KALBOARD 360 STUDENT PERFORMANCE PREDICTION AND EVALUATION BY MACHINE LEARNING

Bhanu Sreekar Reddy Karumuri
Master of Science in Computer Science
North Carolina State University, Raleigh, NC
+1 919-780-1715
bkarumu@ncsu.edu

ABSTRACT

Evaluating and Improving student performance through data mining attracted the attention of many researchers. This led to the development of the research community of "Educational Data Mining(EDM)". Several Papers have come up with analyzing various attributes or factors important in understanding and improving academic performance of students. In this study, we are using a large and feature rich educational dataset which is collected from learning management system(LMS) called Kalboard 360. Kalboard 360 is a multi-agent learning management system which facilitates learning by the use of leading-edge technology. We develop models using the dataset and then evaluate the student performance. The results show that support vector machine is the best predictor with 96.0% accuracy, followed by decision tree with 93.4% accuracy. Naïve Bayes is least accurate with accuracy of 83.3%.

KEYWORDS

Student Performance, Educational Data Mining, Machine Learning, Performance Evaluation

INTRODUCTION

In this fast-growing business world data mining plays a very important role. It helps in decision making process. It became a part in every domain from medical to aerospace. Data mining has very significant applications in educational domain. It is called Educational data mining. Educational data mining is a process of extracting significant patterns from educational databases. It helps instructors, educational institutions predict, improve and evaluate students' academic status. Students can improve learning activities allowing the administration to improve systems performance. Understanding the factors that lead to success or failure of a student is a challenging problem and attracted the curiosity of many researchers. Students performance is hard to define. In this paper we will predict students' performance based on 17 factors. The dataset is obtained from a learning management system called KalBoard 360.

LITERATURE REVIEW

Attributes and Prediction methods are main factors in prediction of student performance. The primary goal is to select the most significant attributes in predicting student performance and next the prediction methods used in predicting student performance. Most of the previous research i.e one-third of the papers have used cumulative grade point average(CGPA) as main attribute to predict students' performance. Other previous studies correlated between parenting styles and academic performance (Attaway and Bry 2004; Steinberg, Lamborn, Darling, Mounts, and Dornbusch, 1994). Some other studies focused on correlating academic performance to socio-economic status of famiy (Goddard, Sweetland and Hoy 2000), impact of teacher aid (Gerber and Fin, 2001). Other previous studies focused on investigating the importance of school types (Carpenter, 1985) in academic success. Some other studies focused on psychological side, they considered the perception of personal control (Stipek, 1981), locus of control (Bain, Boersma, and Chapmen, 1983), study behavior, engage time etc. Other studies considered the demographics like gender (Kelly 1993), age, disability etc. The reason gender became most important is that study by Meit et al (2007) found that female students have positive learning styles , they are more disciplined, self-directed and focused compared to male students. Some studies showed positive correlation to academic performance but were based on limited data.

Some researchers proved that family income has a strong correlation to academic performance (Carneiro, 2008; Yenilmez and Duman, 2008). Some other researches argued that family income indirectly affects academic performance as the families with high income send their children to private schools and provide them with additional tutoring which reflects in their academic performance (Davis-Kean, 2005). Some researchers considered social interaction network, extracurricular activities in predicting the academic performance of students and the studies showed positive correlation to academic performance.

DESCRIPTION OF DATASET

The educational dataset used in this study is collected from learning management system (LMS) KalBoard 360. KalBoard 360 is a multi-agent learning management system. This system is designed to use leading-edge technology to facilitate learning. This system provides synchronous access to educational resources from any device with internet connection. The data is collected from the learner activity tracker tool called experience API (xAPI). The xAPI is a component of the training and learning architecture (TLA). This architecture enables to monitor learning progress and learner's actions like watching a training video or reading an article. The educational dataset consists of data of 480 unique students and 16 features. The Features can be classified into three major categories; 1) Demographic features such as gender and nationality. 2) Academic background features such as educational stage, grade level and section. 3) Behavioral features such as raised hand in class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students from Jordan, 28 students from Palestine, 22 students from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

The dataset is collected through two educational semesters. The dataset also contains school attendance feature the students are classified into two categories based on their absence days that is students exceeding 7 absence days and students under 7 absence days.

