University of Ibadan
Journal of Science and Logics in ICT Research

A Comparative Performance Analysis of Data Mining Classification Algorithms on Predicting Students' Placement into Subject-Area Classes in Senior Secondary Schools

Omoyele, T. D. and Akinola, S.O.

Department of Computer Science,
University of Ibadan, Oyo State, Nigeria

Abstract

The information age has enabled many institutions to gather large volumes of data. However, the usefulness of this data is negligible if knowledge cannot be extracted from it. Data mining attempts to answer this need. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information which can help in decision making. A secondary school with efficient data mining approach can use the prediction model to place junior secondary students into various class categories (Science, Arts or Commercial) in the Senior Secondary Schools in order to improve quality of education and increase success rate. In this study, students' scores in English, Civic Education, Social Studies, Mathematics, Integrated Science, Fine Art, Business Studies and Introductory Technology at the Junior Secondary Classes were obtained. Classification algorithms such as J48, IBK, JRipper and Naïve Bayes were applied on the Junior Secondary final score data in order to determine the best class category of students at Senior Secondary level. The result of this study shows that J48 algorithm performed well for the prediction and can be used as a best predicting model algorithm for the type of data collected. The result of this experiment shows that the marks obtained in the various subjects in the Junior Secondary Schools can determine the best class category a student can be placed in the Senior Secondary School.

***Keywords*- Data mining, Cross validation, Junior and Senior Secondary Schools, Nigeria**

I. INTRODUCTION

The information age has enabled many institutions to gather large volumes of data. However, the usefulness of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it. Data mining attempts to answer this need. Hence in order to get useful information from large amount of data, there is need for mining of data called Data mining. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information which can help in decision making. In order to get the required benefits from such large data and to find hidden relationships between variables, data mining techniques are developed and used [1].

Data mining can be used for extracting models describing important classes or predict future data trends. A secondary school with efficient data mining approach can use the prediction model to assign junior secondary students into various class categories (Science, Arts or Commercial) in the Senior Secondary Schools (SSS) in order to improve quality of education and increase success rate.

With the increasing population in Nigeria, the number of secondary schools in Nigeria using Oyo State as a case study has also increased; causing increase in the number of students enrolled per session. Thus, choice of class categories by most students in Junior Secondary Schools (JSS) is majorly based on parent's opinion and peer influence.

Basically Senior Secondary Schools admit students into various classes based on the marks obtained in some subjects in the JSS final examination; but the parental opinion and/or student preference can affect the placement even if the students do not meet up with the criteria. The use of data mining for prediction can help secondary schools give candid advice to parents and guardians on the right class students should be placed against the parental will or peer pressure.

In this study, data mining classification algorithms (J48, IBK, JRipper and Naïve Bayes) were applied on the final JSS score data of students in order to determine the best class category to place them at SSS level.

The rest of this paper is organized as follows. Related works are presented in Section 2 while our

Omoyele, T. D. and Akinola, S. O. (2018). A Comparative Performance Analysis of Data Mining Classification Algorithms on Predicting Students' Placement into Subject-Area Classes in Senior Secondary Schools, ***University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),*** **Vol. 2, No. 1, pp. 23 - 27**

methodology is presented in Section 3. Experimental results are presented in Section 4

II. LITERATURE REVIEW

Various researches have been done by researchers to explore data mining techniques for the prediction of data in order to extract knowledge which can be useful for decision-making. Few of the research done in the aspect of mining educational data are reviewed below:

Getaneh Berie and Vuda Sreenivasarao [4] conducted a research that involves the prediction of placement of students into different departments. Student's data with information about their entrance exam result was collected and three different classification algorithms (J48, Naïve Bayes and Random Tree) from WEKA software was applied on the data. The result of the research shows that Random Tree algorithm is the best with the type of data collected for the prediction of students' placement into different departments based on their result and preference.

Omprakash Chandrakar and Jatinderkumar Saini [5] carried out research on higher education data wherein association rule mining was applied to analyze the performance of students in their examination and predict the possible outcome in the forthcoming examination. This prediction allows students and teachers to identify the subjects which need more attention even before the commencement of semester.

