Predicting Students’ Academic Performance and Main Behavioral Features using Data Mining Techniques

Suad Almutairi[1], Hadil Shaiba[1], and Marija Bezbradica[2]

[1]Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
{sfalmutairi,hashaiba}@pnu.edu.sa
[2]Dublin City University, Dublin, Ireland
marija.bezbradica@dcu.ie

Abstract. Creating learning environments, where students, parents, and teachers are linked to a learning process, helps study their overall impact on the students’ performance. Data mining can analyze these inter-relationships and thus enable the prediction of academic performance to improve the student’s academic level. The main factors that affect the student’s performance were selected using feature selection methods. An analysis of the crucial features was investigated to better understand the data. One of the main outcomes found is the impact of the behavioral features on the students’ academic performance. Moreover, gender and relation demographical features are another important features found. It was evedent that there is an academic disparity between genders, as females constitute the most outstanding students. Furthermore, mothers have a clear role in student academic excellence. Six machine learning methods were used and tested to predict the studnet’s performance, namely random forest, logistic regression, XGBoost, MLP, and ensemble learning using bagging and voting. Of all the methods, the random forest got the highest accuracy with 10-best selected features that reached 77%. Overfitting was addressed successfully by tuning the hyper-parameters. The results show that data mining can accurately predict the students’ performance level, as well as highlight the most influential features.

Keywords: Educational data mining, machine learning, deep learning, learning analytics.

1 Introduction

Building learning environments, where students have an active interaction in their learning, has become a priority for educational institutions. Learning Analytics (LA) is an emerging area for the collection, analysis, and presentation of learners' data for the purpose of studying the influencing factors on the learning process with the aim of understanding and developing the learning environment [1]. LA provides all parties (parents, teachers, and students) with the appropriate and quick feedback about the educational process. On the other hand, behavior analytics helps us understand the behavior of students and how they interact during the learning period with contributing influences.

Predicting the performance of students has attracted the attention of several authors due to its importance in helping teachers identify and support their students according to the level of difficulty [2]. Considerable works have been done in recent years to analyze the behavior of students and extract the significant patterns that can be used to predict the students’ performance.

This study intends to use data mining methods to: (i) find the strongest features that can help in the prediction of students’ performance, (ii) analyze the most important behavior and demographical features to have a better understanding of the features that affect the students’ performance level, (iii) predict the students’ performance by using data mining techniques and show how feature selection, oversampling, ensemble learning, and parameter tuning can enhance the predictive power of the models and resolve overfitting.

A summary of the previous work has been presented in the literature review section. Explanation of the data used and the data mining methods applied in this resarch have been discussed in the data and methodology sections. In the results section, we show the best selected features, provide a detailed analysis of the main selected features, and evaluate and discuss the results of our prediction models. We finally conclude our research in the conclusion section.

2 Literature Review

2.1 The Use of E-Learning in Education

Online-based learning environments, such as learning management systems (LMSs), allow teachers to study and track students' performance by recording and keeping student information online [2]. E-Learning environments have been used to monitor and record all educational processes and actions done by students, thus it could provide useful information on the progress that each student achieves as mentioned in [3]. In order to achieve the best level of electronic learning, it is necessary to evaluate the processes of learning and teaching continuously by observing all aspects, from the level of interaction between the parties involved in the quality of teaching through the reactions of students and their initiative. In addition to the use of multiple sources, the effect of management, and other aspects, on the development of cognitive skills, should be considered [4].

In recent years, there has been a significant growth of using methods to facilitate the analysis of existing processes related to learners and e-learning systems. In order to provide a more efficient learning environment, data mining techniques have been used in this field, for processing the data and extracting information patterns that can be useful indicators [3].

2.2 Educational Data Mining

Data mining in the educational analysis is described as the process of automatic extraction of a meaningful chain from a large dataset. It is used not only to train the model on the learning process but also to evaluate and develop e-learning systems [3].

The efficiency of machine learning methods has been analyzed in [2] by predicting the difficulty the students will face in the next session to support the students and help them according to the level of difficulty. Five of the well-known machine learning algorithms have been used for the prediction process, namely artificial neural network (ANN), logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), and decision tree (DT). The methods have been selected based on their suitability for the dataset and insensitivity towards overfitting [2]. For evaluation, authors in e.g. [2] used multiple techniques such as root-mean-square error (RMSE) and Cohen's kappa coefficient. Feature selection techniques focus on reducing the dimensionality and avoid unrelated data to the research interest [2]. Alpha-investing feature selection for ranking was used by [2] to minimize the input features to be used in the student prediction model. Their results showed that SVMs and ANNs are the most suitable models to predict student's performance [2].