Table 1: List of Features used in this study

Attribute	Type	Description
Gender	Nominal	Students' gender
Nationality	Nominal	Students' nationality
Place of Birth	Nominal	Students' place of birth
Educational Stages	Nominal	Educational level student belongs
Grade Levels	Nominal	Grade student belongs
Section ID	Nominal	Classroom student belongs
Topic	Nominal	Course topic
Semester	Nominal	School year semester
Parent responsible for student	Nominal	Parent responsible for student
Raised hand	Numeric	Number of times student raises his/her hand on classroom
Visited Resources	Numeric	Number of times a student visits a course content
Viewing announcements	Numeric	Number of times the student checks the new announcement
Discussion groups	Numeric	Number of times the student participates on discussion groups
Parent Answering Survey	Nominal	Parent answered the survey provided from school or not
Parent school satisfaction	Nominal	The degree of parent satisfaction from school
Student absence days	Nominal	Number of absence days for each student

3.METHODOLOGY

In this paper the proposed methodology is as follows

1) Business Understanding: Involves understanding the goal of the research/ research hypothesis.
2) Data Collection: Involves collecting the data from the learning management system (LMS) called KalBoard360.
3) Data Preparation: Involves pre-processing, cleaning, and transforming into a form that can be used for datamining algorithms.
4) Clustering: In this study we clustered the students based on 4 academic performance metric features that is Raised hands, Visited Resources, Announcements viewed, and Discussion participation. The clustering algorithm used in this study is K-Means.
5) Model Building: Involves developing a wide range of models that is selecting the models
6) Evaluating the models: Involves testing the validity of the model against each other and against the goals of the study.
7) Using the Model: Involves making it a part in the decision-making process.

3.1 K-Means Clustering Algorithm:

In this study, k-means clustering algorithm is used for clustering the KalBoard 360 student data. The features considered for clustering are Raised hands, Visited Resources, Announcements viewed, and Discussion participation. These features are metrics for academic performance. K-means clustering algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The output clusters of k-means can differ in every iteration. Hence, to get reliable clusters k-means is run multiple times on the dataset and the clusters are formed based on all the results of the iterations. After clustering the students, the 3 clusters are assigned with the Grades A, B and C depending on the metric values of the 4 features mentioned above that is the cluster with highest metric values is assigned Grade A and

second highest metric values with B and the last cluster with C.

3.2 Repeated k-fold Cross Validation

To minimize the bias associated with the samples we used repeated k-fold cross validation. In k-fold cross validation the complete dataset is divided into k mutually exclusive subsets of equal sizes. The classification and regression models are trained and tested k times, each time it is trained on all except one-fold and tested on this fold. The prediction results from the k experiments are accumulated into a single confusion matrix. Later these confusion matrix is used in the calculation of accuracy and other metrices. In this study we have taken the value of k as 10 that is 10-fold cross validation and it is repeated 3 times.

3.3. Prediction Methods

In this study, three prediction/classification algorithms are used and compared with each other. They are support vector machines, decision trees, and Naïve Bayes. These algorithms are used because they are capable of their superior capability of modeling classification type prediction problems. Here presented below the brief description of the prediction methods used.

3.3.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning method used for classification. Support vector machine belong to the family of generalized linear models which aims to achieve a prediction decision based on a linear combination of features derived from the variables (Pontil and Verri, 1998). Support vector machine uses both linear and nonlinear kernel functions to transform the input data to a high dimensional feature space in which the input data becomes more manageable. In simple terms, support vector machine finds the mathematical definition of a hyperplane that separates the training data into classes that is the data points that belong to same class are within same side of the hyperplane. Once the best hyperplane is identified it can be used to classify new data into one of the classes. In this study, we have used linear kernel support vector machine.

