocational Education and Training (VET) are to prepare students for a better entry into the labour market and for advancement in their chosen careers (Eynon, 2013). It is being incorporated into the Nigerian Educational system to battle the problem of unemployment among the youths. The Polytechnic, Ibadan, Nigeria has been working tirelessly to ensure that students learn vocational skills. The management of The Polytechnic usually face a lot of difficulties in planning for the students' vocational skill training. In this study, students' vocational data were analysed using three classification algorithms to help school management to improve decision making and plan better for more productive execution of the vocational skill programme.

2. LITERATURE REVIEW

Data mining involves extracting hidden analytical information from large databases using multiple algorithms and techniques. Classification and prediction are two forms of data analyses that can be used to extract the models describing important data classes or to predict the future data trends (Krishna and Kiruthika, 2015). The classification task is characterized by a well-defined definition of the classes and a training set, consisting of predefined examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it. Decision trees, Naive Bayes, Artificial Neural Network and Support Vector Machines are the most widely used classification techniques in data mining (Archana and Elangovan, 2014).

Decision Trees

The basic algorithm for decision tree is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The decision algorithms are known to be reasonably fast, accurate and less time consuming. They provide insights into the decision making process (Han and Kamber, 2006). The decision tree is efficient and suitable for both large and small datasets (Quadril and Kalyanker, 2010). Interactive Dichotomizer 3 (ID3), ASSISTANT and C4.5. are three decision tree algorithms mostly used to make decision efficiently (Priya and Kumar, 2013).

ID3 was developed by Quinlan in 1979 and it constructs a decision tree based on information gain/entropy (Baradwaj and Pal, 2011). ID3 starts by splitting the original set of the attribute which provides the maximum gain or the least entropy. This will either generate leaf nodes (Single class) or nodes with multiple classes. The algorithm is applied repeated on the new non-leaf nodes until a leaf node is got and there is no need to split further. C4.5 was developed by Quinlan in 1993 as an extension for ID3 algorithm to handle problems that ID3 could not deal with. ID3 handles only categorical values, does not handle missing values and no pruning is done, where as C4.5 measures numeric values, categorical values, deals with missing values and prunes noisy data (Singh and Giri, 2014). Pruning is used in C4.5 to avoid over fitting to noise in data.

Naïve Bayes Algorithm

The Naïve Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumption. It is robust to noise in input data. It is easy to construct and interpret, so users unskilled in classifier technology can understand why it is making the classification it makes (Archana and Elangovan, 2014). Naïve Bayes classifier simplifies computation and exhibits high accuracy and speed when applied to large databases (David, Saeb and Rubeaan, 2013).

WEKA

Waikato Environment for Knowledge Analysis (WEKA) is a popular suite of machine learning software written in Java developed at the University of Waikato, New Zealand in 1993, first published in 1994. It is a free software licensed under the GNU General Public Licence. GNU is a free unix-like operating system distributed by the Free Software Foundation. WEKA is a workbench that supports many activities of machine learning practitioners. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions (Bouckaert, Frank, Hall, Holmes, Pfahringer, Reutemann and Wilten, 2010).

2.1 The Cross-Industry Standard Process For Data Mining (CRISP-DM)

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a cycle process for development and analysis of data mining models (Leventhal, 2010). This is a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and technology used. It is a framework for recording experience. It allows projects to be replicated, aids project planning and management and helps to obtain reliable results.

2.2 Review of Related Works

Saleh (2015) applied ID3, C4.5 and CART classification algorithms on students' database at different locations of vocational skills and predicted their results for the final assessment. The prediction accuracy and execution time were the criterions used to evaluate and compare the performance of the different predictive models. The prediction accuracies obtained by the predictive models in four experiments varied as follows: ID3 has 60.4% - 64.6%, C4.5 has 64.8% - 65.5% and CART has 62.5% - 65.6% accuracy. The highest accuracy was obtained by CART (65.6%) and ID3 had the lowest accuracy.

