ANALYSIS OF THE FEATURES FOR AUTOMATIC CLASSIFICATION OF ACADEMIC PERFORMANCE

Hakan Alp EREN [1], Efnan SORA GUNAL [2*]

[1] Dept. of Software Engineering, Eskişehir Osmangazi University, Eskişehir, Türkiye
ORCID No: https://orcid.org/0000-0001-6105-158X
[2] Dept. of Computer Engineering, Eskişehir Osmangazi University, Eskişehir, Türkiye
ORCID No: https://orcid.org/0000-0001-6236-174X

Keywords
Academic performance
Data mining
Machine learning
Classification
Feature selection

Abstract

This paper analyzes the contributions of features widely used in the automatic classification of students' academic performance. In this classification problem, the relationship between various features and classifiers is analyzed using an exhaustive feature selection strategy. In this way, the optimal subset of features providing the highest classification performance is obtained. For this purpose, an academic performance dataset consisting of 15 distinct features and 480 samples is used. The features mainly belong to four different categories, including demographic, academic background, parent participation, and behavioral. The samples are from three different classes corresponding to the low, middle, and high levels of students' success. For evaluations, 10 different classification algorithms are employed. Extensive experimental analysis reveals that the accuracy of the classification of students' academic performance can be improved up to 79.40% using only 8 features rather than all.

AKADEMİK PERFORMANSIN OTOMATİK SINIFLANDIRILMASI İÇİN ÖZNİTELİKLERİN ANALİZİ

Anahtar Kelimeler
Akademik performans
Veri madenciliği
Makine öğrenmesi
Sınıflandırma
Öznitelik Seçimi

Öz

Bu makale, öğrencilerin akademik performansının otomatik olarak sınıflandırılmasında yaygın olarak kullanılan özelliklerin katkılarını analiz etmektedir. Bu sınıflandırma probleminde, çeşitli öznitelikler ve sınıflandırıcılar arasındaki ilişki, kapsamlı bir öznitelik seçim stratejisi kullanılarak analiz edilmiştir. Bu şekilde, en yüksek sınıflandırma performansını sağlayan optimal öznitelik alt kümesi elde edilmiştir. Bu amaçla 15 farklı öznitelik ve 480 örnekten oluşan bir akademik performans veri seti kullanılmıştır. Öznitelikler demografik, akademik geçmiş, ebeveyn katılımı ve davranışsal olmak üzere dört farklı kategoriye aittir. Örnekler, öğrenci başarısının düşük, orta ve yüksek seviyelerine karşılık gelen üç farklı sınıftandır. Değerlendirmeler için 10 farklı sınıflandırma algoritması kullanılmıştır. Kapsamlı deneysel analizler, öğrencilerin akademik performansını sınıflandırma doğruluğunun, özniteliklerin tamamı yerine yalnızca 8 tanesi kullanılarak, %79.40'a kadar artırılabileceğini ortaya koymaktadır.

Araştırma Makalesi		Research Article	
Başvuru Tarihi	: 28.12.2021	Submission Date	: 28.12.2021
Kabul Tarihi	: 05.04.2022	Accepted Date	: 05.04.2022

1. Introduction

Data mining is simply the extraction of meaningful information by processing data collected in various ways and from various resources (Han et al., 2011; Aggarwal, 2015). A descriptive or predictive model can be created from the data. While the descriptive models are used to find relationships, trends, clusters, and

* Corresponding Author; e-mail : esora@ogu.edu.tr

anomalies in the data, the predictive models estimate the value of a variable based on the values of other variables. The variable to be estimated is called the target (dependent) variable and the variables used for the estimation are called explanatory (independent) variables. The estimation can be made in the form of classification or regression. Assessment of the academic performance of students using classification estimators, or simply classifiers, is just one of the use cases of these estimations among many other fields. With the help of various features and classification models, one can predict the academic performance of students.

