www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 8 Aug 2016, Page No. 17593-17601

Evaluation of Data Mining Classification Algorithms for Predicting Students Performance in Technical Trades.

Ukwueze Frederick N., Okezie Christiana C.

Department of Computer Education
University of Nigeria, Nsukka
email: frederick.ukwueze@unn.edu.ng
Department of electronic and Computer Engineering
Nnamdi Azikiwe University, Awka
Email: cc.okezie@unizik.edu.ng

Abstract

Measuring trainee's performance in technical/vocational trades involves some peculiar considerations. In this study a survey research method was adopted to generate relevant predictor variables. Our primary data was collected using a simple survey instrument on the regular and sandwich students; the rest was from the Examination Unit of the University Registrars Department and also from various assessment records and instructors' competency test records in the students department. The secondary data set included three year assessment records 2013 to 2015 in a course, VTE 201: Students Industrial Work Experience of the faculty of Vocational and Technical Education, University of Nigeria Nsukka. The raw data was preprocessed and converted to a required format. A total of 187 student records were obtained. This was used to train four selected classification algorithms whose accuracies were compared. Results showed that Decision Tree Algorithms performed best in predicting student's terminal performance in four categories of technical trades.

Introduction

Students performance in technical trades depends on many variables related to their academic, socio-economic and cultural backgrounds. These variables can be used as inputs to various analytical tools needed to predict their ability to go through a successful training programme in this area. Through such predictive analysis, those students capable of completing a desired programme can be identified. Those that might experience challenges can also be assisted in a planned fashion. For this to be done data relating to student's performance need to be available in large quantities. But as volumes of data in various fields are continually on the increase, data mining tools can be employed particularly, in this area. Data mining is defined as a method by which valuable information can be divulged from a large mass of data, (Vijayarani etal 2011). Data mining has also been in use in educational field, for many decades past, and is called Educational Data Mining. According to Saurabh Pal etal (2011), it can be employed in developing methods that discover knowledge from educational data warehouse. By this, we can discover

information which can be useful for prediction regarding students' performance in terms of final achievement or programme completion.

Data mining had provided many tools for studying student academic performance. Galit [2007] used three classification methods to study the learning behaviour of students and predicting their failure risk level before their terminal assessment. Of the three algorithms he used, viz: ID3, C45 and Naive Bays. Decision tree (C.45), outperformed the other two. Cortez and Silva [2008] carried out the same study using the following algorithms: Decision Tree, Random Forest, Neural Network and Support Vector Machine. Decision Tree, here again, outperformed others with a predictive accuracy of 93%. Garfinkel et al. (2010), and Marturana et al. (2012b), evaluated accuracies of various algorithms and obtained results showing that decision trees and Bayesian networks were better respectively in their studies. Both however worked on different areas of application. One on multi-user data ascription problem and the other on forensic analysis of hard disks.

It is of note that, not much had been done using any variable set closely relevant to performance evaluation of students in technical/vocational trade programmes. Moreover conflicting results from various authors about the accuracies of algorithms demand a closer study of the well-performing algorithms in relation to technical trades. In this study, therefore, the various classification algorithms were used to evaluate student's performance and the classification efficiencies of those algorithms evaluated. Relevant predictor variables were collected from various assessment records and instructor's competency test records to predict their performance at the end of the session. This study investigates the accuracy of K- Nearest neighbor, Neural Networks, Decision Trees, and Naïve Bayes, techniques for predicting student performance.

Study Objectives

(a) Generating data from source for predictive variables, and identifying those variables that influence the academic performance of students in technical trades,

(b) Application of selected machine learning algorithms to predict students' performance and

(c) Comparing the classification accuracies of the selected algorithms.

Methodology

This study involves a practical experimental design. A data mining methodology will be applied on Students performance records which had been determined by the assessment of continuous practical and theoretical competency tests. The continuous assessment was carried out by the instructors based upon student's performance in learning activities such as class test, assignments, general proficiency, attendance and laboratory work. The end semester examination is one that is scored by the student in semester examination. Each student has to get minimum marks to pass a semester in internal as well as end semester examination.