3.3.2 Decision Trees

Decision tree is one of the most popular technique used for prediction. Most researchers prefer decision trees because of the following reasons: 1) Decision tree outputs are more transparent to the end user that is they produce outputs that are easily readable and understandable.2) They can be easily used and converted into a set of IF-THEN rules to integrate them into the decision support system. This technique recursively separates observations in branches to construct a tree for highest possible prediction accuracy. In constructing the tree different algorithms are used like information gain, chi square statistic etc. based on these values the variables are selected for the node. This process is repeated for each node and the entire tree is constructed. Often, decision trees generate results that are more accurate in decision making and easier to digest. Initial node of the decision tree is called root node and intermediate nodes of the tree are called leaf nodes. The last node of the tree is called end node. The number of branches of the decision tree depends on the specific algorithm used and number of values of the selected variable.

3.3.3 Naïve Bayes

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features. Naïve Bayes classifier is highly scalable. It requires many parameters linear in the number of variables in the learning problem. Naïve Bayes models are known with other names like simple Bayes and independence Bayes. Naïve Bayes assumes the value of feature to be independent of any other feature given the class variable. A Naïve Bayes classifier considers each of these features to contribute independently to the probability, regardless of any possible correlations between the features.

4.Parameter Tuning or Hyperparameter Optimization

Algorithm Parameter Tuning is an important step for improving algorithm performance right before presenting the results or preparing a system for production. It is sometimes called Hyperparameter optimization. The goal of the machine learning is to make the machine system that can automatically build models from data without requiring tedious and time consuming human involvement. One of the difficulties is that learning algorithms require you to set parameters before you use the models. Parameter Tuning or Hyperparameter optimization can be phrased as a search problem, different search strategies can be employed to find a good and robust parameter. In this paper, we are employing grid search to find the hyperparameter.

4.1 Grid Search

Grid Search or Parameter Sweep is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in the grid. In simple terms it is searching through a specified subset of the hyperparameter space of a learning algorithm. A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set. Since the parameter space of the machine learner may include real-valued or unbounded value spaces for certain parameters, manually set bounds and discretization may be necessary before applying grid search. In this search the parameter grid is fed to the respective predictive methods and the method selects the best hyperparameter with high accuracy value.