In the literature, there are several works on the classification of academic performance. As an example, Huang and Fang (2013) developed a set of mathematical models and then identified the most appropriate models for predicting student academic performance. Four types of mathematical modeling techniques (multiple linear regression, multilayer perception network, radial basis function network, and support vector machine) and six combinations of predictor variables were used to develop a total of 24 predictive mathematical models based on the dataset collected from 323 undergraduates in four semesters. The outputs of the models were the students' scores on the final exam of the engineering dynamics course. The inputs of the models were the student's cumulative GPA, grades earned in four pre-requisite courses (statistics, calculus I, calculus II, and physics), and the scores on three mid-term exams of the engineering dynamics course. Amrieh et al. (2015) developed a model that predicts students' academic achievement by using the attributes that were divided into three categories, namely demographic, academic background, and behavioral. The demographic features consist of nationality, gender, place of birth, and the parent responsible for the student. The academic background features include school level, grade, section, semester, topic, and teacher ID. The behavioral features consist of raised hands, visited resources, joining discussion groups, and viewing announcements. The students were divided into three levels as low, medium, and high based on their grades. Decision tree, artificial neural network, and naive Bayes were used as the classification algorithms. In another work (Amrieh et al., 2016), features were divided into four categories: demographic features, academic background features, parent participation in the learning process, and behavioral features. Demographic features were nationality, gender, place of birth, and the parent responsible for the student. Academic background features included school level, grade, section, semester, topic, and the number of absences. Whether the parent answered the survey about school and whether they were satisfied with the school were the features in the parent participation category. Behavioral features included participation in discussion groups, access to course resources, number of raised hands in class, and viewing announcements. The dataset contains 15 features and 500 students. In this study, filter-method were applied for feature selection using an information gain-based algorithm. The models were created using various classifiers and ensemble methods on two datasets with and without behavioral features. In another work, the performances of several feature selection algorithms were analyzed on the student academic performance dataset (Zaffar et al., 2017). The best combinations of feature selection and classification algorithms were obtained. Rahman et. al used the classification algorithms, including naive Bayes, artificial neural network, decision tree, and *k*-nearest neighbor as well as ensemble filtering methods to classify the academic performance of students (Rahman and Islam, 2017). Hussain et al. (2018) developed a model for classifying the academic performance of university students. The attributes included gender, social class, family size, marital status, income, and attendance. Four classifiers were used in this work. These classifiers and their accuracies are as follows: random forest (99%), PART (74.33%), J48 (73%), and BayesNet (65.33%). Finally, a new model and features for predicting the academic performance of students were introduced in (Sana et al., 2019). Ha et al. (2020) investigated the machine learning techniques to predict the grade point average of students based on personal characteristics, university entry scores, gap year, and their academic performance of the first and second year. Zhang et. al (2021) provided a systematic review of the student performance prediction studies from the perspective of machine learning and data mining considering five stages, i.e., data collection, problem formalization, model, prediction, and application.

In all the abovementioned studies, the contributions of the features to the performance of the classification of academic performance were analyzed by applying either suboptimal feature selection algorithms or no feature selection algorithm. On the other hand, in our work, the relationship between various features and classifiers is analyzed for the same classification task using an optimal feature selection algorithm rather than a suboptimal one. For this purpose, 10 different classifiers were used on the Students' Academic Performance dataset (Amrieh et al., 2016). The classifiers include naive Bayes (Gaussian), naive Bayes (Bernoulli), *k*-nearest neighbor, support vector machine (kernel: radial basis function), support vector machine (kernel: polynomial), logistic regression, decision tree (criterion: entropy), decision tree (criterion: Gini), random forest (criterion: entropy), and random forest (criterion: Gini) (Bishop, 2006; Theodoridis and Koutroumbas, 2009). The dataset consists of 15 features and 480 samples in total. The best feature subsets offering the highest classification accuracy for each classifier were obtained using the exhaustive search (Gunal et al., 2009), which is an optimal feature selection algorithm. In this way, classification accuracies of up to 79.40% were achieved using as few as 8 features out of

15.

The rest of the paper is organized as follows: The materials and methods are described in Section 2. The experimental results and discussion are presented in Section 3. Finally, the conclusions and future directions are given in Section 4.

2. Materials and Methods

As mentioned earlier, various classification algorithms were used to classify the Students' Academic Performance dataset. The contributions of the features to the performance of each classifier were then analyzed using the exhaustive feature selection algorithm. In the following subsections, first, the dataset is introduced. Then, the preprocessing steps, classifiers, and success metrics are described. Finally, the optimal feature selection algorithm is explained. In this work, research and publication ethics were followed.