Method of Data Collection

A survey instrument was used to generate a dataset for this study. Primarily data was collected using a simple survey instrument on the regular and sandwich students; the secondary data was gathered partly from the Examination Unit of the University Registrars Department and also from various assessment records and instructors' competency test records in the students department. The secondary data set included three year assessment records 2013 to 2015 in a course, VTE 201: Students Industrial Work Experience of the faculty of Vocational Technical Education, University of Nigeria Nsukka. The designed questionnaire was used to collect student details. And based on this information, relevant variables were generated. The predictor variables are presented in Table 1 as elicited from the completed questionnaires**.**

Data Preparations

Data preparation involved conversion to ARFF format, necessary in the WEKA workbench. The data was extracted from result sheets collected from instructors. Attributes were checked for relevance, determining the possible impact on the learners), and redundant ones cleaned off. Noise was removed and missing values replaced. Here, derived variables were selected which can be used in the chosen data mining tool.

Feature/Category	**Feature List**	**Numeric/Nominal**	**Possible (Domain) Values**
Personality Attributes	sex	*Nominal*	*{male, female}*
	body weight	*Nominal*	*{underweight, normal, overweight,}*
	visual acuity	*Nominal*	*{normal, defective}*
Academic/Attributes	physically challenged?	*Nominal*	*{yes, no}*
	Last semester letter grades	*Nominal*	*{A > 70% , B>60&<70% , C>50% <60 %, D>44<50% F<45}*
	Class Assignment	*Nominal*	*{Good, Average, Poor}*
	Field Trip Participation	*Nominal*	*{Good, Average, Poor}*
	Technical proficiency	*Nominal*	*{Good, Average, Poor}*
	Attendance/Punctuality	*Nominal*	*{Good, Average, Poor}*
	Workshop Practicals	*Nominal*	*{Yes, No}*
Social Background	Industrial work Experience scheme (SIWES) marks	*Nominal*	*{A > 70% , B>60&<70% , C>50% <60 %, D>44<50% F<45}*
	student's accommodation category	*Nominal*	{*Hostel, off campus*}
	residential status	*Nominal*	{*individual, joint*}
	means of transportation to school	*Nominal*	{ *campus shuttle, Hired bike, personal car* }
	Category of primary school attended	*Nominal*	{*private, government*}
	Category of post-primary school attended	*Nominal*	Commercial, Technical, Secondary.
	Area of subject study	*Nominal*	*{Sciences,Arts,Technical,Business}*
	marks/grade obtained at secondary level	*Nominal*	*{A > 70% , B>60&<70% , C>50% <60 %, D>44<50% F<45}*

Table 1: Definition of Predictor Variables

Tools and Techniques

Classification

Saurabh Pal etal (2011) defined classification as '*the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large*'. The technique, according to the authors frequently employs machine learning classification algorithms. The data classification process involves discovering and sorting data into groups based on similarities of data; three essential stages are involved: Learning, accuracy evaluation and classification. During the Learning stage, the training set is created and classified, then, the classification rules (model) is analysed to evaluate its learning accuracy. If the accuracy meets the requirements, the rules will then be applied to classify new data instances. This model is called a classifier. The following is the data processing workflow encompassing the sequence of operations illustrated in Figure1.

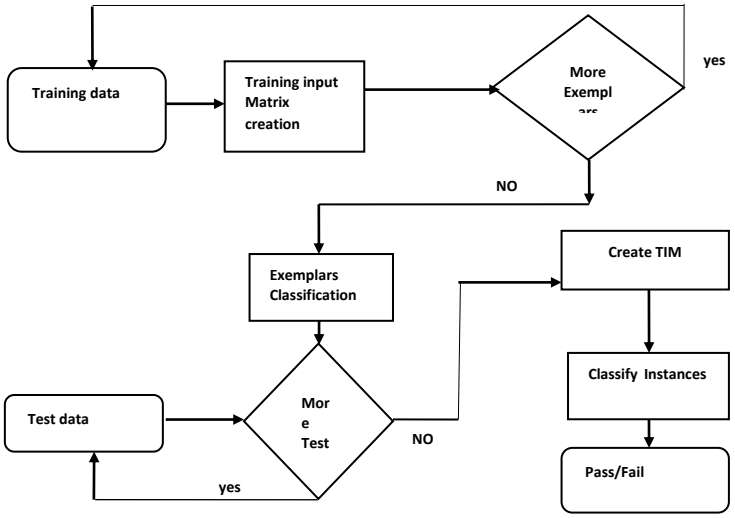

Figure 1. Data Classification Process Flowchart.