Two datasets were analyzed in [5] to predict students failing in the early stages, exactly after the first exam in introductory programming courses available from a Brazilian Public University. A noticeable result was shown by using the SVM algorithm and F-measure for evaluation. They claimed that female students had less in-class participation than male students. In addition to that, after the first exam application, the F-measure value reached approximately 0.92 and 0.83 with distance and on-campus datasets, respectively.

Other research [6] focused on a new side of educational analytics called behavioral features. The authors chose ANN, DT, NB and ensemble models (bagging, boosting, and random forest) to build a performance predictive model. The information gain algorithm has been used to build the students' performance model. The results were presented with and without behavioral characteristics using 10-fold cross-validation. The result showed how behavioral characteristics had a strong effect on the students' academic achievement. In addition, the ANN technique overcomes other methods, as its accuracy was around 79.1%. The result and quality of the classifiers were measured by four common measures; Accuracy, Precision, Recall, and F-Measure [6]. The lack in this paper is that it did not predict early enough student performance due to the lack of information about midterm exams and assignments, the prediction was after the finals.

Using data mining to predict the student's dropouts was explored with a dataset of 165, 715 high school students from different schools [7]. The authors [7] selected significant features that were presented by using the random forests model with out of bag (OOB) estimate. The unauthorized absence was the most significant variable in predicting students' dropouts, followed by unauthorized lateness. The random forests model predicts students' dropouts with a high accuracy of 95% using 10 folds cross-validation. AUC score got 97% which represents an outstanding performance. The work in this article was excellent, but they noted some shortcomings with the calculation of the model features, that used inaccuracy weights [7].

The researches which were discussed in this section varied by using data mining techniques to predict at-risk students. Findings cannot be generalized due to their limited domain, but their work is worthy of praise. The following section concentrates on one main stage in data analysis which is feature selection and its effect on the predicted results and data visualization and its importance in better understanding the data..

3 Dataset

A LMS called Kalboard 360 has been used to collect educational dataset. This system gives users (students, teachers and parents) synchronous access to reach the educational resources from all devices by using an internet connection [6]. The original source of the dataset is found in [6]. The dataset is available on Kaggle.com under the name of BStudents' Academic Performance Dataset. In total 480 students with 16 features are analyzed in this project which can be divided into four basic categories. Details of these features and their number of instances have been presented in Table 1.

Table 1. Features' description of the dataset used in this study adapted from [7].

Features Category	Feature	Description	Number of Instances
Demographical Features	Gender	Male/Female	2
	Nationality	Kuwait/Lebanon/Saudi Arabia etc.	14
	Place of birth	Kuwait/Lebanon/Saudi Arabia etc.	14
	Relation	Parent responsible for student (Mother/father)	2
Academic Background Features	Educational Stages	Lower level/Middle School/High School	3
	Grade Levels	G-01/G-02/……. /G-12	10
	Section ID	Classroom student belongs to (A/B/C)	3
	Topic	Course topic (English/Spanish/French/IT etc.	13
	Student Absence Days	above-7/under-7	2
	Semester	First/Second	2
Behavioral Features	Raised hand	How many times the student raises his/her hand in classroom (numeric:0-100)	101
	Visited resources	How many times the student visits a course content (numeric: 0-100)	101
	Viewing announcements	How many times the student checks the new announcements(numeric:0-100)	101

	Discussion groups	How many times the student participate in discussion groups (numeric:0-100)	101
Parents Participation on learning process	Parent Answering Survey	Yes/No	2
	Parent School Satisfaction	Yes/No	2

The class feature contains the performance level which is the total mark of the student in a subject decided in each record. This performance level is categorized into three levels (High, Medium, and Low). Marks below 70 are belonging to the low level, marks between 70 and 89 are belonging to the medium, and marks higher than 89 considered as a high level [6].