Gorikhan (2016) implemented classification techniques for a vocational institute known as IAT to predict the number of students who were likely to pass or fail and the prediction analysis was applied for newly enrolled students. The results of the analysis was given to teachers and steps were taken to improve the academic performance of the weak students. The classification techniques used were decision tree, KNN, logistic regression, Support Vector Machine and Neural Network. Two datasets were examined, the first dataset consist of the students details with their maths and science scores in the previous exams while the dataset two contains only the students details. It was found that the decision tree proved out to be the most accurate prediction model and least in accuracy is K-NN for data set 1 while K-NN has highest accuracy and Neural Network has lowest accuracy for data set 2. Ukwueze and Okezie (2016) applied various data mining classification algorithms to predict students' performance in Technical Trades and the classification efficiencies of those algorithms were evaluated. The study investigated the accuracy of K-Nearest Neighbour, Neural Network, Decision trees and Naive Bayesian techniques. The attributes considered were assessment of continuous practical and theoretical competency test, class test, assignments, general proficiency, attendance, laboratory work and the end of semester examination.

The results show accuracy ranging from 73.30% to 98.33%. The decision tree (precisely C4.5) algorithm performed highest in predicting students' performance while K-Nearest Neighbour gave the least. Demirhan and Hacioglu (2017) used data mining methods to create models that helped to make meaningful deductions by establishing relationships within records. The data set used was created through questionnaire by voluntary participation of students from Vocational School (Distance Education) of Trakya University, Tunca. The experiments were performed using WEKA 3.8.1. Naive Bayes, MLP, SMO, J48 Rep-tree, Random tree and Decision table were used as the classifiers. 10 folds cross validation was the test method used and performance was based on confusion matrix. They classified vocational data such as applied vocational course data and theoretical vocational. The result of classification analysis on applied vocational data had accuracies as follows; J48 also called C4.5(80%), Rep-tree(78%), Decision table(76%), MLP(71%), Naive Bayes(71%), SMO(63%) and Random table(59%). The result of classification accuracies on theoretical vocational data analysis were J48(79%), Rep-tree(78%), Decision table(79%), MLP(78%), Naive Bayes(69%), SMO(74%) and Random table(60%). For both analyses, J48/C4.5 performed better than all other classifiers.

3. METHODOLOGY

The approach adopted by this study is based on Cross Industry Standard Process for Data Mining (CRISP-DM) and was implemented with WEKA 3.8 in window 10 environment. The methodology is made up of five phases as shown in figure 1below

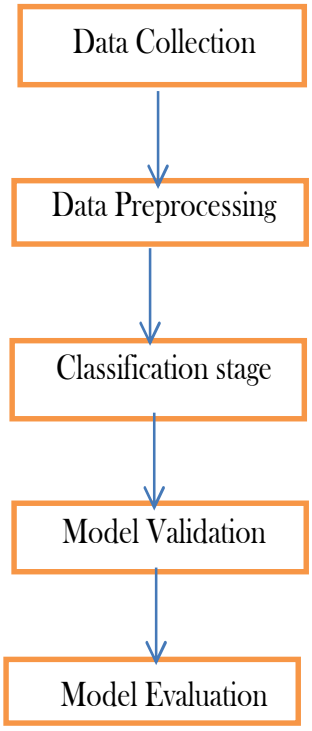

Figure 1: Phases of analysis of students' vocational data using ID3, C4.5 and Naïve Bayes

Data Collection

The data for this study was collected from the students' database of Vocational Skills and the Entrepreneurship Study Centre (VSESC) of The Polytechnic, Ibadan, Nigeria for the period of five years: 2012 to 2016. The data comprised of students' general data and academic data in vocational skills chosen by the students. The students' general data such as names, gender and course of study among others were used. The academic data included scores of students in the different vocational skills. The scores comprised the continuous assessments, practical and semester examination.

Data Pre-processing

This stage comprises of two major steps, which are data cleaning and data transformation.

Cleaning of Data

The initial data collected was a total of 19,021 but after the cleaning by removing noise and missing values, the data left was 17,582. The fields that were removed are: name, email, form number, date of registration and date of birth. The name field was removed in order to protect the identity of the students. The form number field was removed because it is irrelevant to the study. The email, date of registration and the date of birth fields were removed because there were a lot of errors in these fields. This was necessary in order to get good mining results.