Nirmala and Mallikarjuna [6] consider the number of parameters for the derivation of performance prediction indicators needed for faculty performance assessment, monitoring and evaluation. In order to predict the quality, productivity and potential of faculty across various disciplines which will enable higher level authorities to take decisions and understand certain patterns of faculty motivation, satisfaction, growth and decline. The analysis depends on encompassing student's feedback, organizational feedback, institutional support (finance), administration, research activity etc. Data mining methodology used to extract useful patterns from the institutional database was able to extract certain unidentified trends in faculty performance when assessed across several parameters.

Bayes classification algorithm was applied by Pandey U. K and Pal S. [7] on 600 students' data from different Colleges of Awadh University, Faizabad, India. Language and background qualification were found to affect the performance of students in the University.

Khan [8] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyse. It was found that girls with high socioeconomic status had relatively higher academic achievement in science stream and boys with low socioeconomic status had relatively higher academic achievement in general.

III. THE PRESENT SYSTEM

Placing JSS students into different class categories in the Senior Secondary School is not an easy task. The students choose the class category they intend to go but this has to be based on the past performance in various subjects. Teachers/counselors do advise students to choose their appropriate class categories, but this advice may be heeded or not. Such problem happened in student's placement into different classes in various schools in Nigeria. This study is mainly concerned on the prediction of the number of students that are correctly placed into different classes based on the marks obtained in various subjects in their JSS. On the other hand, the classification algorithm that performs well for predicting the number of students correctly placed in the right category can help build a model for classifying students into different class categories based on the marks obtained in various subjects.

IV. METHODOLOGY

A. Data Source

The data used for this study was obtained from both private and public secondary schools in Ibadan, Oyo state Nigeria. A total of 62 records was gotten from a government owned federal secondary school, a private secondary school supplied us with 34 records, another private secondary school supplied us with 23 records. The data set is a nominal data set which contains nine columns and 126 rows (data instances), the columns are identification column, English, Civic Education, Social Studies, Mathematics, Integrated Science, Fine Art, Business Studies, Introductory Technology and Remarks (which indicate the various class categories the students were placed). Figure 1 shows the sample data set.

ID	ENGLISH	CIVIC	S/STUDIES	MATHS	INT SCI	FINE ART	BSTUDIES	INTRO TE	REMARK
ai1	68.67	75.33	68	69.67	57.83	62.67	68.33	60.33	COMMERCIAL
ai2	64.67	60.33	60.67	53	57.33	58.67	58.67	53.67	ART
ai3	66.67	81.33	71.33	63	64.5	65	73.67	61.33	COMMERCIAL
ai4	62	71	73	62.67	57.67	58.67	65.67	46.33	COMMERCIAL
ai5	66.33	69.33	67.33	53	60.5	56.33	62.67	56.67	COMMERCIAL
ai6	63.67	80	73.67	62.33	59.67	58.67	70	51.33	COMMERCIAL
ai7	74.33	81.67	82	74	66.67	61	79.33	61.33	COMMERCIAL
ai8	58.67	76.67	65.33	65	58.83	53	60.67	57.67	COMMERCIAL
ai9	69.33	71	71.67	79.67	71.67	55.67	70.33	60.33	SCIENCE
ai10	71.33	93.67	86	79.67	78.33	68.67	74.67	65.33	SCIENCE
ai11	74	83	81.33	87.33	77.33	72	79.67	72.67	SCIENCE
ai12	84	82.67	83.33	81.33	81.33	67.33	79	71.33	SCIENCE
ai13	74	81	72.67	79	73.17	67.33	75.33	71	SCIENCE
ai14	68.67	52	48.33	83	68.17	38	72.33	58.33	COMMERCIAL

Figure 1: Sample Dataset

B. Data Preparation

After the data was collected from the various secondary schools, processing of the original data was carried out so that the data can be suitable for use with the data mining software used in this study. The data of the students include scores in English Language, Civic Education, Social Studies, Mathematics, Integrated Science, Fine Art, Business Studies and Introductory Technology. The scores in all these subjects were used for this experiment to discover the number of students in the Senior Secondary Schools that are in the right placement based on the marks obtained in English Language, Civic Education, Social Studies, Mathematics, Integrated Science, Fine Art, Business Studies and Introductory Technology. The student's name or other means of identification was intentionally left out because of privacy of individual sensitive data. After the data was collected, noisy data were removed, missing data were accounted for and the data file was saved in CSV format in order for it to be acceptable by Waikato Environment for Knowledge Analysis (WEKA) software. To partition the dataset into training and testing data sets, cross validation was used.