4 Methodology

The process of feature selection is considered as a major step in the classification process. This is due to two great benefits: the first is to reduce the complexity of the computation [16][17], and the second is to enhance the classifier's generalization to correctly classify unseen instances [17]. In this research, the wrapper method and the filter-based method have been used for feature selection. A subset of features is selected and evaluated based on the predictor's accuracy [16] in a wrapper method which guarantees accurate results [17]. Recursive Feature Elimination (RFE) have been used as a wrapper method.

The filter method requires a lower computational effort [17]. Information Gain algorithm (IG) and K-best feature selection have been used. IG ranks the features separately in decreasing order based on how relevant it is to the class label (student academic performance level). K-best (SelectKBest) is a univariate method that selects features based on the K highest scores. It calculates the Analysis of Variance (ANOVA) F-value between each feature and the class which is the target vector [18].

Four popular classification algorithms are used in this research in addition to ensemble models. The discription of each method is explained in the following:

• **Random forest (RF)** algorithm has its immunity against exaggeration [7]. It is also considered as an effective tool for prediction [8]. Moreover, it is known for its performance and ability to dealing with corrupted data and overfitting [9].

• **Extreme gradient boosting (XGBoost)** is a scalable implementation of gradient boosting framework which works efficiently [10]. Its scalability refers to many algorithmic optimizations, and it is trained in an additive manner [11].

• **Logistic regression** reduces the impact of confounding factors by analyzing multiple explanatory variables at the same time [12]. In addition to that, the allowance of continuous variables has a useful smoothing impact on the model or on the estimates [13].

• **Multilayer perceptron artificial neural network (MLP)** is one type of neural network with a number of neuron layers such that every layer is connected to the following layer. In this way, it can classify data that is not separated linearly [14]. This algorithm is known for reducing the estimation error by calculating and updating the weights in the network in order to acquire the great configuration of the neural network [15].

- **Ensemble models** are the new trend in machine learning, which are based on training multiple classifiers, therefore the choice of the best result. **Bagging** uses one model over various and random sampling (with replacement) of the dataset. Bagging using a decision tree has been applied in this research. **Voting** combines multiple classifiers and selects the best performance by voting. Soft voting was chosen in this research by Ensemble Vote Classifier function with mixing logistic regression and XGBoost.

The best practices to avoid overfitting were introduced in this paper. Along with feature selection, oversampling, ensemble learning, and parameter tuning were used. This could be as road map to researchers who would like to resolve the problem of their predictive models facing overfitting.

5 Results and Discussion

The result and performance of the model depend on the techniques used in the preprocessing. The original data is converted into a form that is suitable for use with data mining such as data cleaning and data conversion [2]. The whole 480 records in the dataset are clean from any missing values, and outliers. The class feature has three values, High, Medium, and low levels which contain 141, 211, and 127 cases, respectively. This slight difference in the number of cases is not considered as imbalance dataset [19]. Normalization is applied to prevents misleading variance between features' values. After these steps, the dataset is ready to be used with classification methods.

To analyze the features and their relationships, the dataset has been explored using data visualization. Then, evaluation using train/test/split and cross-validation with the feature selection are studied next.

5.1 Evaluation Measures

Four different measures for evaluating the predictive models were used, which are precision, recall, accuracy, and F-measure. Equations of these measures are defined in the following [6]:

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} \qquad (1)$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}} \qquad (2)$$

$$\text{Accuracy} = \frac{\text{tp+tn}}{\text{tp+tn+fp+fn}} \qquad (3)$$

$$\text{F} - \text{measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision+recall}} \qquad (4)$$

Those four measures are calculated using the confusion matrix shown in Table 2.

Table 2. Confusion Matrix [7].

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

5.2 Feature Analysis

Three feature selection methods have been used in this research. Information gain ranking and 10-best feature selection chose the same features except that information gain selected nationality and place of birth with the most high-ranking features while the 10-best method selected gender and semester as shown in Table 3. On the other hand, RFE has chosen different collection with each classifier and focused on academic background features. Figure 1 shows that students who did more behavioral features have acquired the highest grades. Since this paper attempts to discover which behavioral features have the greatest impact on the class, thus, more focus is presented about these features and their relation to the other features that got higher ranks, which are: **(1)** The parent responsible, **(2)** Parents' involvement, **(3)** Students' absence days and **(4)** The gender of the student.