2.1. Dataset

The Students' Academic Performance dataset contains 15 features and 480 instances. The students were divided into three classes namely Low (L), Middle (M), and High (H) according to their success levels. Students with grades 0-69 are in L, 70-89 are in M and 90-100 are in H class. 30% of the students are in H, 44% are in M and 26% are in L class. The features were divided into four categories including demographic, academic background, parent participation in the learning process, and behavioral.

The student's nationality, place of birth, gender, and parent (father or mother) features are in the demographic feature category. The nationalities of the students are 37% Kuwait, 36% Jordan, 6% Palestine, 5% Iraq, and 16% others. 64% of the students are male and 36% are female. Their birthplaces are 37% Kuwait, 36% Jordan, 6% Palestine, 5% Iraq, and 16% other countries. The distributions of the demographic features are illustrated in Figure 1.

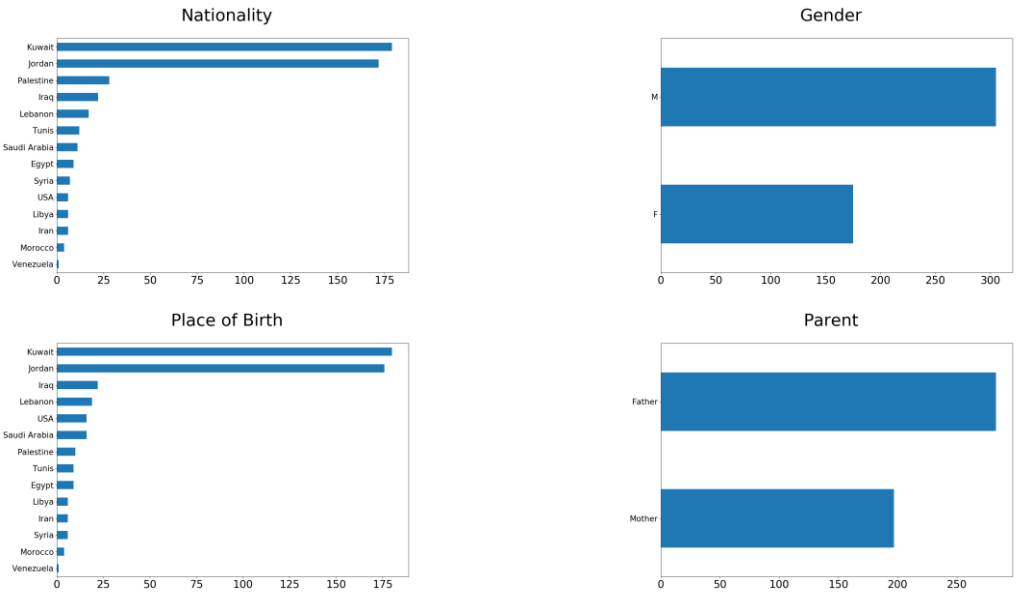

Figure 1. Distribution of the demographic features.

In the academic background category, the features are stage (lower, middle, or high school), grade (from 2 to 12), section (A, B, C), semester, topic (English, math, biology, etc.), and the number of absences (above or under 7 days). 52% of the students attend secondary school, 41% attend primary school and 7% attend high school. 31% are 2nd grade, 24% are 8th grade, 21% are 7th grade, 10% are 4th grade and 14% are other grades. Section distributions are 59% A, 35% B, and 6% C. 51% of the data were collected in the first semester and 49% in the second semester. The distribution of the topics is 20% IT, 14% French, 12% Arabic, 11% science, and 44% others. 60% of students are absent under 7 days and 40% are over 7 days. The distributions of the academic background features are illustrated in Figure 2.

Whether the parent participated in the survey about the school and parent's satisfaction with the school is in the parent's participation category. More than half of the parents participated in the survey. While 61% of the parents are satisfied with the school, 39% are not. The distributions in this feature category are illustrated in Figure 3.

The category of behavioral features includes participation in discussion groups, access to lecture resources, raising hand in class, and viewing

announcements. The features in this category consist of numerical values.