Training set creation.

This involved collecting a uniform set of previous class observations with a known class. Collection of exemplars is usually according to their class. Population of the training input matrix was a critical procedure. This was done by using a combination of data from as much exemplars as possible which is necessary to make it sufficient for getting a credible result. These constitute the input to the classifiers. The number of exemplars used are shown in fig 2-5.

Training set classification.

This stage processed the training input matrix. Two classification methods are usually employed for this: the multiclass categorization, and binary categorization. In multiclass method, each instance is classified according to the more likely category in the training set; a classifier per each category is trained. In binary categorization the classes are assigned binary values (pass or fail; viable/not viable etc.), depending on whether the student may likely pass or fail; or can complete the programme or not. In our case study, binary categorization has been chosen as the classification method.

Test set creation.

Once the training step was over, and learning accuracy had been computed, the model wass ready to classify new test instances populated with data. A test instance was created per each student and the test input matrix was populated as a collection of all the available test instances.

Classifying the test set

The test input matrix was fed into each learner for classifiying the instances.

Selection of Classification Rule Algorithms.

Four algorithms have been chosen for this study:

C4.5 Algorithm

Although C.45 has been superseded by a much improved. C.50 wel retained the use of C.45 being the version implemented in the WEKA package used in the analysis. Decision tree algorithms are known to be reasonably fast and accurate and less time consuming Chanchal etal. (2013). Rules generated by this algorithm are easily understandable by non technical professionals.

Instance Based Learning with K-Nearest Neighbor

Instance based learning is usually referred to as lazy classifiers because it takes the longest possible time to learn. A new instance of data is assigned to a class based on some distance metrics. Most common metric is the Euclidean distance. Each new instance is compared to each of the existing instances. The nearest neighbor, i.e. the instance with the shortest distance becomes the class to which it is assigned. Once the k nearest training instances has been located, the test instance is classified accordingly to the majority class in the neighborhood.

Nearest Neighbor Search is Expensive when searching object is in high dimensional space. It grows exponentially with the size of the searching space. Chanchal etal (2013), but we considered this algorithm because of the relative small size of attributes we have.

Neural network

Neural network is a weighted set of connected input/output units having layers in between the input and the output. The layers may be single (SLP), or many (MLP). Multi layer Perceptron classifiers particularly, are based on backpropagation algorithm which is used in the learning process.

Neural Network algorithms have been proved highly accurate in recognizing data; they also show very high classification accuracy, Chanchal etal. (2013)

DOI: 10.18535/ijecs/v5i8.29

Naive Bayes

Naive Bayes is a simple form of Bayesian classifiers, a set of statistical classifiers based on Bayes' Theorem. They are effective in predicting the probability that a record belongs to a particular class. Naive Bayesian classifiers are comparable in performance to decision trees and exhibit high accuracy and speed when applied to large datasets.

Data Classification in WEKA

Weka is an open source machine learning suite developed in University of Waikato New Zealand. It has four different application modes which can be worked within; they are the Explorer, Experimenter, KnowledgeFlow, and lastly SimpleCLI (command line interface). The Explorer mode provides an intuitive user interface for loading, filtering, clustering, classifying, and visualization of test data. The Experimenter, KnowledgeFlow, and SimpleCLI were not used in this experiment. After the data preparation was done, WEKA was used to run its suite of algorithms on the test data.

Experimental Results

The following tables show the experimental results for the four trade areas.

Experimental Results

The following tables show the experimental results for the four trade areas.