1. The parent responsible: Figure 2 shows that students were more active when their mother was responsible for them. Fathers make up two-thirds of the dataset, yet the parent responsible for the most excellent students were their mothers. Figure 3 shows that mothers were responsible for the most high-level students.

2. Parents' involvement: After addressing the effect of parents' involvement on the behavioral features we found a strong relationship between these features. Most of the students who were active during the learning period; their parents answered the survey and were satisfied with the school.

3. Student' absence days: There is clear evidence that the absence of a student negatively affects participation, visiting resources, viewing announcements, and raising hands during class.

4. The gender of the student: Most of the students who got the highest grades are females with 52%, compared with male students which are 47%. This means that 42% of the whole female students got high-performance level, while just 21% of the whole male students are excellent. The discussion did not show a significant influence on female students as presented in Figure 4.

Table 3. Top 10 features using k-best feature selection (were k=10) and information gain ranking methods.

Best 10 features using k-best method	**Best 10 features using information gain ranking**
Visited Resources	Visited Resources
Student’ Absence Days	Student’ Absence Days
Raised hands	Raised hands
Announcements View	Announcements View
Parent’ Answering Survey	Parent’ Answering Survey
Gender	Nationality
Semester	Place of Birth
Relation	Relation
Discussion	Discussion
Parent’ School Satisfaction	Parent’ School Satisfaction

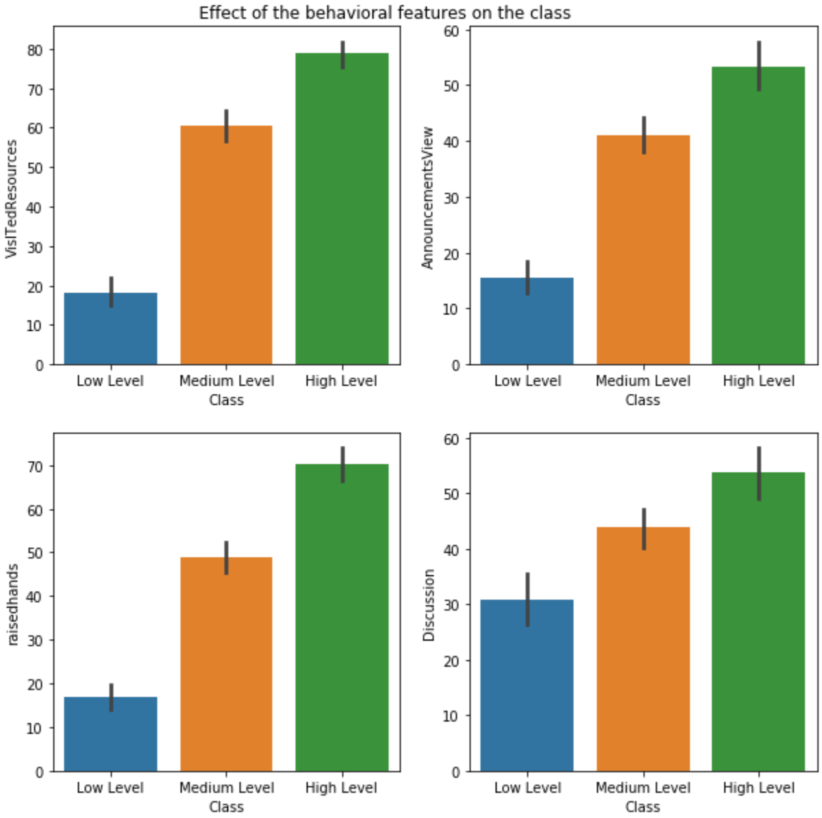

Figure 1. Comparison between the four behavioral features and the students’ academic performance (the class feature).