2.2. Pre-Processing

There are both categorical and numeric values in the dataset. While categorical features can be nominal or ordinal, numeric features are either interval or ratio (Han et al., 2011; Aggarwal, 2015). The nominal scale is a labeling scale, where features are simply labeled, with no specific order. On the other hand, the ordinal scale has all its variables in a specific order, beyond just labeling them. Interval features can be categorized and ranked, and evenly spaced. Ratio feature can be categorized, ranked, evenly spaced, and has a natural zero. Hence, in the dataset, label encoding should be used for ordinal values like low, middle, and high. However, for nominal values like section (A, B, C), one-hot encoding should be used. Nationality, place of birth, and topic features are not included in this work since the dimension would be very high for exhaustive search after the one-hot encoding. The section feature was one-hot encoded as shown in Table 1. Other categorical features were expressed using 0 and 1 with one variable each since the values are binary (parent: mother/father, gender: male/female, and so on). Finally, a 15-dimensional feature vector with numeric values was obtained. Also, a standard scaler (zero mean and unit variance) was used to normalize the data.

2.3. Classifiers and Success Metrics

In this study, decision tree (DT), naive Bayes (NB), *k*-nearest neighbor (kNN), logistic regression (LR), random forest (RF), support vector machine (SVM) classifiers were used. Accuracy, precision, recall, and F1-score were the success metrics used to select the best feature subsets. First, the subset with the highest accuracy was selected. If the accuracies of any two subsets were equal, then their F1-scores were compared. The confusion matrix including true positive, true negative, false positive, and false negative values is given in Table 2. Accuracy, precision, recall, and F1-score are calculated using the formulations in (1-4).

2.4. Feature Selection

Feature selection methods are mainly divided into two categories: filters and wrappers (Guyon and Elisseeff, 2003; Gunal and Edizkan, 2008). While most of the feature selection algorithms offer suboptimal results due to the considerations on processing time, the exhaustive search strategy can provide the optimal feature subset, but with a burden of significantly higher processing time. Therefore, researchers may not prefer the exhaustive search for large numbers of features. As mentioned earlier, the exhaustive feature selection algorithm providing the optimal feature subset was utilized in our work due to the relatively small size of the initial feature set. In general, all of 2^{n-1} feature combinations are tested in an exhaustive search for n features. Since there were 15 distinct features in our work, 32,767 feature combinations were compared to each other for every single classification algorithm.

