Scientific and Practical Cyber Security Journal (SPCSJ) 3(2): 18- 35 ISSN 2587-4667
Scientific Cyber Security Association (SCSA)

Using Data Mining Classifiers to Predict Academic Performance of High School Students

Yecheng Yao, The University of Chicago (Chicago, USA)
Zebang Chen, The University of California, San Diego (San Diego, USA)
Sumin Byun, Hankuk Academy of Foreign Studies (Yongin, Korea)
Yizhu Liu, Pius XI Catholic High School (Milwaukee, USA)

ABSTRACT: The use of data mining techniques for educational datasets is being referred to as educational data mining. This study uses popular classifier algorithms in data mining with secondary school student data to estimate their success rate. Student success depends on various factors related to the student's personal, family and surrounding environment, among others. This study's dataset has attributes related to parental education, job information, student travel time, study time, financial status, extracurricular activities, access to the Internet, family relationship, alcoholic consumption, student health condition, regular school attendance. This study analyzes the correlations between these attributes and identifies the attributes that contribute to students' test achievement for better prediction and management of student performance. The study also compares the performance of top classification algorithms in data mining and concludes J48 classifier and oneR to outperform the other classifiers.

KEYWORDS: Data Mining, Educational Data Mining, Classifier, High School Data Mining

I. Introduction

Data mining entails extracting knowledge from large volumes of data. This work attempts to uncover insightful patterns and relationships useful for the decision-making process. Data from various sources are processed using various methods and algorithms to uncover useful patterns and insights. Data mining and knowledge discovery have gained increasing attention for their usefulness for decision making and now have become an essential part of any organization, including educational institutions. Schools and teachers need to know their students' academic achievement in terms of what factors influence that achievement. To better estimate of student performance, data mining algorithms and statistical methods have been used for analysis purposes.

The application of data mining to educational datasets is referred to as educational data mining (EDM), which uses data mining techniques, machine learning methods, and statistical analysis methods. EDM is expected to innovate novel methods to mine educational data for useful insights on student learning and achievement and the factors that influence them.

Educational data mining focuses on analysing students' academic data sets, classifying students using decision trees, generating some association rules for better decisions for enhancing

academic performance. The main goal is to predict students' future learning and academic performance, and this goal depends on identifying all attributes related to learning characteristics and behaviors. Another goal is to identify students' interests so that the curriculum could be planned to accommodate their educational standards and learning styles.

The rest of this paper is organized as follows: Section II provides a literature review. Section III discusses the artificial neural network and the multi-layer perceptron, and Section IV provides the materials and methods. Section V discusses the results, and Section VI concludes.

II. Literature Review

Saibaba et al. [1] analyzed student performance using various clustering techniques with the WEKA tool. The proposed system considered student data with 10^{th} percentage, intermediate percentage, and B.Tech I Year, II Year and III Year marks using decision trees. They analyzed these decision trees and forecasted the likelihood of students getting jobs after graduation. They also proposed to consider attributes such as social networking interests, parental economic status, and parental educational achievement.

Ahmed et al. [2] compared four data mining techniques including J48 decision trees, MLP, NB, and SMO and found that the attribute evaluation method to be helpful in predicting instructor performance. The results for the attributed evaluation method show that SMO outperformed other algorithms with 85.5% accuracy. J48 DT outperformed other algorithms with 84.8% accuracy.

Kalpana and Venkatalakshmi [3] presented a case study on educational data mining and discussed how data mining can be used in higher education for graduate student performance. They used engineering students' data for 5 years and applied data mining techniques, using various clustering methods along with distance- and density-based approaches to improve prediction accuracy for graduate student performance.

Brijesh Kumar and Saurabh [4] used various data mining techniques to achieve high quality in the higher education system by extracting insights into students' exam performance. They used a dataset from 50 students from VBS Purvanchal University (Jaunpur of MCA course) from 2007 to 2010.

Kalpesh et al. [5] proposed an architecture to more accurately predict student performance. They considered 17 student attributes using K-means clustering and Naïve-Bayes algorithms and concluded that Naïve Bayes outperformed with 96% accuracy.

Taier et al. [6] discussed the effectiveness of data mining in improving the performance of graduate students by using data mining techniques such as association rules, classification, clustering, and outlier detection.

Cortez et al. [7] proposed a method for predicting student grades by analyzing past grades and found that school, family, and social environments to be key factors. They employed binary and five-level classifications and regression using decision tree, random forest, neural network, and support vector machine.

Jovanovic et al. [8] used classification techniques to analyze and predict student academic performance and applied clustering to student groups based on e-learning, concluding that their model can help identify students' academic strength.

Pandey and Pal [9] examined student performance using 600 students from Dr. R.M.L. Awadh University, Faizabad, India, and used Bayes classification in category, language and background qualifications to predict new students' academic performance.

Hijazi and Naqvi [10] analyzed student performance using 300 students (225 males and 75 females) from Punjab University of Pakistan and considered student attributes including student attitudes towards class attendance, student family income, maternal age, maternal education level, and hours spent studying. They used simple linear regression and concluded maternal education level and student family income to be highly correlated with academic performance.

Khan [11] conducted a performance analysis of 400 students composed of 200 males and 200 females from Senior Secondary School of Aligarh Muslim University, Aligarh. These students were selected using cluster sampling where the whole population was divided into groups or clusters. They concluded that females with high socioeconomic status showed higher science achievement and males of low socioeconomic status, higher academic achievement in general.