C. Tools and Techniques

The tools used in this research work are Microsoft Excel (which was used for data preprocessing) and WEKA software (this contains the classification algorithms that were used for this study). The WEKA is a collection of machine learning algorithms for data mining tasks written in Java. The algorithms in WEKA can either be applied directly or called from its own Java code. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. The classification techniques of data mining help to classify the data on the basis of certain rules [2]. For the purpose of this study, classification algorithms such as J48, IBK, JRipper, and Naïve Bayes algorithms were used to discover the number of students that are in the right classes based on the fulfillment of the conditions of the rules generated by the algorithms.

V. RESULTS AND DISCUSSIONS

Classification is 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 [2]. This study attempt to discover the number of students that were correctly placed into various departments based on the different subjects passed in the junior secondary schools. The classification algorithms selected for this research work are JRipper, IBK, J48 and Naïve Bayes. The data has been experimented by cross validation with the total data set using 10 fold. To build the predictive model, the arff format of the selected dataset was given to Weka and one experiment was performed for each selected algorithms. Summary of the results is presented in Table 1.

Table 1: Summary of the result

ALGORITHM	CORRECTLY PLACED	INCORRECTLY PLACED	PRECISION	DEPARTMENTS
J48	68	4	0.971	COMMERCIAL
	29	2	0.879	SCIENCE
	23	0	1.000	ART
JRIPPER	67	5	0.944	COMMERCIAL
	27	4	0.844	SCIENCE
	23	0	1.000	ART
IBK	68	4	0.919	COMMERCIAL
	26	5	0.867	SCIENCE
	22	1	1.000	ART
NAIVE BAYES	63	9	0.913	COMMERCIAL
	28	3	0.757	SCIENCE
	20	3	1.000	ART

5.1 J48 Decision Rules

```
B/STUDIES <= 52.33: ART (20.0)
B/STUDIES > 52.33
|  INTRO TECH <= 60
|  |  S/STUDIES <= 51.666667: ART (3.0)
|  |  S/STUDIES > 51.666667: COMMERCIAL (53.0)
|  INTRO TECH > 60
|  |  MATHS <= 65: COMMERCIAL (16.0)
|  |  MATHS > 65
|  |  |  INT SCI <= 60.333333: COMMERCIAL (2.0)
|  |  |  INT SCI > 60.333333: SCIENCE (32.0/1.0).
```

The result above shows that social studies is a required subject to be passed for both art and commercial department student, while Mathematics is required to be passed also for both science and commercial department student.

Results show that J48 algorithm correctly classified 120 of the 126 instances. The student whose marks obtained fulfil the condition of the rule generated are

68, 29, 23 in Commercial, Science and Art respectively. It shows that:

- For students in Commercial class, it is only the marks of 68 of them that fulfil the condition of the rule to be in the class
- For Science, the scores of 29 students fulfil the condition to be in science, based on the rule generated and the scores of 23 students in Art class fulfil the condition generated by the rule.

5.2 JRIPPER

(B/STUDIES <= 51.67) => REMARKS=ART (19.0/0.0)
(CIVIC EDUCATION <= 57.33) => REMARKS=ART (4.0/0.0)
(INTRO TECH >= 64.67) and (MATHS >= 68.33) => REMARKS=SCIENCE (19.0/0.0)
(ENGLISH >= 72) and (B/STUDIES >= 72.67) => REMARKS=SCIENCE (9.0/1.0)

The result above shows that business studies and civic education are required subject to be passed for art department student, while Mathematics is required to be passed also for both science and commercial department student.
From the results obtained, JRipper algorithm correctly classified 117 out of the 126 instances. The students whose marks obtained fulfil the condition based of the rule generated are 67, 27, 23 in Commercial, Science and Art respectively. It shows that:

- For students in Commercial class, it is only the marks of 67 of them that fulfil the condition of the rule generated to be in Commercial class
- For Science, the scores of 27 students fulfil the condition to be in Science, based on the rule generated and the scores of 23 students in Art class fulfil the condition generated by the rule.