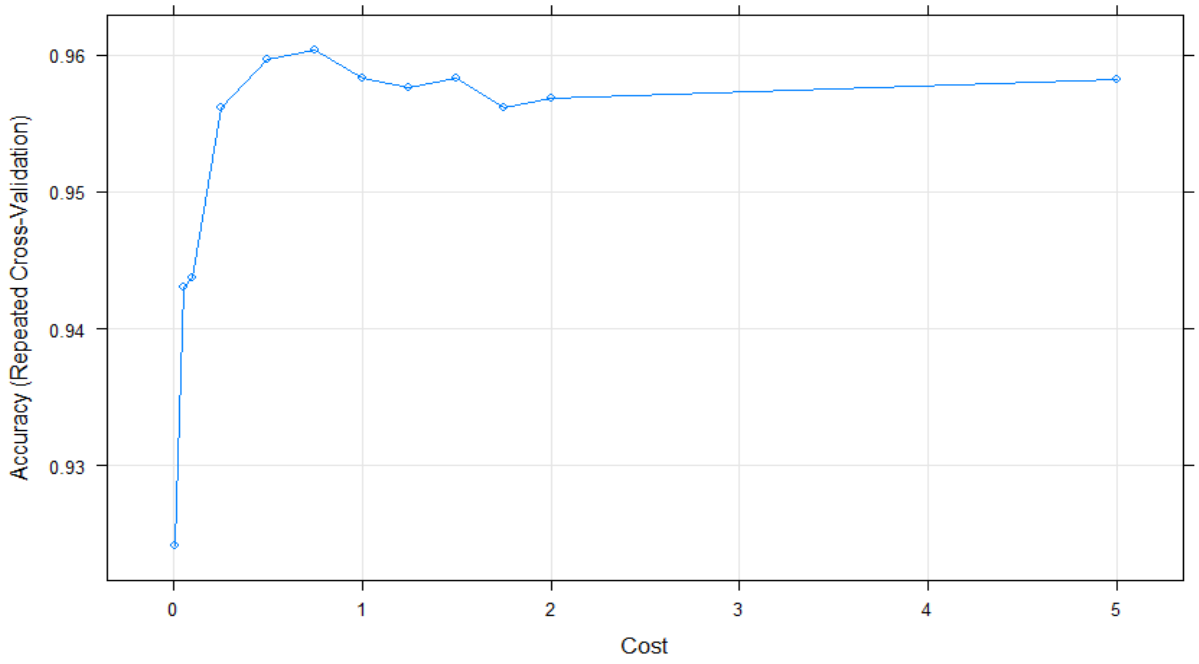

Fig 1: Plot for parameter tuning output of SVM Linear prediction method using Grid Search.

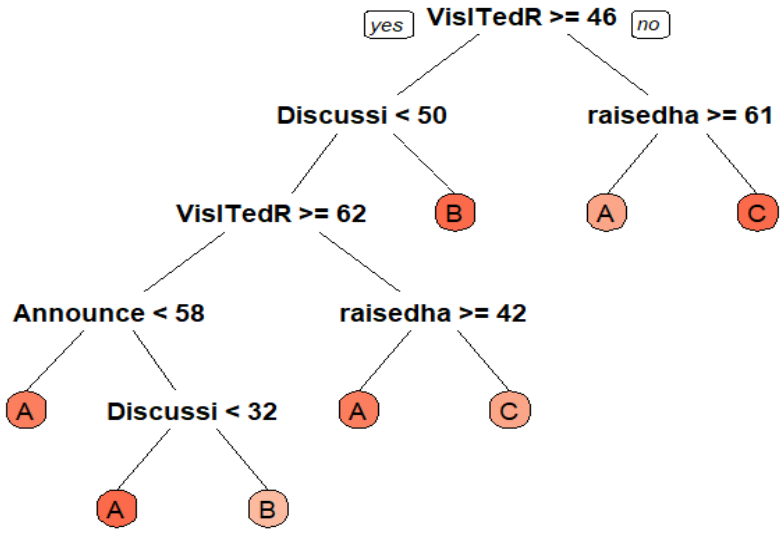

Fig 2: Decision Tree with attributes with high information gain (Most important features/attributes)