Pandey and Pal [12] conducted a performance analysis of 60 students from Dr. R. M. L. Awadh University, Faiza bad, India, using association rule mining and found students' interest in class teaching language.

Bray [13] examined private tutoring and observed the percentage of students receiving private tutoring in India to be higher than that in Malaysia, Singapore, Japan, China, and Sri Lanka. They also observed enhanced academic performance through private tutoring.

Ayesh et al. [14] used the K-means clustering algorithm to predict students' learning activities and concluded the model to be useful in predicting academic performance.

Al-Radaideh et al. [15] used the decision tree model to predict final grades of students taking a C++ course at Yarmouk University, Jordan. They compared three classification techniques including ID3, C4.5 and Naïve Bayes and concluded the decision tree model to outperform other models.

III. Materials and Methods

Dataset:

This study considers data on student achievement in secondary education of two Portuguese schools. Data attributes include student grades, demographic, social, and school-related factors, and the data were collected using school reports and questionnaires. Two datasets focused on two distinct subjects: Mathematics (mat) and Portuguese language (por). These two datasets were combined. Here the target attribute G3 has a strong correlation with attributes G2 and G1, which is due to G3 being the final year grade (issued in the 3rd period), while G1 and G2 correspond to 1st and 2nd period grades. Predicting G3 without G2 and G1 is more difficult, but this kind of prediction is considered to be much more useful.

Table 1 shows the attributes.

Table 1. Various attributes

Attribute	Description (Domain)
Sex	student's sex (binary: female or male)
Age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4a)

Mjob	mother's job (nominalb)
Fedu	father's education (numeric: from 0 to 4a)
Fjob	father's job (nominalb)
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10
failures	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
Gout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
Health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first-period grade (numeric: from 0 to 20)
G2	second-period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Table 1 provides the following insights:

i. There are 1044 instances and 33 attributes for both Mathematics (mat) and Portuguese language (por) datasets.

ii. G3 is the output label (all 32 attributes other than G3 are independent variables for the dependent variable G3).

iii. G3 has a range of [0, 20], and the classification model must predict 1 class out of the possible class labels.

iv. There is a mix of numerical and nominal attributes. There are no missing values for any given attributes.

IV. Results & Discussion

The downloaded dataset was in .csv format, and this was converted into.arff to accommodate the WEKA environment. The converted .arff file was given as input for the WEKA explorer, as shown as follows:

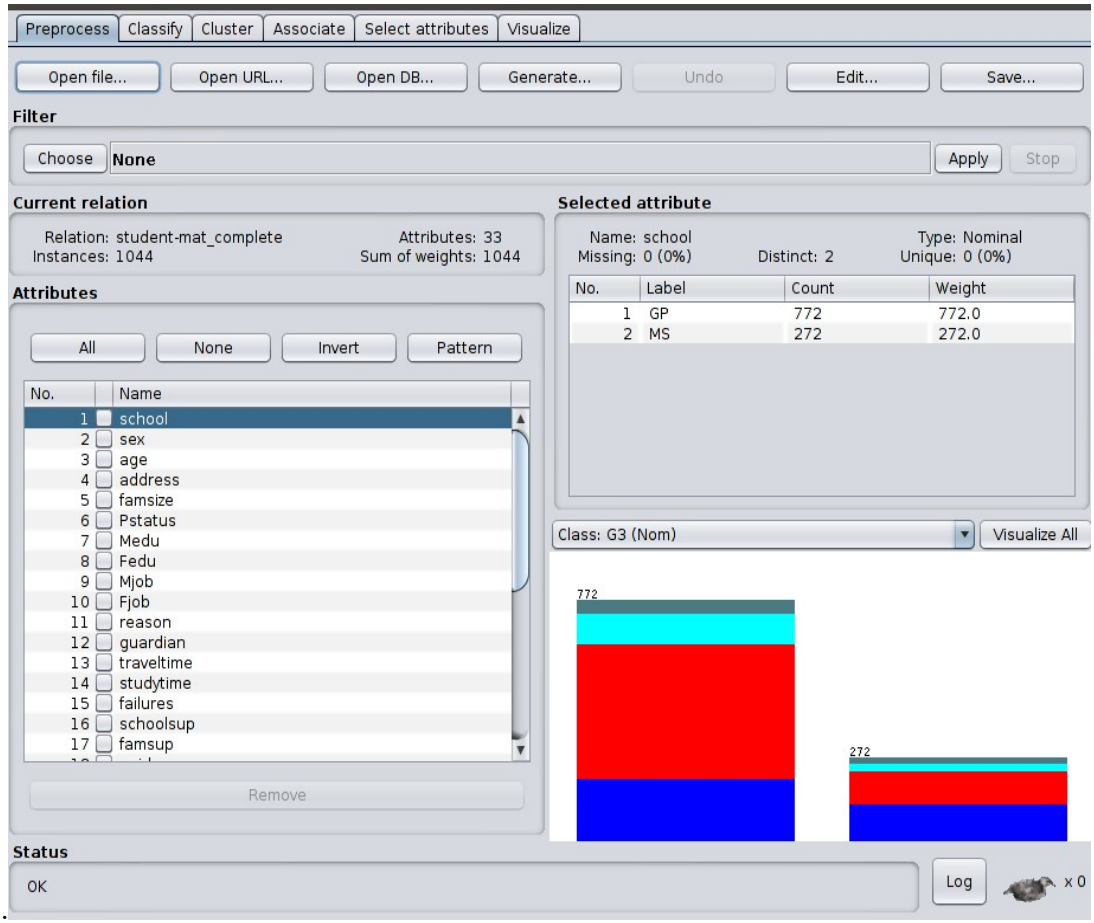


Figure 1. Student performance dataset on WEKA

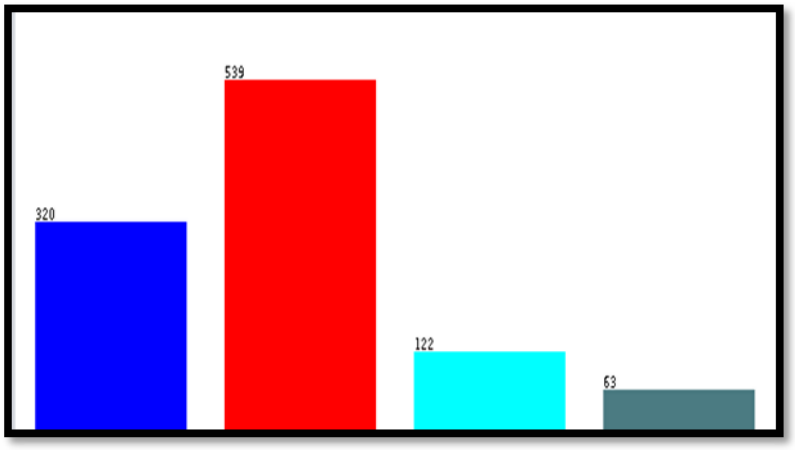
The experiment was conducted in three stages: data preprocessing, data classification without feature selection, and data classification with feature selection.

1. **Data Preprocessing:**

In the data preprocessing phase, target attributes are identified, and the dataset is loaded into WEKA for cleaning and preprocessing. The target attribute is categorized into 4 classes A, B, C, D, as shown in Table 2.

Table 2. Mapping of clusters based upon initial values

The range of initial class	**New Class Label**
0 ~ 5	D
6 ~ 10	C
11 ~ 15	B
16 ~ 20	A



2. **Implement of classifiers on full dataset**

Classifiers	**Correctly Classified Instances (%)**	**In-correctly Classified Instances (%)**	**Kappa statistic**	**Mean absolute error**	**Root mean square error**
ZeroR (baseline)	51.63	48.37	0	0.3114	0.3944
OneR	**84.19**	15.81	0.7421	0.079	0.2811
J48	81.32	18.68	0.6971	0.1171	0.2839
Naive bayes	76.53	23.47	0.6332	0.1303	0.2964
IBk (k-nearest neighbor)	45.98	54.02	0.1283	0.2706	0.5186
Decision Tree	83.81	16.19	0.7351	0.1388	0.2477
Logistic Regression	81.61	18.39	0.7023	0.1125	0.2663
PART	78.83	21.17	0.6587	0.1135	0.3133
RandomForest	82.38	17.62	0.7099	0.1618	0.2615
JRip	81.90	18.10	0.7113	0.1273	0.2715

Table 3. Comparison of results for different classifiers

Table 3 shows the results for different classifiers for correctly classified instances, incorrectly classified instances, Kappa statistic, mean absolute error, and root mean square error. Using the above method, OneR shows a higher accuracy value of 84.19, while ZeroR (baseline) shows the lowest value of 51.63. Figures 3 and 4 provide the graphical representation of results for different classifiers.