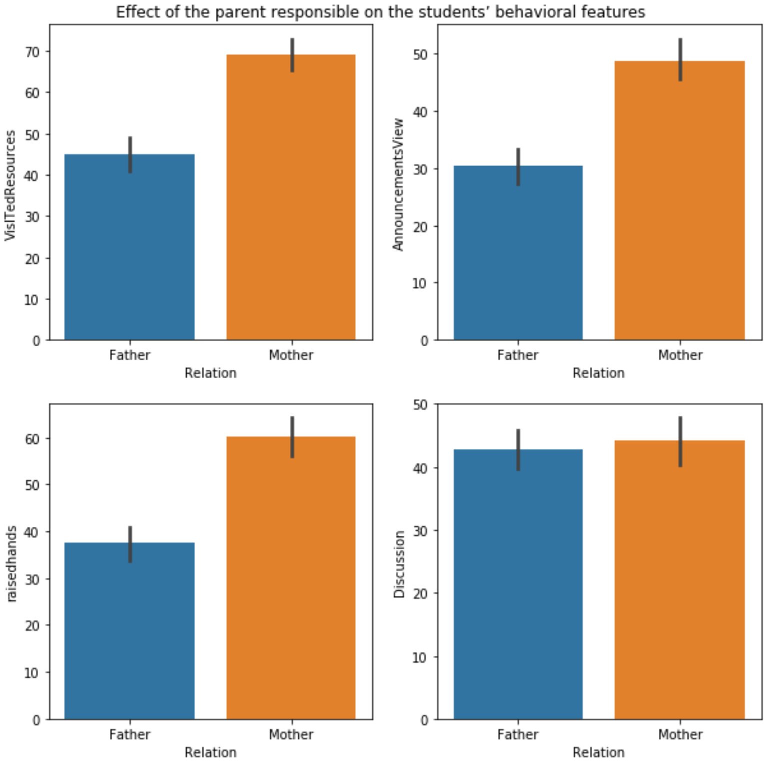

Figure 2. Effect of the parent responsible of the student on the four behavioral features.

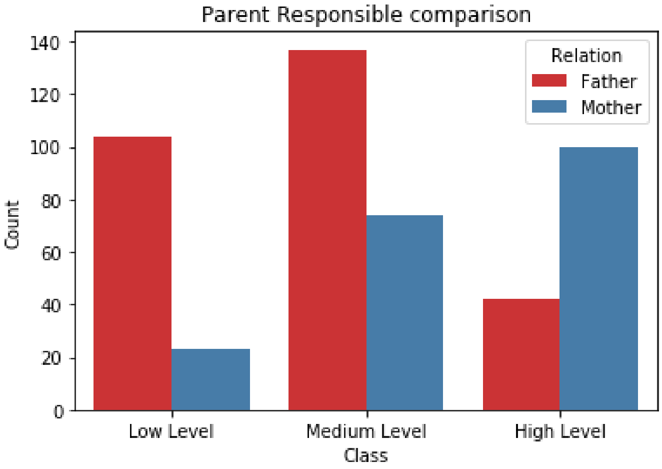

Figure 3. Influence of the parent responsible of the student on the students' academic performance (the class feature).

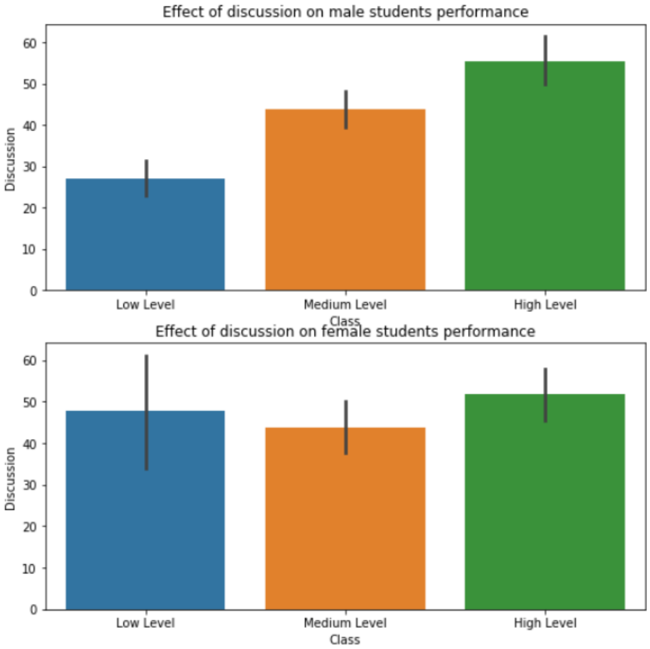

Figure 4. Effect of discussion feature on the academic performance of both students' genders.

Four of the behavioral features (visited resources, raised hands, discussion and announcements view) were ranked by information gain method as the top features which influenced the students' performance. Showing those behavioral features visually have made a clear vision of the influence of these features on the student's academic success.