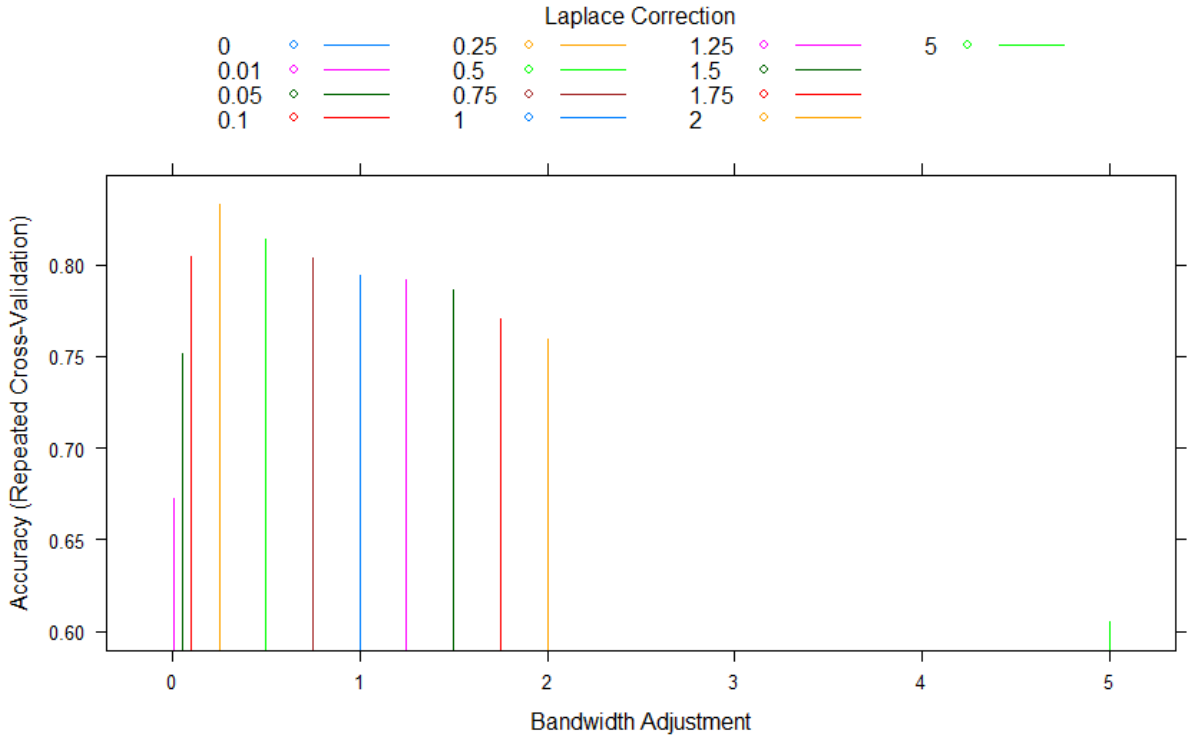

Fig 3: Plot for parameter tuning output of SVM Linear prediction method using Grid Search.

RESULTS AND DISCUSSION

The prediction results for three prediction models used in this study are presented in tables 2, 3, 4, 5 and 6. Table 2 indicates the results of Prediction methods before tuning the parameters. As the results indicate, SVM Linear produced the best prediction results before parameter tuning that is 95.4% accuracy and decision tree produced the prediction results with 90.9% accuracy and Naïve Bayes is least accurate with prediction results of 77.3% accuracy. The results after parameter tuning significantly improved the accuracy of the 3 prediction methods. The prediction accuracy of SVM Linear improved from 95.4% to 96.0%. The accuracy of decision tree improved from 90.9% to 93.4%. The prediction accuracy of Naïve Bayes improved the most that is from 77.3% to 83.3%.

The prediction accuracy of Naïve Bayes is least when compared to other prediction methods. This can be attributed as Naïve Bayes assumes strong independent relationship between the features.

Tables 4, 5 and 6 indicate the results of parameter tuning or hyperparameter optimization.

Table 2: Prediction results of all methods before parameter tuning

Prediction Method	Accuracy	Kappa statistic
SVM Linear	0.9541612	0.9305368
Naïve Bayes	0.7738843	0.6545808
Decision Tree	0.9099080	0.8632813

Table 3: Hyperparameter results of SVM Linear

C	Accuracy	Kappa
0.00	NaN	NaN
0.01	0.9241623	0.8848968
0.05	0.9430182	0.9135244
0.10	0.9436819	0.9146216
0.25	0.9561695	0.9335403
0.50	0.9597002	0.9389256
0.75	0.9603368	0.9398497
1.00	0.9582824	0.9367526

1.25	0.9575879	0.9356860
1.50	0.9582534	0.9366832
1.75	0.9561405	0.9334777
2.00	0.9568350	0.9345166
5.00	0.9582239	0.9366233

Table 4: Hyperparameter Results of Naïve Bayes

fL	**adjust**	**Accuracy**	**Kappa statistic**
0.00	0.00	NaN	NaN
0.01	0.01	0.6721758	0.4965221
0.05	0.05	0.7515342	0.6180692
0.10	0.10	0.8048879	0.6991404
0.25	0.25	0.8332935	0.7428229
0.50	0.50	0.8139754	0.7126672
0.75	0.75	0.8041504	0.6973605
1.00	1.00	0.7943679	0.6818869
1.25	1.25	0.7922408	0.6784013
1.50	1.50	0.7860758	0.6684990
1.75	1.75	0.7707797	0.6442733
2.00	2.00	0.7596230	0.6265340
5.00	5.00	0.6054627	0.3727474