Algorithm	KNN			J48	NN	BN
	k=5	k = 3	k = 1			
Number of Instances	64	64	64	64	64	64
Correctly classified	95.30%	95.91%	97.22%	**98.33%**	97.37%	88.51%
Incorrectly classified	4.70%	4.08%	2.78%	1.67%	2.63%	11.49%

Table 2: Classification results for **Business Education**

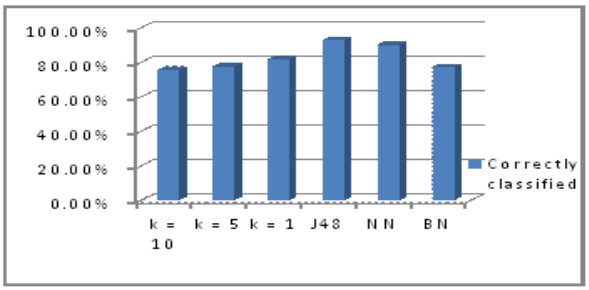

Fig 2: Classification results for **Business Education**

Algorithm	KNN			J48	NN	BN
	k = 10	k = 5	k = 1			
Number of Instances	17	17	17	17	17	17
Correctly classified	70.25%	73.90%	75.30%	**93.77%**	91.45%	86.00%
Incorrectly classified	29.75%	26.10%	24.70%	6.23%	8.55%	14.00

Table 3: Classification results for **Computer Education**

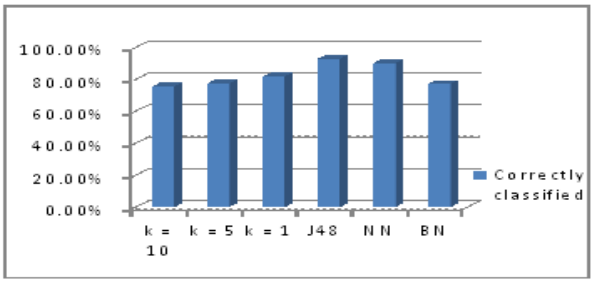

Fig 3: Classification results for **Computer Education**

DOI: 10.18535/ijecs/v5i8.29

Algorithm	KNN					
	k = 10	k = 5	k = 1	J48	NN	BN
Number of Instances	70	70	70	70	70	70
Correctly classified	82.44%	85.63%	86.3%	**96.88%**	95.11%	87.32%
Incorrectly classified	17.56%	14.37%	13.70%	3.12%	4.89%	12.67%

Table 4: Classification results for **Home Economics Education**

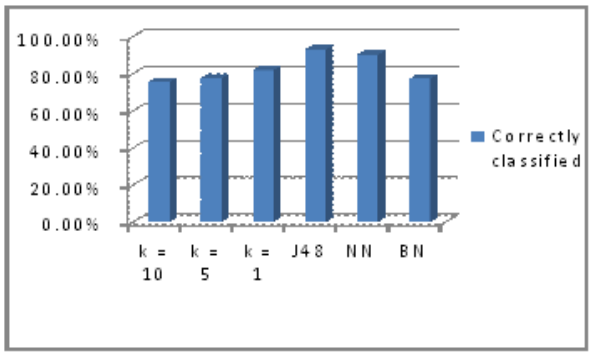

Fig 4: Classification results for **Home Economics Education**

Algorithm	KNN					
	k = 10	k = 5	k = 1	J48	NN	BN
Number of Instances	36	36	36	36	36	36
Correctly classified	75.56%	77.54%	81.85%	**92.97%**	90.20%	77.23%
Incorrectly classified	24.44%	22.46%	18.15%	7.03%	9.80%	22.76%