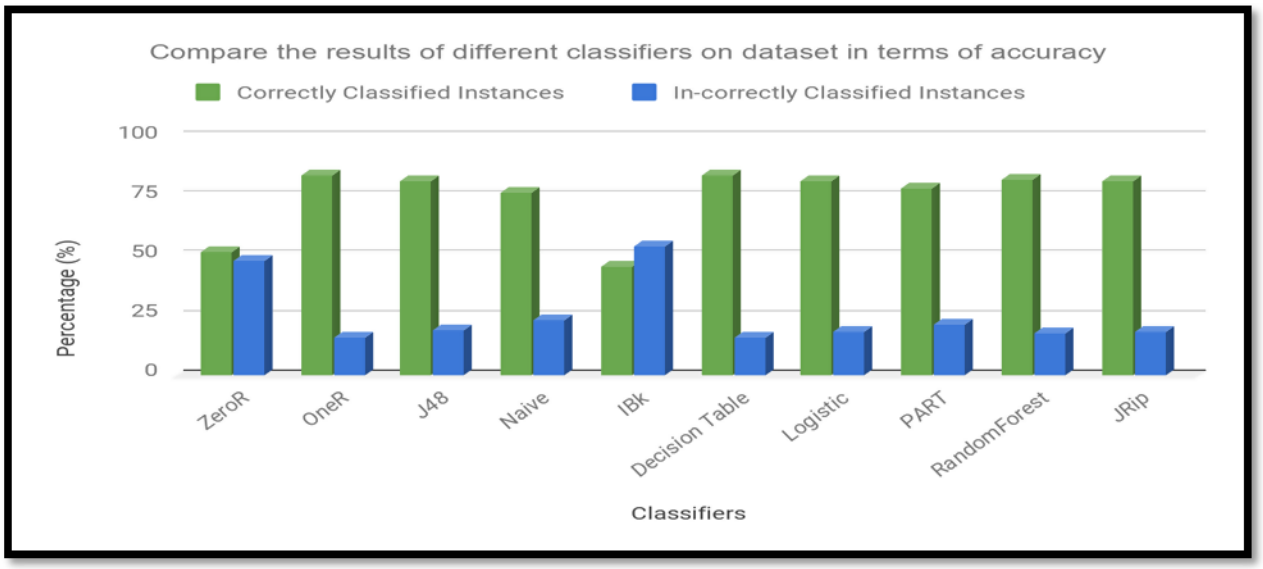


Figure 3. Graphical representation of different classifiers

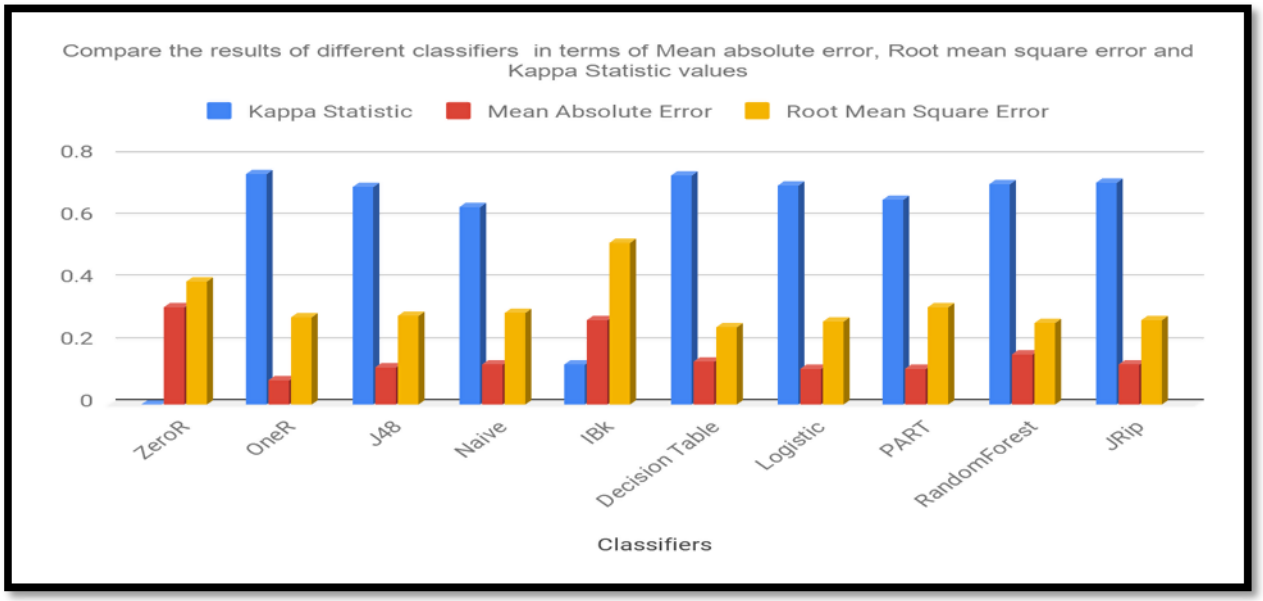


Figure 4. A comparison of accuracy for MASE, RMSEs, Kappa statistic

Table 4. A comparison of different classifiers for execution time

Classifiers	**Time taken to build a model (in a sec)**
ZeroR	0.001
OneR	0.01

J48	0.14
Naive Bayes	0.03
IBk	0.001
Decision Table	0.38
Logistic	0.54
PART	0.19
RandomForest	0.47
JRip	0.31

Table 4 compares different classifiers in terms of execution time (seconds) without using feature selection for the given dataset. ZeroR and IBk (k-nearest neighbor) take the shortest time, and logistic classifiers take the longest execution time. Figure 5 graphically shows the execution time for different classifiers.