Table 5: Hyperparameter results of Decision tree

cp	Accuracy	Kappa statistic
0.000000	0.9341451	0.9004462
0.05019157	0.9099080	0.8632813
0.10038314	0.9099080	0.8632813
0.15057471	0.9099080	0.8632813
0.20076628	0.9099080	0.8632813
0.25095785	0.9099080	0.8632813
0.30114943	0.9099080	0.8632813
0.35134100	0.9099080	0.8632813
0.40153257	0.9099080	0.8632813
0.45172414	0.5486518	0.2758555

Table 6: Prediction results of all methods after parameter tuning

Prediction Method	Accuracy	Kappa statistic
SVM Linear	0.9603368	0.9398497
Naïve Bayes	0.8332935	0.7428229
Decision Tree	0.9341451	0.9004462

CONCLUSION

Success in data mining depends on the methodology employed. In this study, we employed repeated k-fold cross validation to reduce the bias associated with the samples. This is one of the reason for good prediction results. Next, we further improved the accuracy of the prediction models by Hyperparameter optimization or Parameter tuning. We have seen in the results the improvement in the accuracy value before and after parameter tuning.

As this study illustrated, data mining techniques can predict the grades of the students provided sound methodology is employed. Hence, the study provided us good insights into predicting KalBoard 360 Student Performance Prediction given the features of the student.

FUTURE WORK

This study can be extended to other domains (eg: banking, healthcare, aerospace, medicine/biology, marketing like predicting the sales of a product by ingredients used in it, predicting the success of a student in extracurricular activities etc.).

ACKNOWLEDGEMENT

I Sincerely Thank professor Dr. Collin Lynch for his guidance and support in successful competition of this study.

REFERENCES

[1] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

[2] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[3] Attaway, N. M., & Bry, B. H. (2004). Parenting style and black adolescents' academic achievement. *Journal of Black Psychology*, *30*, 229–247.

[4] Baha Sen, Emine Ucar, Dursun Delen, Predicting and Analyzing Secondary Education Placement-test Scores: A data mining approach, Expert Systems with Applications, V 39 I 10, 0957-4174.

[5] Bain, H. C., Boersma, F. J., & Chapman, J. W. (1983). Academic achievement and locus of control in father-absent elementary school children. *School Psychology International*, *4*(2), 69–78.

[6] Carneiro, P. (2008). Equality of opportunity and educational achievement in Portugal. *Portuguese Economic Journal, 7*(1), 17–41.

[7] Carpenter, P. (1985). Type of school and academic achievement. *Journal of Sociology, 21*(2), 219–236.

[8] Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19*(2), 294–304.

[9] Gentilucci, J. L. (2007). Principals' influence on academic achievement: The student perspective. *NASSP Bulletin, 91*(3), 219–236.

[10] Gerber, S. B., & Fin, J. D. (2001). Teacher aides and students' academic achievement.*Educational Evaluation and Policy Analysis, 23*(2), 123–143.

[11] Goddard, R. D., Sweetland, S. R., & Hoy, W. K. (2000). Academic emphasis of urban elementary schools and student achievement in reading and mathematics: A multilevel analysis. *Educational Administration Quarterly, 36*(5), 683–702.

[12] Kelly, K. (1993). The relation of gender and academic achievement to career self- efficacy and interests. *Gifted Child Quarterly, 37*(2), 59–64.

[13] S. S. Meit, N. J. Borges, B. A. Cubic, H. R. Seibel, Personality differences in incoming male and female medical students., Online Submission.

[14] Steinberg, L., Lamborn, S. D., Darling, N., Mounts, N. S., & Dornbusch, S. M. (1994). Over-time changes in adjustment and competence among adolescents from authoritative, authoritarian, indulgent, and neglectful families. *Child Development, 63*, 754–770.

[15] Stipek, D. J. (1981). Perceived personal control and academic achievement. *Review of Educational Research, 51*(1), 101–137.

[16] Yenilmez, K., & Duman, A. (2008). Interviewing with students about the factors that affect the achievement of mathematic in primary school. *Sosyal Bilimler Dergisi, 19*, 251–268.