Data Transformation

The remaining datasets were transformed into the form suitable for the data mining tasks by converting them into Comma Separated Value (CSV) format for WEKA workbench. The final attributes considered for this analysis comprises of eleven (11) attributes and their data types are shown in table 1 below, alongside with their descriptions. From table 1, ND represents Ordinary National Diploma and HND represents Higher National Diploma. The department includes Arts and Design, Mechatronics, Electrical Electronics, Mechanical Engineering, Civil Engineering, Building Technology, Quantity Survey, Estate, Management, Public Administration, Purchasing and Supply, Local Government, Marketing, Business Administration, Insurance, Banking and Finance, Accountancy, Survey and Geoformatics, Geology, Town Planning, Architecture, Office Technology Management, Mass Communication, Microbiology, Biology, chemistry, Applied Chemistry, Science Laboratory Technology, Computer Science, Mathematics and Statistics. The states of origin comprised of 31 states of the federation which includes Kebbi, Kano, Kaduna, Niger, Kwara, Oyo, Ogun, Osun, Ekiti, Ondo, Kogi, Benue, Nassarawa, Plateau, Bauchi, Taraba, Bornu, Adamawa, Ebonyi, Rivers, Cross River, Imo, Lagos, Edo, Delta, Bayelsa, Anambra, Enugu, Abia, AkwaIbom and Sokoto states. The vocational skills include bead making, website design and software development, car repairs and servicing, textile technology, electrical installation, catering and confectioneries, barbing, computer cloning and networks, garment making, cosmetology, GSM repairs and servicing, pure water production and marketing, brick and block moulding, refrigeration and air conditioning, household products, welding and fabrication, carpentry and panelling.

Table 1: Attributes with their data types and descriptions

No	Attributes	Data type	Description and possible values
1	Gender	Text	(Male, female)
2	Programme status	Text	(fulltime, daily part time (DPP) part-time and Annex)
3	Class	Text	(ND I, ND II ND III, HND I HND 2, HND 3)
4	Semester	Text	(First, Second)
5	Department	Text	(29 departments)
6	State of Origin	Text	(31 states of the federation)
7	Vocational skill	Text	(17 vocational skills)
8	Score 1	Integer	(continuous assessment + practical) (0 – 60)
9	Score 2	Integer	(semester examination) (0 – 40)
10	Score 3	Integer	(Score 1 + score 2)
11	Grade	Alphanumeric	70 – 100 = Excellent, 60-69 = Average 50 – 59 = Good, 0-49 = poor

Classification Stage

The data fed into the WEKA data mining tool was randomly divided by WEKA validation operator into training set and testing sets into:

Experiment 1: 66½% training set and 33½ testing set (Hold-out method)

Experiment 2: 10 folds cross-validation

Model Validation

Two validations were performed. In the first experiment 33½ % of the data were used as test data to validate the algorithms and 10 folds cross-validation were used to validate the algorithms in the experiment 2.

Cross-Validation and Holds-out Methods are some forms of validation, where data is split into training and testing sets. Cross-validation is a statistical method of evaluating model by dividing data into two segments: one used to train or build model and the other segment is used to validate the model in data mining; 10-fold cross-validation is the most common. Hold-out approach involves randomly splitting data into training and test data, usually two third (2/3) for training, one-third (1/3) for testing. Hold-out estimate can be made more reliable by repeating the process with different subsamples (Stefanowski, 2010).

Model Evaluation

The performance of the models was examined using performance measure based on confusion matrix. The classification algorithms ID3, C4.5 and Naive Bayes were compared using the following performance matrices: accuracy, sensitivity and time taken to build the models.

Evaluation Measure

Several measures such as accuracy, specificity, sensitivity or recall and precision have been proposed to evaluate model performance in classification problems.

A confusion matrix is a table that contains information about actual and predicted classifications for any given classification algorithm (Chawla, 2005). It is often used to describe the performance of a classification model on a set of test data for which the true value is known. The equations for calculating Accuracy, Specificity, Sensitivity and Precision are described below in equations 1.1, 1.2, 1.3 and 1.4 respectively.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad \textit{equation 1.1}$$

$$\text{Specificity} = \frac{TN}{TP + FN} \qquad \textit{equation 1.2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \textit{equation 1.3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \textit{equation 1.4}$$

Where

TP = the proportion of positive cases that were correctly identified. For instance, predict as positive when actual positive.