5.3 IBK

Results show that IBK correctly classified 68 out of the 126 instances in Commercial class while for Science class, 26 students were correctly placed. For Art class, 22 of the students were correctly placed; all this with a total precision of 0.921.

5.4 Naive Bayes

It is observed that Naïve Bayes correctly classified 63 out of the 126 instances in Commercial Class, based on the marks obtained in various subjects in JSS while for Science, 28 students were correctly placed. For Art class, 23 of the student were correctly placed based on the marks obtained in their various subjects in the JSS. All these are predicted with a total precision of 0.890.

- Thus, from the results, J48 algorithm with precision accuracy of 95.4% using 126 attributes is more appropriate to build the predicting model for discovering the number of students in Senior Secondary Schools that were correctly placed in various SSS Classes. On the other hand, Naïve Bayes algorithm performed less in accuracy.
- Therefore, the results obtained from the study show that the best prediction model for discovering the number of incorrectly placed students in various classes in secondary school is J48 algorithm.

VI. Conclusion

This study employed data mining techniques to predict the number of students that are correctly placed into various classes in Senior Secondary Schools based on the marks obtained in various subjects in the Junior Secondary School. In this work, we applied four classification algorithms on students' data i.e. JRIPPER, J48, IBK, and Naïve Bayes.

From the results obtained from the experiments J48 algorithm performs best with the precision accuracy of 95.4%, followed by JRIPPER with precision accuracy of 92.9%, IBK with precision accuracy of 92.1% and Naïve Bayes with precision accuracy 89.0% in that order.

It is hereby concluded that J48 algorithm is the best suitable for predicting the number of students to be placed correctly into various class categories at the Senior Secondary School. This study demonstrates that marks obtained in various courses by Junior Secondary School students can be used in data mining tasks to determine the best class categories the students can be placed in Senior Secondary School.

This study can however be extended by employing more data from various school categories (private and government based) so as to obtain accurate predictions.

REFERENCES

[1] Samrat S. and Vikesh K. (2013). Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques, *IJCSET*, Vol 3, Issue 2, 31-37.

[2] Ramesh V., Parkavi P. and Yasodha P. (2011). Performance Analysis of Data Mining Techniques for Placement Chance Prediction, *International Journal of Scientific & Engineering Research* Volume 2, Issue 8, http://www.ijser.org/researchpaper/Performance-Analysis-of-Data-Mining-Techniques-for-Placement-Chance-Prediction.pdf

[3] Samrat S. and Vikesh K. (2012). Classification of Student's Data Using Data Mining Techniques for Training & Placement Department in Technical Education, *International Journal of Computer Science and Network (IJCSN)* Volume 1, Issue 4, http://ijcsn.org/IJCSN-2012/1-4/IJCSN-2012-1-4-63.pdf

[4] Getaneh B.T. and Vuda S. (2016). Application of Data Mining Techniques to Predict Students Placement into Departments, *International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)* Volume 3, Issue 2, pp. 10-14.

[5] Omprakash C. and Jatinderkumar R. S. (2015). Predicting Examination Results using Association Rule Mining. *International Journal of Computer Applications* (0975 – 8887) Volume 116 – No. 1.

[6] Nirmala G. and Mallikarjuna P. B. (2014). Faculty Performance Evaluation Using Data Mining, International *Journal of Advanced Research in Computer Science & Technology* (IJARCST 2014) Vol. 2, Issue 3.

[7] Pandey U. K. and Pal S. (2011). Data Mining: A prediction of performer or underperformer using classification, *International Journal of Computer Science and Information Technology*, Vol. 2(2), pp.686-690, ISSN:0975-9646.

[8] Khan Z. N. (2005). Scholastic achievement of higher secondary students in science stream, *Journal of Social Sciences*, Vol. 1, No. 2, pp. 84-87.