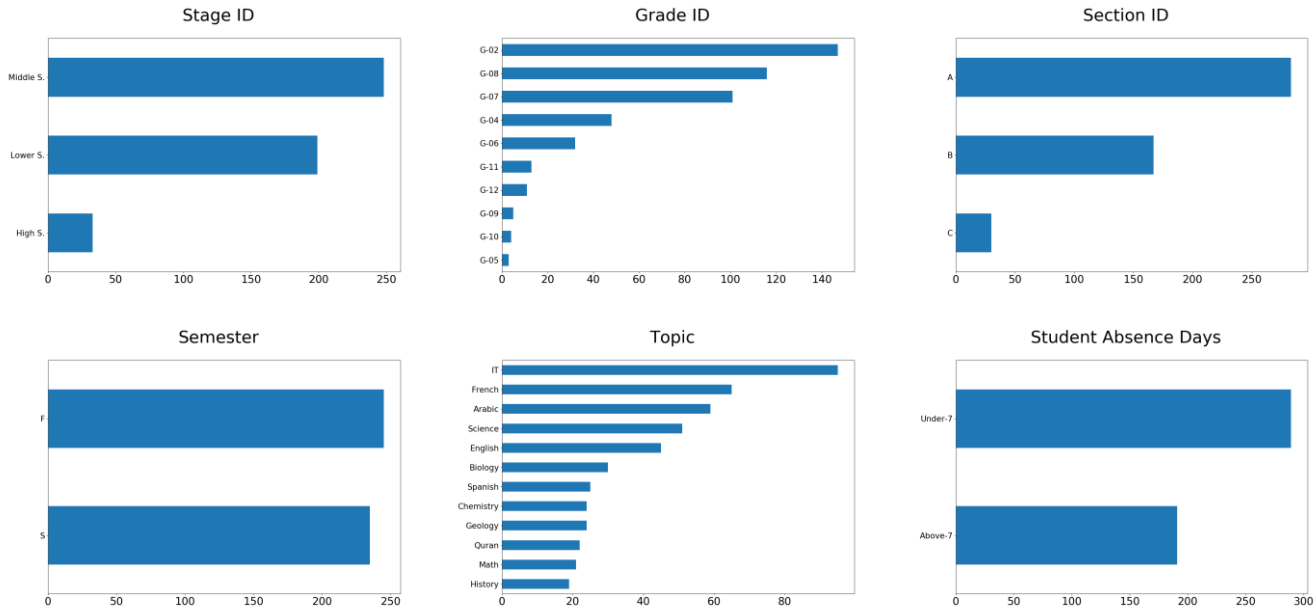

Figure 2. Distribution of the academic background features

Figure 3. Distribution of the parent participation features

Table 1

One-hot encoded section feature

Section	Section A	Section B	Section C
A	1	0	0
B	0	1	0
C	0	0	1

Table 2

The layout of a confusion matrix

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3. Experimental Results and Discussions

During the experimental work, 1/3 of the instances were used for testing. The remaining 2/3 were used to train the models with 10-fold stratified cross-validation.

The best feature subsets were obtained for different numbers of features ranging from 1 to 15. The test dataset was used to calculate the classification accuracies of the best feature combinations found by cross-validation. The best feature subset was determined by comparing the accuracies, or F1-scores if the accuracies were the same. The combination of fewer features was selected if both the accuracy and F1-score were equal.

After executing the exhaustive feature selection algorithm, the optimal feature subsets providing the highest classification accuracies for each classifier were obtained. These feature subsets are listed in Table 3, where the selected features are indicated with "x". It is seen from the table that parent, parent answering survey, and the number of absences are important features for all of the classifiers.

The classification accuracies of each classifier and class-specific precision, recall, F1-score values for the best feature subsets in each case are listed in Table 4, where the highest values are indicated in bold. As shown in the table, the F1-score of class M is lower than those of the other two classes. The highest precision, recall, F1-score, and accuracy were all achieved by the random forest (criterion: Gini) classifier, whereas the lowest ones were obtained with the support vector machine (kernel: RBF).

The classification accuracies with and without feature selection are comparatively given in Table 5, where the highest values are indicated in bold. Also, the results were compared with the related work (Amrieh et al., 2016) using the same dataset. It is clear from the table that naive Bayes and random forest classifiers with the corresponding optimal feature subsets achieved superior results than those of the related work. In our work, the most successful classifier was found to be the random forest (criterion: Gini) with an accuracy of 79.40%. Moreover, only 8 of the initial feature set was selected with this classification algorithm. These 8 features were found to be gender, parent, raised hands, visited resources, announcements view, parent answering survey, the number of absences, and section B.

Also, the changes in the classification accuracy for the best feature subsets consisting of different numbers of features ranging from 1 to 15 are illustrated in Figure 4. While the naive Bayes (Gaussian) classifier provided the best performance in most of the feature dimensions, the highest performance (an accuracy of 79.40% with only 8 features selected) was attained with the random forest classifier (criterion: Gini). In the meantime, the lowest performances for each classifier were achieved when only a single feature was selected. In that case, the accuracy dropped even under 50% for some of the classifiers.

Table 3

The best feature subsets for each classifier, where the selected features are indicated with "x".

Feature Category	No	Feature	GNB	BNB	kNN	SVM (RBF)	SVM (Poly)	LR	DT (Ent.)	DT (Gini)	RF (Ent.)	RF (Gini)
Demographic	1	Gender	x	x	x	x	x		x	x	x	x
	2	Parent	x	x	x	x	x	x	x	x	x	x
Academic Background	3	Stage	x				x	x	x	x	x	
	4	Grade						x	x	x	x	
	5	Section A	x					x	x	x		
	6	Section B	x	x	x		x	x	x	x	x	x
	7	Section C			x		x	x	x	x		
	8	Semester	x						x	x		
	9	The number of absences	x	x	x	x	x	x	x	x	x	x
Parent Participation	10	Parent answering survey	x	x	x	x	x	x	x	x	x	x
	11	Parent school satisfaction	x				x	x	x	x	x	
Behavioral	12	Discussion	x					x	x		x	
	13	Visited resources	x	x				x	x	x	x	x
	14	Raised hands						x	x	x	x	x
	15	Announcements view	x					x	x	x	x	x