TN = the proportion of negative cases that were classified correctly. Predict negative when actual negative.

FN = the proportion of positive cases that were incorrectly classified as negative. Predict negative when actual positive.

FP = the proportion of negative cases that were incorrectly classified as positive. Predict positive when actual negative.

4. RESULTS AND DISCUSSION

The results of the classification of the students' vocational data are explained below.

Classification of gender data

From the graph shown in figure 3 below, the male students were 9656 representing 54.92% of data, while the female students were 7926, representing 45.08% of data.

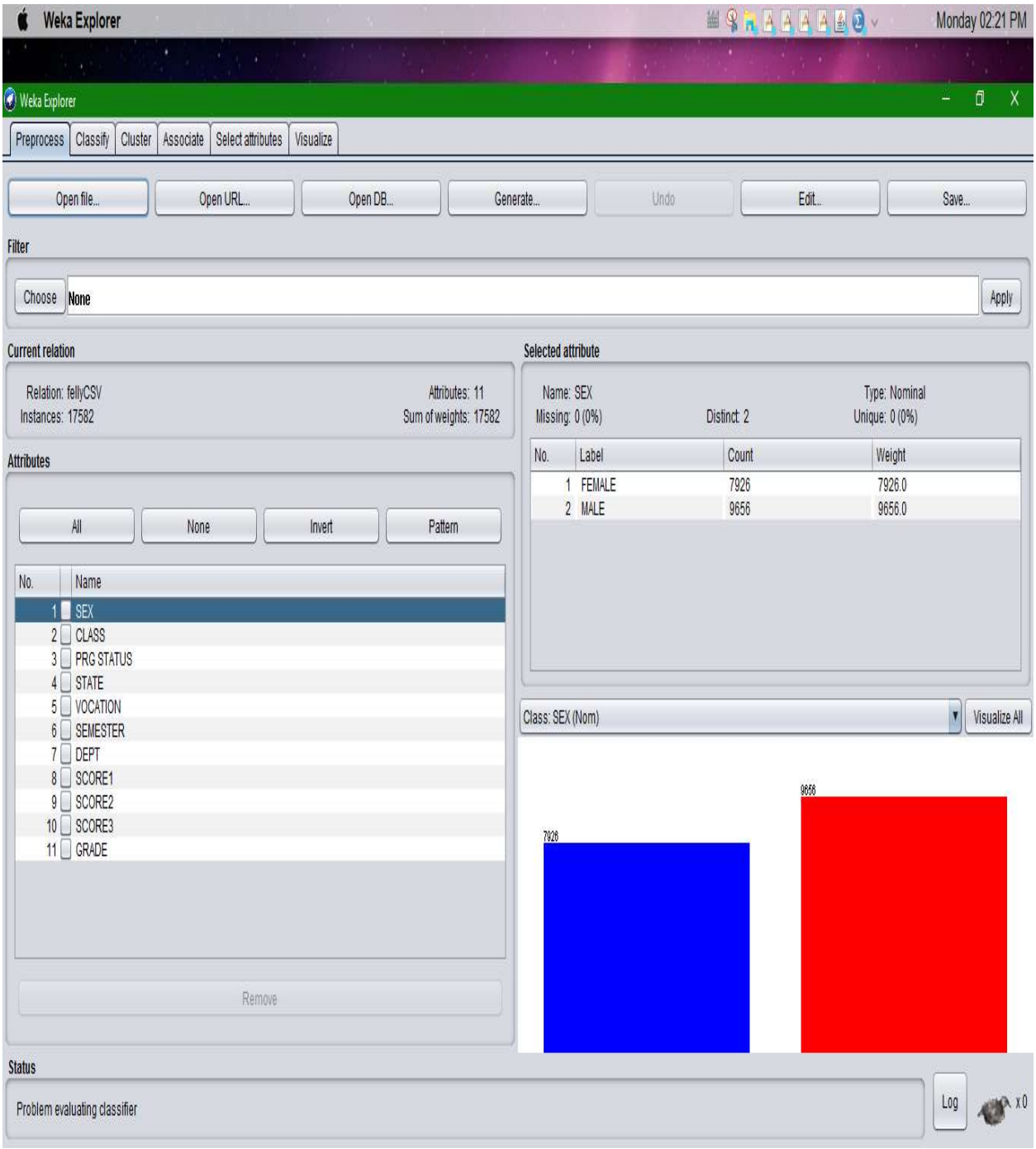

Figure 2: Classification of Students 'Gender