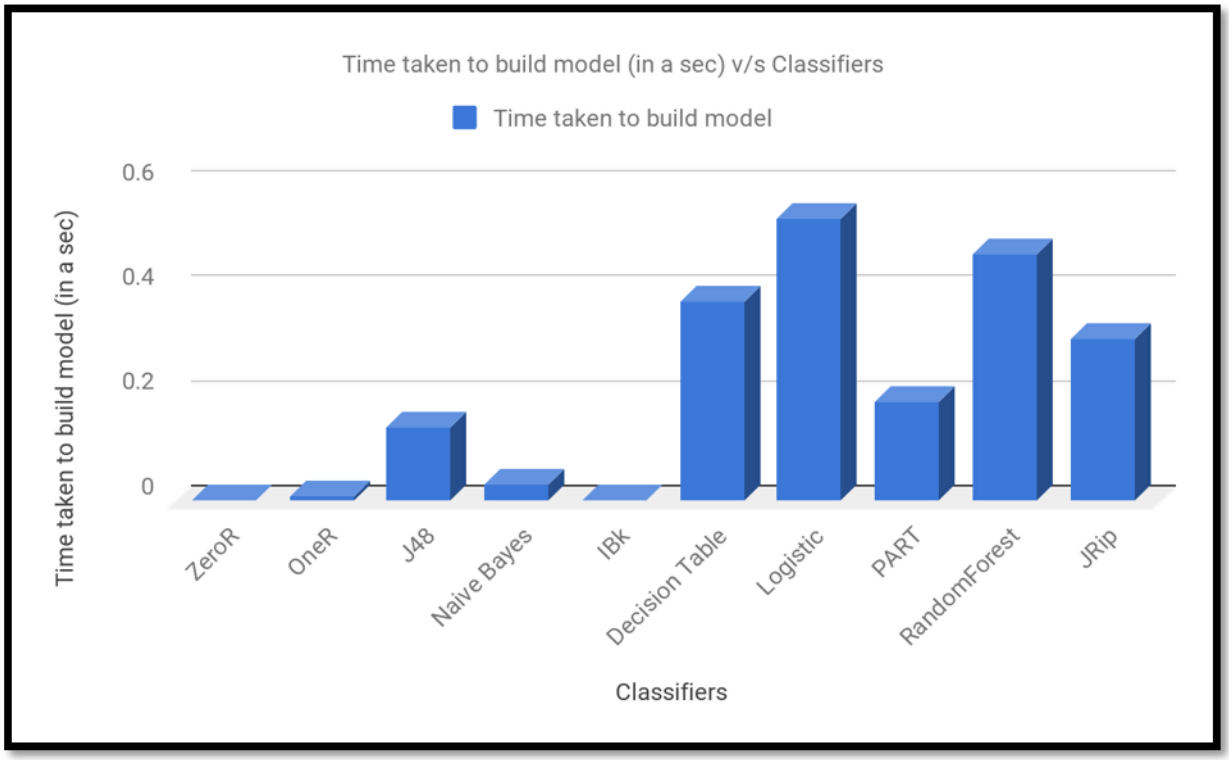


Figure 5. Time taken to build the classifier

3. Feature Selection

Feature selection is used to select attributes from the dataset. The feature evaluator and search method was employed to perform feature selection, and based on selected attributes techniques, the following 6 features/attributes were selected:

(i) famsize
(ii) Fedu
(iii) failures
(iv) absences
(v) G1 and
(vi) G2 for better results from the dataset

After selecting these attributes, the dataset was prepared to implement different classifiers to compare with and without feature selection method.

Table 5. A comparison of results for different classifiers for feature selection

Classifiers	Correctly	In-correctly	Kappa	Mean	Root mean

	Classified Instances (%)	**Classified Instances (%)**	**statistic**	**absolute error**	**square error**
ZeroR (baseline)	51.63	48.37	0	0.3114	0.3944
OneR	84.19	15.81	0.7421	0.079	0.2811
J48	**84.39**	15.61	0.7465	0.1142	0.2515
Naive bayes	81.23	18.77	0.7026	0.1197	0.2672
IBk (k-nearest neighbor)	76.53	23.47	0.6185	0.1193	0.3316
Decision Tree	84.20	15.80	0.7434	0.1353	0.2429
Logistic Regression	84.10	15.90	0.7403	0.1147	0.2413
PART	81.03	18.97	0.6948	0.1153	0.2728
RandomForest	80.84	19.16	0.6912	0.1153	0.263
JRip	83.24	16.76	0.7305	0.1271	0.263

Table 5 shows the results for different classifiers with the feature selection method for correctly classified instances, incorrectly classified instances, Kappa statistic, mean absolute error, and root mean square error. With this method, J48 shows a higher accuracy value of 84.39, and ZeroR (baseline), the lowest value, 51.63. Figures 6 and 7 graphically represent different classifiers results.