5.3 Evaluation Using Train/Test/Split

For evaluation, the dataset was first split to train, test, and validation such that each part has 60%, 20%, and 20%, respectively. The results are listed in Table 4.

Table 4. Train/test/split evaluation results including all features presented in Table 1.

Accuracy Score	Logistic Regression	Random Forest	MLP	XGBoost	Ensemble Voting	Ensemble Bagging
Accuracy on training set	78.1%	100%	50.6%	96.5%	78.4%	80.9%
Accuracy on Validation set	68.7%	72.9%	54.1%	69.7%	68.7%	64.5%
Accuracy on test set	65.6%	70.8%	56.2%	77.08%	66.6%	70.8%

Table 4 shows accuracy scores with a significant sign of overfitting problem with all classifiers and a slight underfitting with MLP. Overfitting means that the predictive model gives satisfactory results, but when applied with new data, it gave significantly lower results. In other words, the model will not be able to generalize, because it fits much noise while training [7]. Multiple techniques were used to overcome overfitting issues such as: **(1)** Examining the collinearity between features, **(2)** Oversampling to gain more data, **(3)** Fine-tuning of hyperparameters and **(4)** Feature selection. After applying each technique, classification algorithms were used to examine their effect on solving the overfitting problems.

1. Examining the Collinearity Between Features: To address the collinearity, two methods have been used which are the heat map and Reduce VIF (variance inflation factor) function. Heat Map showed that there is no significant correlation higher than 0.7 between the features except the correlation between the nationality and the place of birth which shows 0.9 as expected. The VIF measures the collinearity between features in regression analysis [20]. After applying this method, four features have been chosen, one of them was PlaceofBirth feature which we deleted. The remaining three were behavioral features that we combined into one feature. Addressing the collinearity showed a slight improvement with the overfitting problem, but it was not solved completely.

2. Oversampling with SMOTE: To oversample the minority class, SMOTE-NC (Synthetic Minority Oversampling Technique Nominal Continuous) algorithm is used with the training part of the original dataset [5]. This technique takes each sample of the minority class and creates synthetic examples based on nearest neighbors [19]. After doing oversampling, classification methods are applied, random forests got 72.9% but the overfitting was not solved with all classifiers.

3. Hyper-Parameter Tuning: Each classifier has its hyper-parameters that are used to control the learning process. Choosing a specific set of values for those hyper-parameters will change the performance of the model [7][21]. The grid search algorithm is used to accelerate the selection of values. Some hyper-parameters proved their effect on solving overfitting such as the maximum depth of the trees, the higher value means the deeper tree, and this will lead to capturing more knowledge about the data. Moreover, the minimum number of samples that need to split an internal node. As increasing its value, each tree will be more constrained because of the growing number of samples that each node needs to consider. In logistic regression, the C parameter (inverse of regularization strength) controls the regularization by adding information to solve problems such as overfitting and increase the accuracy. As the value becomes smaller, the stronger the regularization is. Table 5 shows the hyperparameters that prove to solve the overfitting problem.

Table 5. Hyper-parameter tuning to solve overfitting problem.

Classifier	Random forest	Decision tree	Logistic regression	XGBoost

Parameter	max_depth = 6
	min_samples_split = .2
		max_depth = 2
		min_samples_split = .02
			C = .09	max_depth = 2

Table 6 shows the result after tuning the hyper-parameter with information gain as a filter-based feature selection. The score of training for logistic regression improves to reach 80.5%. Moreover, the random forest training score decreases to reach 82.2% to solve the overfitting problem. While the test score for ensemble voting increases to 71.8%.

Table 6. The result of train/test/split evaluation with the top 10 features using filter-based information gain ranking and after hyper-parameter tuning.

Accuracy Score	Logistic Regression	Random Forest	MLP	XGBoost	Ensemble Voting	Ensemble Bagging
Accuracy on training set	80.5%	82.2%	77.7%	79.1%	80.5%	74.6%
Accuracy on Validation set	73.9%	72.9%	77.08%	68.7%	72.9%	67.7%
Accuracy on test set	71.8%	75%	68.7%	68.7%	71.8%	67.7%

In Table 7 the result of 10-best feature selection which shows that overfitting problem has been solved except that there is a little overfitting with XGBoost. Comparing the result of both feature selection techniques, the k-best shows the best result with random forest which got 77.08%.