Classification of students' distribution over vocational skills

From figure 3 below, more male students enrolled for vocations like electrical installation and barbing when compared to the number of the female students who enrolled for those vocations. Whereas, more female students populated in vocations like bead making, catering and confectioneries and garment making than male students as depicted in figure 4. Very few students inrespective of their gender enrolled for carpentry, welding, brick and block moulding.

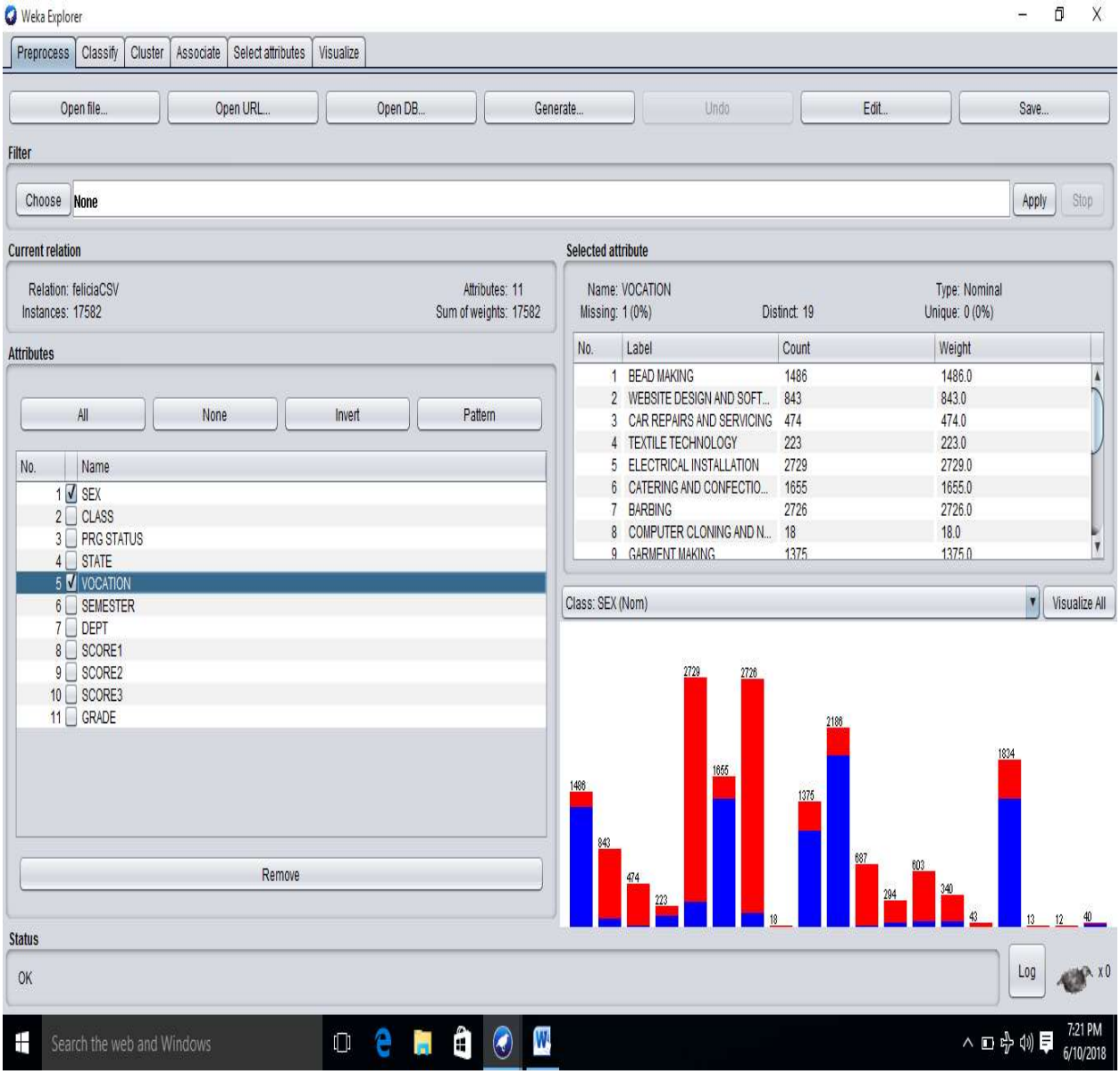

Figure3:Classification of students' distribution over vocational skills

Table 3:Tabulation of results of vocational data using different data size as training set and testing set