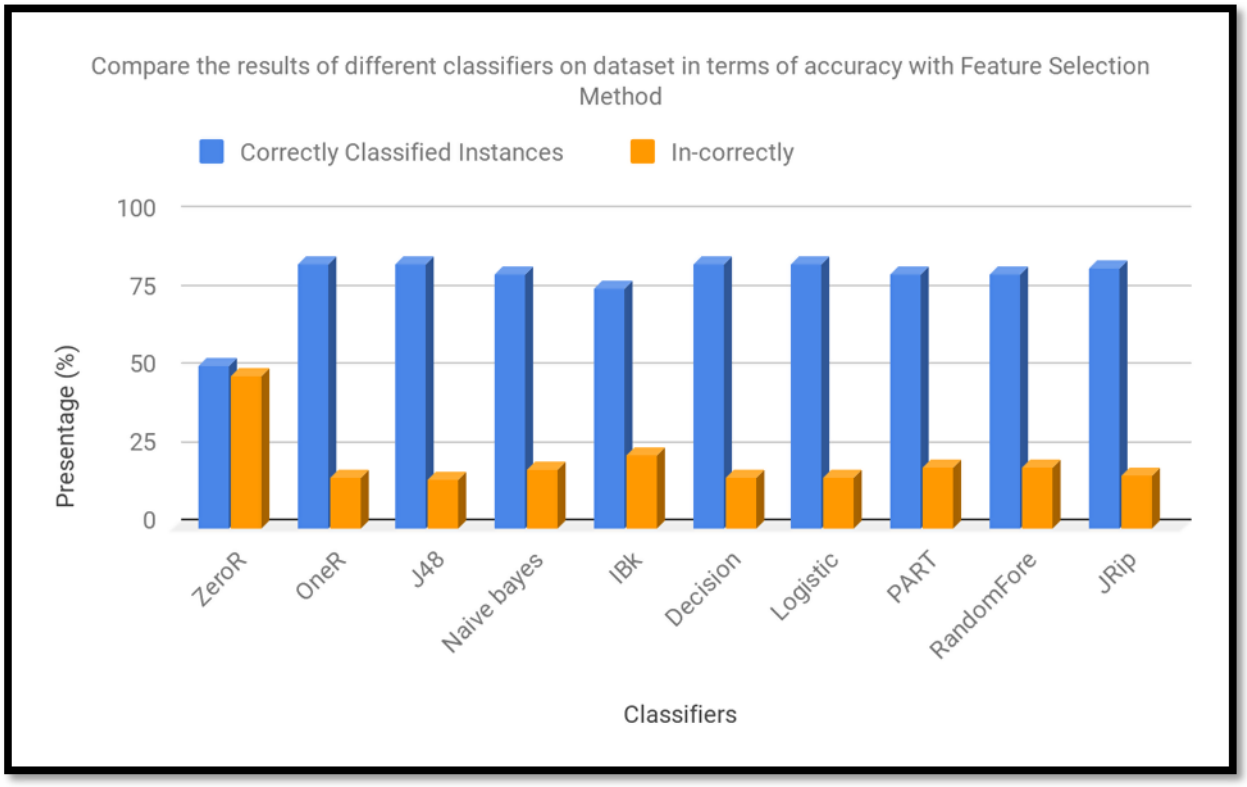


Figure 6. Number of correctly and incorrectly classified instances

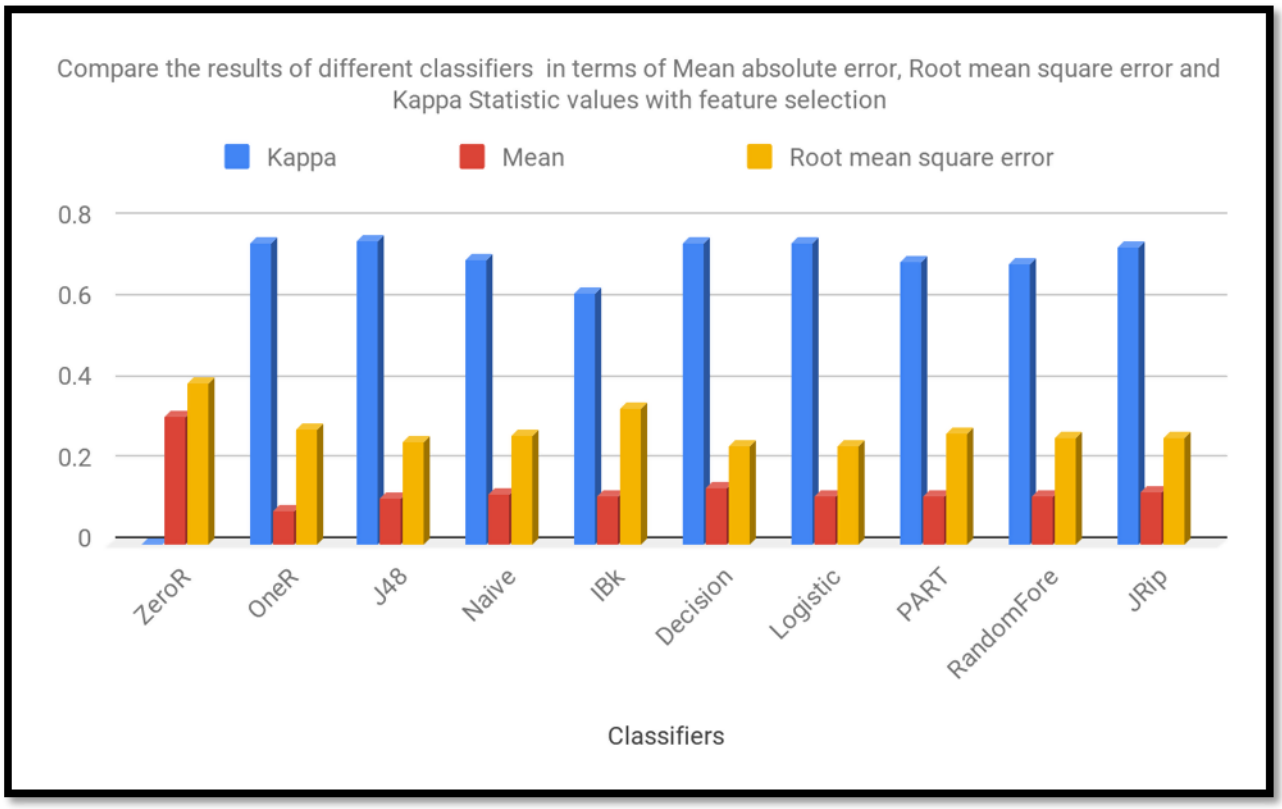


Figure 7. Accuracy for different classifiers

Table 6 compares different classifiers for execution time (seconds) with and without feature selection. Table 6 shows that ZeroR and IBk (k-nearest neighbor) take the shortest time with and without feature selection. The longest time is taken by Logistic classifiers without feature selection, and RandomForest takes the longest time with feature selection.