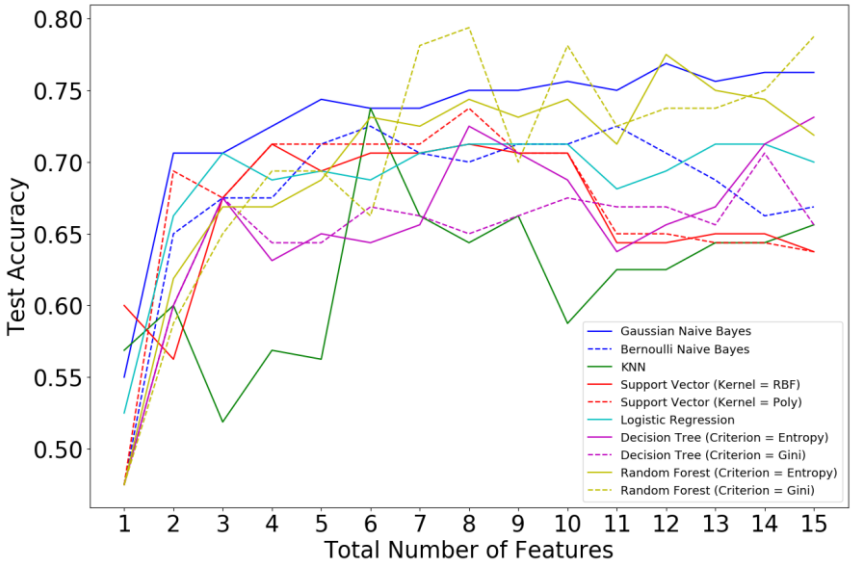

Figure 4. The accuracies of each classifier for the best feature subsets with different numbers of features

4. Conclusions

In this work, the contributions of features for the automatic classification of students' academic performance were analyzed extensively. With the help of an exhaustive search strategy, the optimal subset of features providing the highest classification performance is obtained for 10 different classification algorithms. Extensive experimental analysis on an academic performance dataset revealed that the accuracy of the classification of students' academic performance can increase up to 79.40% using only 8 features rather than all. These features were found to be gender, parent, raised hands, visited resources, announcements view, parent answering survey, the number of absences, and section B. Also, a classification accuracy of around 60% was achieved even with a single feature. As future work, analyses of different features and classification algorithms can be carried out.

Conflict of Interest

No conflict of interest was declared by the authors.

Author Contributions

H. A. Eren contributed to the implementation of the research, to the analysis of the results, and to the writing of the manuscript. E. Sora Gunal contributed to the design of the research, to the analysis of the results, and to the writing and reviewing of the manuscript.