Table 7. The result of train/test/split evaluation with filter-based 10-best feature selection method and after hyper-parameter tuning.

Accuracy Score	Logistic Regression	Random Forest	MLP	XGBoost	Ensemble Voting	Ensemble Bagging
Accuracy on training set	77.08%	82.2%	74.3%	80.2%	77.7%	74.6%
Accuracy on Validation set	72.9%	72.9%	68.7%	70.8%	72.9%	64.5%
Accuracy on test set	67.7%	77.08%	68.7%	68.7%	67.7%	68.7%

There is a decrease in the score of training, but this to solve the problem of overfitting, which is more important for the generalization of the models. Random forest shows the highest testing score with train test evaluation using both feature selection methods.

5.4 Evaluation Using Cross Validation

The k-fold cross validation is considered as the most popular technique to avoid overfitting [7]. The results of using 10-fold cross-validation with other techniques are shown in Table 8. MLP, ensemble voting, and ensemble bagging algorithms cannot work with RFE, because they do not have any metric for feature importance.

Table 8. 10-Fold cross validation results scores with all feature selection methods.

Accuracy	Logistic Regression	Random Forest	MLP	XGBoost	Voting	Bagging
With all features (original dataset)	70%	72.08%	68%	70.8%	71.2%	70.8%
With combining features and removing correlation	70.4%	71.6%	68%	70.6%	70%	70%
With information gain (the best 10 ranks)	74.1%	75%	75%	71.6%	72.5%	68.9%
With 10-best feature selection	73.5%	75.2%	74%	71.8%	72.7%	70.4%
With RFE	72.9%	74.1%	Not used	72.08%	Not used	Not used

Most classifiers show their best performance with 10-best feature selection. Applying the filter method with information gain for feature selection shows great accuracies for random forest and MLP as they reach 75%. On the other hand, XGBoost got its best score with RFE feature selection. Comparison with other evaluation measures using cross-validation based on 10-best selected features is listed in Table 9.

Table 9. Evaluation scores of 10-fold cross validation based on 10-best selected features.

Scores	Logistic Regression	Random Forest	MLP	XGBoost	Ensemble Voting	Ensemble Bagging
Accuracy	73.5%	75.2%	74%	71.8%	72.7%	70.4%
F1	73.2%	71.4%	73.1%	70.9%	73.5%	67.5%
Precision	74.7%	74.3%	75.3%	73.5%	74.91%	72.5%
Recall	73.5%	72.1%	73.6%	71.2%	73.8%	68.3%

Other evaluation metrics are addressed in addition to accuracy which are recall, precision, and F1 measures. Recall means the ratio of the total relevant results that are classified correctly by the classifier. Ensemble Voting showed 73.8% percentage in the recall, while MLP got 73.6%. Moreover, precision expresses the ratio of the instances that the algorithm classifies them as relevant, and they were relevant. Ensemble Voting is precise with 74.9%, while logistic regression got 74.7% in precision.

6 Conclusion

Academic achievement has become widely concerned by academic institutions around the world. With the widespread use of e-learning management system, a lot of hidden knowledge can be extracted and analyzed to improve the students' performance level. Students' performance prediction model has been presented in this research. Predicting the academic level of students with a small amount of data shows clear evidence of overfitting problems. The best ways that worked better with this research to overcome overfitting and to increase the accuracy of the model is by:

- Tuning the hyperparameters especially the ones that affect overfitting.
- Using feature selection methods such as filter based k-best feature selection.

Data mining techniques proved that they can predict the student academic level which in turn answers the first question of the paper. In this research, six predictive models have been used. Random forest shows the highest testing score with train test evaluation which reaches 77.08% with the k-best feature selection method. With cross-validation, it got 75.2% based on 10-best feature selection. Thus, it confirms what has stated in [9] that it fits the data facing the problem of overfitting.

Visited resources, raised hands, announcements view, and the discussion shows a clear impact on students' final scores which brings us to answer the second question. These behavioral features show a significant relationship with the academic performance of the students. Moreover, they got the highest rank using information gain ranking and was chose by the 10-best method as the most features that influence the academic performance of the students.