Table 6. A comparison of different classifiers for execution time

Classifiers	**Time taken to build model (in a sec)**	**Time taken to build model with feature Selection (in a sec)**
ZeroR	0.001	0.001
OneR	0.01	0.01
J48	0.14	0.11
Naive Bayes	0.03	0.02
IBk	0.001	0.001
Decision Tree	0.38	0.10
Logistic Regression	0.54	0.11
PART	0.19	0.06
RandomForest	0.47	0.29
JRip	0.31	0.18

Figure 8 graphically shows different classifiers for execution time using both methods.

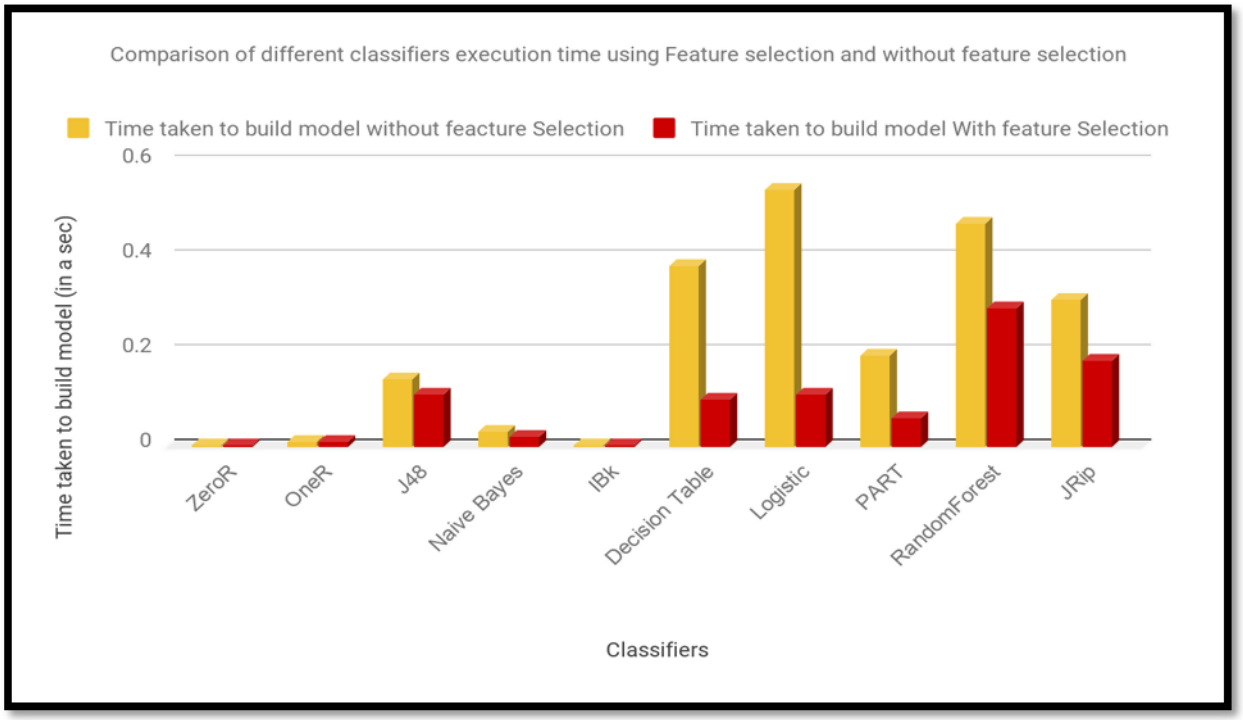


Figure 8. Time taken to build a model with and without feature Selection

Table 7 shows the results for different classifiers for accuracy with and without feature selection. The highest accuracy is shown for J48 with feature selection.

Table 7. A comparison of different classifiers for accuracy

Classifiers	**Correctly Classified Instances with feature selection (%)**	**Correctly Classified Instances without Feature Selection (%)**
ZeroR	51.63	51.63
OneR	84.19	84.19
J48	**84.39**	81.32
Naive bayes	81.23	76.53
IBk	76.53	45.98
Decision Tree	84.2	83.81
Logistic Regression	84.1	81.61
PART	81.03	78.83

RandomForest	80.84	82.38
JRip	83.24	81.9

Figure 9 graphically shows the results for different classifiers for accuracy (correctly classified instances).