Table 5: Classification results for **Industrial Technical Education**

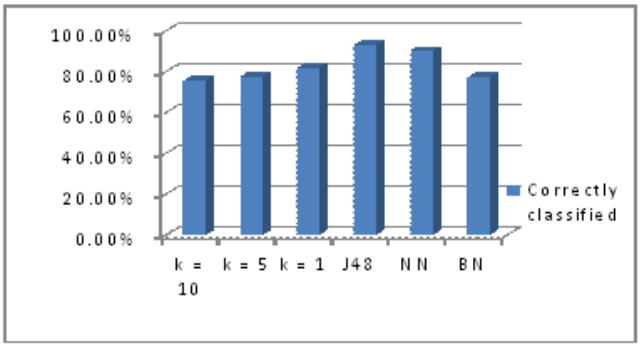

Fig 5: Classification results for **Industrial Technical Education**

Discussion of Results

The results shown in tables 2-5 above for the four selected trade areas show accuracy ranging from 75.30% to 98.33% accuracy. With J48, a highest accuracy was obtained, while k-nearest neighbor gave the least. It is observed that as the value of k increased, our accuracy slightly declined for all trade areas tested. Tests were run for each with 10, 5, and 1 nearest neighbours. This result agrees with Al-Radaideh, et al [2006] who applied a decision tree model to predict the final grade of students and obtained an outcome indicating that Decision Tree model had better prediction than other models. It also corroborates other results from Galit (2007) and Cortez Etal (2008). However, the result contradicts with the accuracy results obtained by Sumanetal (2012) who found that Multilayer Perceptron is a better algorithm across five different data sets; and also Kotsiantis, et al. [2004**]** who applied a filter based variable selection technique to select highly influencing variables and it was reported that the Naïve-Bayes algorithm yielded highest predictive accuracy for a binary dataset.

Conclusion and Future Work

In this study, an evaluation of the performance of four selected algorithms in terms of classification accuracy has been carried out using percentage of accuracy measure. Accuracy has been measured on each datasets. The results show that the decision tree algorithm performs better in predicting the performance of students in technical/vocational trade programs. The conflict of the obtained results with earlier results on the same subject suggest a closer examination of the possible effects of overfitting and imbalaced data problem on this category of classification and to see if mitigating these effects can bring any harmony in the results. Moreover better variable selection technique can also be employed for a possible improvement.

References

1. *AI-Radaideh Q. A., E. W. AI-Shawakfa, and M. I. AI-Najjar, (2006). Mining student data using decision trees, International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.*
2. *Chanchal Yadav, Shuliang Wang and Manoj Kumar (2013), Algorithm And Approaches To Handle Large Data-A Survey. International Journal of Computer Science and Network, Vol 2, Issue 3, 2013*
3. *Cortez P., and A. Silva, (2008), Using Data Mining To Predict Secondary School Student Performance, In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.*
4. *Simson L. Garfinkel (2010), An Automated Solution To The Multiuser Carved Data Ascription ProblemIEEE Transaction On Information Forensics And Security vol 5 (4) 868-882*
5. *Kotsiantis S., C. Pierrakeas, and P. Pintelas, (2004) Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques, Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.*
6. *Marturana, F., Berte', R., Me, G., and Tacconi, S. (2012b). Triage-based automated analysis of evidence in court cases of copyright infringement. In IEEE International Workshop on Security and Forensics in Communication Systems. 2012.*
7. *Saurabh Pal and Brijesh Kumar Baradwaj (2011) Mining Educational Data To Analyze Students" Performance. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011*
8. *Suman, Rohit Arora (2012) Comparative Analysis of Classification Algorithms on Different Datasets using WEKA International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012*
9. *Vijayarani S. and Divya M. (2011) An Efficient Algorithm for Classification Rule Hiding. International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011.*

Authors Profiles:

Ukwueze Frederik N, a Lecturer and former Coordinator, Computer Education at University of Nigeria, Nsukka, holds a bachelors' degree in Electronic Engineering and a Masters degree in Computer Science of the University of Nigeria. He is currently working in the area of Data Mining and Digital Forensics. His other research interests include workskill development and Engineering/Technology Education. He is actively engaged in teaching of undergraduate and postgraduate courses including, Computer Graphics,

Data communication and Networks, Computer Architecture and Hardware Management. Prior to the University career, Engr. Ukwueze, had worked in various corporate organizations involved in computer, communication and project management consultancy.

Okezie Christian C. is a professor and current head, Department of Electronic and computer Engineering Nnamdi Azikiwe University Awka Nigeria. She holds a Bachelor of engineering (B.Eng) degree in electronic and computer Engineering and also a masters (M.Eng) and doctorate (PhD) degrees in Computer Engineering. Starting a career as a Maintenance Engineer with Shell Petroleum Development Company (SPDC), she joined the academic faculty in 1998 where she has been engaged in teaching and research, having published over 70 academic journal articles and several conference papers, textbooks and monographs. She has served as editorial board member of many international journals, including Journal of Electrical and Telecommunications Systems Research (Electroscope). She specializes in the areas of Computer Engineering, Microprocessors, Data Communication Engineering and Database systems.