Male and female students have the same academic behavior except with discussion which did not show a clear influence on female students' academic performance. An interesting finding has been shown by the parent responsible feature, as fathers make up two-thirds of the dataset, yet for most high-performance students their mother was responsible for them during their studies. Female students in this dataset prove their high academic performance, such that 42% of the whole female students got high-performance level, compared with male students who represent just 21% of them have got high academic performance.

In our future work, more investigation in hyper-parameter tuning will be performed. Moreover, increasing the size of the dataset by merging with other data from another school will make the predictive model more convenient to be generalized. Different academic performance between both genders needs more attention to know beyond this disparity. This is also the case with the parent responsible, is there an education strategy followed by mothers and thus enable students to excel?

References

1. Sin, K., Muthu, L.: Application of Big Data in Education Data Mining and Learning Analytics – a Literature Review. 6956, 1035–1050 (2015)
2. Hussain, M., Zhu, W., Zhang, W., Muhammad, S., Abidi, R., Ali, S.: Using Machine Learning to Predict Student Difficulties from Learning Session Data. Artif. Intell. Rev. (2018)
3. Romero, C., Ventura, S.: Educational Data Mining: a Survey from 1995 to 2005. Expert Syst. Appl. 33, 135–146 (2007)
4. Rodrigues, M.W., Isotani, S., Zárate, L.E.: Educational Data Mining: a Review of Evaluation Process in the E-Learning. Telemat. Informatics. 35, 1701–1717 (2018)

5. Costa, E.B., Fonseca, B., Santana, M.A., De Araújo, F.F., Rego, J.: Evaluating the Effectiveness of Educational Data Mining Techniques for Early Prediction of Students' Academic Failure in Introductory Programming Courses. Comput. Human Behav. 73, 247–256 (2017)
6. Amrieh, E.A., Hamtini, T., Aljarah, I.: Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods. Int. J. Database Theory Appl. 9, 119–136 (2016)
7. Chung, J.Y., Lee, S.: Dropout Early Warning Systems for High School Students Using Machine Learning. Child. Youth Serv. Rev. 96, 346–353 (2018)
8. Hänsch, R., Hellwich, O.: Random Forests. Handb. Der Geodäsie. 1–42 (2016)
9. Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., Montenegro, M.: Centralized Student Performance Prediction in Large Courses Based on Low- Cost Variables in an Institutional Context. Internet High. Educ. 37, 76–89 (2018)
10. Chen, T., He, T.: Xgboost : EXtreme Gradient Boosting. 1–4 (2019)
11. Krauss, C., Do, X.A., Huck, N.: Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. Eur. J. Oper. Res. 259, 689–702 (2017)
12. Sperandei, S.: Lessons in Biostatistics Understanding Logistic Regression Analysis. (2014)
13. Roalfe, A.K., Holder, R.L., Wilson, S.: Standardisation of Rates Using Logistic Regression: a Comparison with the Direct Method. BMC Health Serv. Res. 8, 1–7 (2008)
14. Latham, A., Crockett, K., Mclean, D.: Profiling Student Learning Styles with Multilayer Perceptron Neural Networks. 2013 IEEE Int. Conf. Syst. Man, Cybern. 2510–2515 (2013)
15. Kayri, M.: an Intelligent Approach to Educational Data : Performance Comparison of the Multilayer Perceptron and the Radial Basis Function Artificial Neural Networks. 15, 1247–1256 (2015)
16. Chandrashekar, G., Sahin, F.: a Survey on Feature Selection Methods q. Comput. Electr. Eng. 40, 16–28 (2014)
17. Maldonado, S., Weber, R.: a Wrapper Method for Feature Selection Using Support Vector Machines. Inf. Sci. (Ny). 179, 2208–2217 (2009)
18. Pe, M., Barrag, M.: Feature Engineering Based on ANOVA , Cluster Validity Assessment and KNN for Fault Diagnosis in Bearings. 34, 3451–3462 (2018)
19. Chawla, N. V, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE : Synthetic Minority Over-Sampling Technique SMOTE : Synthetic Minority Over-Sampling Technique. (2002)
20. Robinson, C., Schumacker, R.E.: Interaction Effects: Centering, Variance Inflation Factor, and Interpretation Issues. 35, 6–11 (2009)
21. Mustaffa, Z., Yusof, Y.: LSSVM Parameters Tuning with Enhanced Artificial Bee Colony. 11, 236–243 (2014)