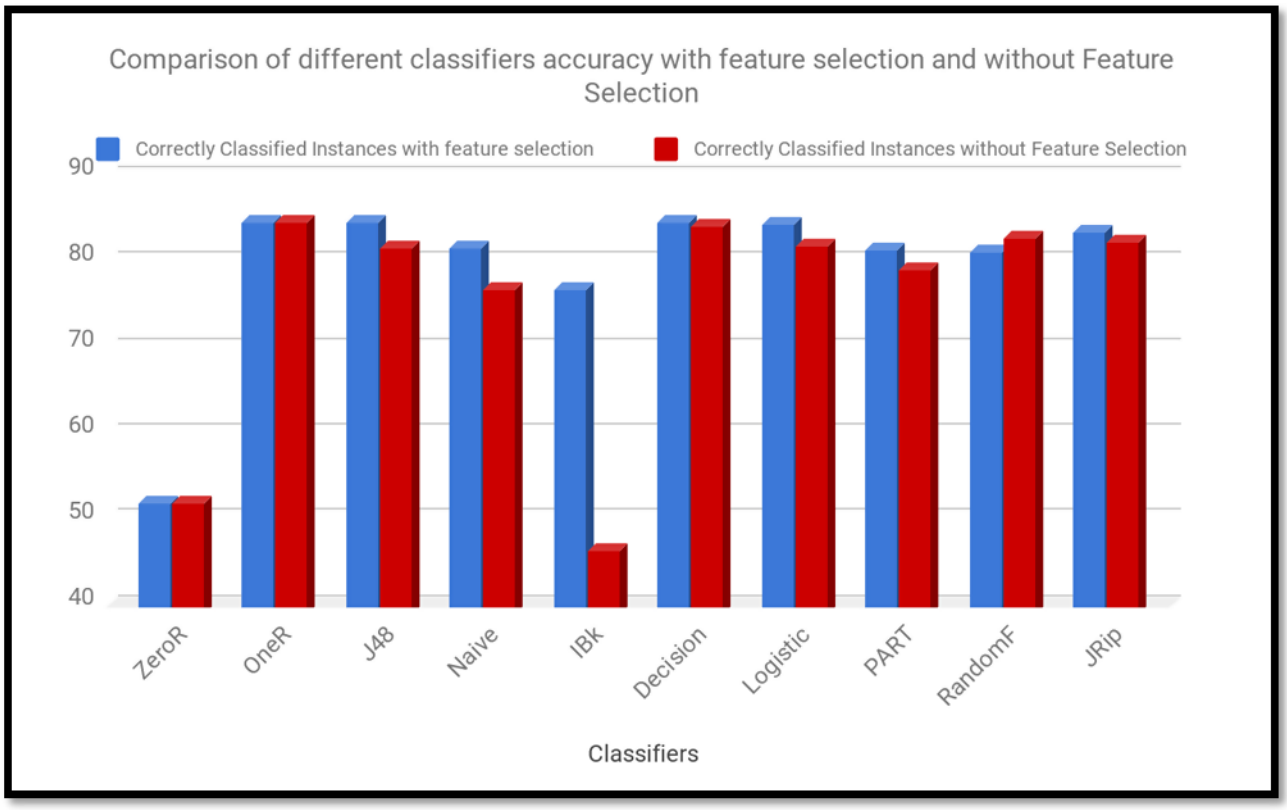


Figure 9. A comparison of different classifiers

V. Conclusions

This paper mines data from high school students to predict their academic performance. The dataset included relevant educational and personal attributes. Top 10 classifiers were identified based on previous research. The results show that without feature selection, oneR provided the best performance, followed by the decision tree, random forest and J48. The results were tabulated, and corresponding plots were obtained. Then the feature selection method was applied to identify famsize, fedu, failures, absence, G1 and G2 as the required attributes. According to the results, J48 showed the best performance with fewer attributes, followed by oneR, decision tree and logistic regression. The tabulated and plotted results suggest that J48 and oneR provided good results for classifying the dataset and predicting student performance.

References

[1] Sai Baba et al, "Student Performance Analysis Using Classification Techniques", International Journal of Pure and Applied Mathematics, Vol.115, No.5 2017, pp.1-7

[2] Mohamed Ahmed et al., "Using Data Mining to Predict Instructor Performance", 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, pp. 137-142

[3] Kalpana and Venkatalakshmi, "Intellectual Performance Analysis of Students by Using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technology, Vol.3, 2014.

[4] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications", Vol. 2, No. 6, 2011.

[5] Kalpesh et al., "Student Performance Prediction System using Data Mining Approach", International Journal of Advanced Research in Computer and Communication Engineering, Vol.6, Issue 3, 2017.

[6] Tair and El-Halees 2012. Mining Educational Data to Improve Student's performance: A Case Study. International Journal of Information and Communication Technology Research.

[7] P. Cortex and A. Silva. Using data mining to predict secondary school student performance.

[8] M. Jovanovic, M. Vukicevic, M. Milovanovic and M. Minovic (2012). Using data mining on student behaviour and cognitive style data for improving e-learning systems: a case study. International Journal of Computational Intelligence Systems.

[9] U. K. Pandey and S. Pal, Data Mining: A Prediction of performer or under performer using classification , (IJCSIT), International Journal of Computer Science and Information Technology, Vol 2(2), pp. 686-690.

[10] S. T. Hijazi and R.S.M.M. Naqvi, "Factors affecting students performance: A case of Private Colleges". Bangladesh e-Journal of Sociology, Vol. 3 No.1, 2006.

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

[12] U. K. Pandey and S. Pal , " A Data mining view on class room teaching language", (IJCSI), International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, 2011.

[13] M. Bray, The Shadow education system : Private tutoring and its implications for planners, (2nd edition), UNESCO, PARIS, France, 2007.

[14] Sheela Ayesha et. al , "Data mining model for higher education system", European Journal of Scientific Research", Vol. 43, No.1. pp. 24-29, 2010.

[15] Q. A. AI-Radaideh, et. al., " Mining Student data using decision trees", International Arab Conference on Information Technology (ACIT2006), Yarmouk University, Jordan, 2006.