Sample size	Algorithms	Correctly classified instances %	Incorrectly classified instances %	Sensitivity/ Recall	Time taken to build the model	Mean absolute error
S1 66½ - 33½	ID3	73.5545	26.4455	0.736	0.09	0.1271
S1 66½ -33½	C4.5	99.981	0.019	1.000	0.33	0.0003
S1 66½ -33½	NB	91.8483	8.1517	0.918	0.08	0.0422
S2 10CV	ID3	73.0406	26.9594	0.730	0.05	0.1278
S2 10CV	C4.5	99.9488	0.05I2	0.999	0.2	0.0004
S2 10CV	NB	91.7757	8.2243	0.918	0.05	0.0416

The tabulation of the student vocational data using different data size as training set and testing as shown in table 3 reveals that when the hold-out method (S1) was used, C4.5 obtained the highest accuracy of 99.981% and the least accuracy of 73.5545 was obtained by ID3. C4.5 also had highest sensitivity of 1.000 and least in sensitivity was obtained by ID3 (0.736). C4.5 took the longest time to build the model (0.33 second) while Naïve Bayes took the shortest time (0.08 second) to build the model.

When 10 fold-cross validation method (S2) was used to split the data C4.5 obtained the highest accuracy of 99.9488%, highest sensitivity of 0.999 but took the longest time of 0.2 second to build the model. The least accuracy (73.0406%) and lowest sensitivity of 0.730 was obtained by ID3. Both ID3 and Naïve Bayes took 0.05 second each to build the model.

5. CONCLUSION

The accuracies recorded in the analysis were above 70% and sensitivity lies between 0.73 and 1.0 revealing that all three algorithms are good classifiers for analyzing students' vocational data to uncover patterns and discover knowledge, but C4.5 performed better than others in terms of accuracy and sensitivity. The eleven (11) attributes selected for the analysis were relevant hence the high accuracy and sensitivity obtained from results of analysis. Classification of students' distribution over vocational skills (Figure 4.2) reveals that some vocations such as website design and software development, electrical installation and barbing, computer cloning and networks, car repair and servicing and GSM repair and servicing were dominated by the male students while vocations like bead making, cosmetology and household products were chosen by most female students showing that gender can influence students' choice of vocation.

The thriving vocations were bead making, website design and software development, electrical installation, catering and confectioneries, barbing, garment making, cosmetology and household products. Vocations like welding and fabrication, carpentry and panelling were not viable while the remaining vocations like computer cloning and network, pure water production, brick and block moulding were not doing too badly.

6. RECOMMENDATIONS

The Polytechnic management is advised to direct more fund towards the thriving vocations like barbing, cosmetology, household products and electrical installation. The students should be encouraged by organising frequent promotional and empowering those vocations like pure water production, computer cloning and networks and car repair and servicing that are still coming up by providing more equipment and manpower. Vocations like welding and fabrication, carpentry and panelling that are not viable can be removed from programme to save fund, space and time.

The vocational skills and training programmes are also recommended to other institutions of higher learning like the Universities and Colleges of Education that are yet to embrace vocational training because of the increasing unemployment rates among young graduates and also to curb crime and suicide among the youths. For subsequent analyses, C4.5 can be used for the classification of vocational data because of its accuracy.