Table 4

Accuracy, precision, recall, and F1-scores of the classifiers

Classifier	Class	Precision	Recall	F1-score	Accuracy
Naive Bayes (Gaussian)	L	0.76	0.91	0.83	0.77
	M	0.76	0.65	0.70	
	H	0.78	0.82	0.80	
	Weighted Avg.	0.77	0.77	0.76	
Naive Bayes (Bernoulli)	L	0.75	0.86	0.80	0.73
	M	0.65	0.71	0.68	
	H	0.82	0.67	0.73	
	Weighted Avg.	0.73	0.72	0.73	
kNN	L	0.72	0.83	0.77	0.74
	M	0.68	0.68	0.68	
	H	0.82	0.75	0.78	
	Weighted Avg.	0.74	0.74	0.74	
SVM (kernel: RBF)	L	0.74	0.80	0.77	0.71
	M	0.63	0.71	0.67	
	H	0.82	0.67	0.73	
	Weighted Avg.	0.72	0.71	0.71	
SVM (kernel: Poly)	L	0.72	0.89	0.79	0.74
	M	0.68	0.68	0.68	
	H	0.83	0.72	0.77	
	Weighted Avg.	0.74	0.74	0.74	
Logistic Regression	L	0.79	0.89	0.84	0.71
	M	0.64	0.68	0.66	
	H	0.75	0.65	0.70	
	Weighted Avg.	0.71	0.71	0.71	
Decision Tree (criterion: Entropy)	L	0.72	0.83	0.77	0.73
	M	0.69	0.63	0.66	
	H	0.77	0.78	0.78	
	Weighted Avg.	0.73	0.73	0.73	
Decision Tree (criterion: Gini)	L	0.57	0.77	0.66	0.71
	M	0.70	0.57	0.63	
	H	0.82	0.82	0.82	
	Weighted Avg.	0.72	0.71	0.71	
Random Forest (criterion: Entropy)	L	0.77	0.86	0.81	0.78
	M	0.71	0.77	0.74	
	H	0.86	0.73	0.79	
	Weighted Avg.	0.78	0.78	0.78	
Random Forest (criterion: Gini)	L	0.79	0.86	0.82	**0.79**
	M	0.74	0.78	0.76	
	H	0.87	0.77	0.81	
	Weighted Avg.	0.80	0.79	0.79	

Table 5

Comparison of the classification accuracies with the related work

Classifier	Without Feature Selection	With Feature Selection	Related Work (Amrieh et al., 2016)
Naive Bayes (Gaussian)	0.762	0.769 (12 features)	0.670
Naive Bayes (Bernoulli)	0.669	0.725 (6 features)	
k-Nearest Neighbor	0.656	0.738 (6 features)	-
Support Vector Machine (kernel: RBF)	0.638	0.712 (4 features)	-
Support Vector Machine (kernel: Poly)	0.638	0.738 (8 features)	
Logistic Regression	0.700	0.712 (13 features)	-
Decision Tree (criterion: Entropy)	0.731	0.731 (15 features)	0.750
Decision Tree (criterion: Gini)	0.656	0.706 (14 features)	
Random Forest (criterion: Entropy)	0.719	0.775 (12 features)	0.750
Random Forest (criterion: Gini)	**0.788**	**0.794** (8 features)	

References

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Amrieh, E.A., Hamtini, T.M., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies (AEECT), 1-5. doi: https://doi.org/10.1109/AEECT.2015.7360581

Amrieh, E.A., Hamtini, T.M., & 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. doi: http://dx.doi.org/10.14257/ijdta.2016.9.8.13

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Gunal, S., & Edizkan, R. (2008). Subspace based feature selection for pattern recognition. Information Sciences, 178(19), 3716-3726. doi: https://doi.org/10.1016/j.ins.2008.06.001

Gunal, S., Gerek, O. N., Ece, D. G., & Edizkan, R. (2009). The search for optimal feature set in power quality event classification. Expert Systems with Applications, 36(7), 10266-10273. doi: https://doi.org/10.1016/j.eswa.2009.01.051

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An empirical study for student academic performance prediction using machine learning techniques. International Journal of Computer Science and Information Security, 18(3), 21-28.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining concepts and techniques.* Elsevier.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Computers & Education, 61, 133-145. doi: https://doi.org/10.1016/j.compedu.2012.08.015

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459. doi: http://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Rahman, M. H., & Islam, M. R. (2017). Predict student's academic performance and evaluate the impact of different attributes on the performance using data mining techniques. IEEE 2nd International Conference on Electrical & Electronic Engineering (ICEEE), 1-4. doi: https://doi.org/10.1109/CEEE.2017.8412892

Sana, Siddiqui, I. F. & Arain, Q. A. (2019). Analyzing students' academic performance through educational data mining. 3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, 402–421. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.402-421

Theodoridis, S., & Koutroumbas, K., (2009). *Pattern recognition (4th ed.)*. Academic Press.

Zaffar, M., Hashmani, M. A., & Savita, K. S. (2017). Performance analysis of feature selection algorithm for educational data mining. IEEE Conference on Big Data and Analytics (ICBDA), 7-12. doi: https://doi.org/10.1109/ICBDAA.2017.8284099

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in Psychology, 12, 698490-698490. doi: https://doi.org/10.3389/fpsyg.2021.698490