REFERENCES

1. Adebayo, N. A. (2016): Rethinking Nigeria's Development Paradigm. The Entreprenuership – poverty Nexus. Inaugural lecture series 9, The Polytechnic, Ibadan, Nigeria.
2. Archana, S. and Elangovan, K (2014): A survey of classification Techniques in Data -Mining. International Journal of Computer Science and Mobile Applications, Vol 2 Issue 2, pg 65-71. ISSN: 2321-8363.
3. Baradwaj, B., K. and Pal, S.(2011): Mining Educational Data to Analyse students' performance: (IJACSA) International Journal of Advanced Computer science and Applications Vol. 2, No 6, pg 63-69.
4. Berson, A., Smith, S and Thearling, K. (2011): An overview of Data Mining Techniques. http//www.thearling.com/text/dmtechniques/dmtechniques,htm
5. Bouckaert, R.R., Frank, E., Hall, M. A, Holmes, G., Pfahringer, B., Reutemann, P. And Wilten, I. H.(2010): WEKA- Experiences with a Java Open- source Project. Journal of Machine Learning Research 11. Pg 2533-2541.
6. David, S. K., Saeb, A. T. M and Rubeaan, K. A. (2013): Comparative Analysis of Data Mining Techniques using WEKA in Medical Bioinformatics. Computer Engineering and Intelligent Systems. Vol. 4, No 13. ISSN 2222- 1719(paper).ISSN 2222- 2863(on line).
7. Demirhan, T and Harioglu, I. (2017): Performance and Achievement analysis of a dataset of distance education with WEKA. International Conference on Science and Education (IconSE). Eurasia Proceeding of Educational and Social Sciences (EPESS). Vol 8, pg 9-18.
8. European Centre for the Development of Vocational Training (CEDEFOP) (2011): The benefits of vocational education and training. Paper No.10. Luxembourg publication office of European Union.
9. Eynon, R. (2013): The rise of Big Data. What does it mean for education technology and media research? Hearing, media and Technology. Vol 38, No 3, pg 237-240.
10. Gorikhan N. A. (2016): A study on Implementation of classification techniques to predict students' results for institutional analysis. Dissertation submitted in partial fulfilment of the requirements for the degree of MSc IT. Faculty of Engineering and IT, The British University in Dubai.
11. Han, J and Kamber, M. (2006): Data concepts and techniques. Second edition. Morgan Kaufmann Publishers.

12. Jasser, M. B., Sidi, F., Mustapha, A. and Binhamid, A.K.T. (2013): Mining Students' characteristics and effects on University Preference choice: A case study of Applied Marketing in Higher Education. International Journal of Computer Applications. Vol 67, No 21, pg 1-5.
13. Leventhal, B. (2010): An introduction to data mining and other techniques for advanced analytics. Journal of Direct Data and Digital Marketing Practice, Vol 12, No 2, pg 137-153.
14. Okolocha, C. C. (2012): Vocational Technical Education in Nigeria: challenges and the way forward. Business Management Dynamics. Vol. 2, No 6, pp 01-08.
15. Priya, K. S. and Kumar, A.V. (2013): Improving the student's performance using Educational Data Mining. Int. J. Advanced Networking and Applications. Vol. 04, Issue 04, pg 1680- 1685.
16. Quadril, M. N. and Kalyankar, N. V. (2010): Drop out feature of student data for academic performance using decision tree techniques. GJCST. Vol. 10, Issue 2.
17. Quinlan, J. R. (1979): Discovering rules from large collection of examples: A case study, expert systems in the Micro Electronic Age. Edinburgh Press, Edinburgh.
18. Ramageri, B. M. (2017): Data Mining Techniques And Applications. Indian Journal of Computer Science and Engineering. Vol. 1, No 4, pg 301- 305. ISSN: 0976 – 5166.
19. Saleh, A. A. (2015): Education Data Mining to predict exam grades in vocational institutes. Dissertation submitted in partial fulfilment of the requirements for the degree of MSc Information Technology Management. Faculty of Engineering and IT, The British University in Dubai.
20. Singh, S and Giri, M. (2014): Comparative Study of ID3, CART and C4.5 Decision Tree Algorithms: A survey. International Journal of Advanced Information Science and Technology. Vol. 3, No 7.
21. Stefanowski, J. (2010): Data Mining-Evaluation of classifiers. Institute of computing science. Poznan University of Technology Poznan, Poland SE Master Course 2008/2009.
22. Ukwueze, F. N. and Okezie, C. C. (2016): Evaluation of Data Mining classification Algorithms for predicting students performance in Technical trades: International Journal of Engineering and Computer Science. Vol. 5, No 8, pg 17593-17601. ISSN: 2319-7242.