Bartın University Journal of Faculty of Education
BUEFAD Volume 14, Issue 1

181-195
dergipark.org.tr/buefad
DOI: 10.14686/buefad.1390209

Prediction of Academic Success and Selection of Effective Features with Machine Learning

Furkan AYDIN [a*]

a Teacher, Ministry of National Education, https://orcid.org/0000-0003-0610-8744 *frknaydn44@gmail.com

Research Article
Received: 13.11.2023
Revised: 16.6.2024
Accepted: 13.10.2024

Abstract

This study aims to determine whether the success or failure of students after the measurement process can be predicted in terms of some variables, with the help of machine learning algorithms. The research problem statement is "As a result of the evaluation and selection of some variables affecting the student's academic success using machine learning algorithms, are the determined variables more effective than other variables in predicting student success?". The study was conducted with causal-comparative research, one of the quantitative research designs. In this paper, machine learning algorithms have been applied to the dataset that consists of major features to predict students' performances. Thus, the most significant features and the highest-performance machine learning algorithm have also been detected. To this end, univariate, tree-based, and L1-based feature selection methods have been used for the feature selection process. Classification and regression trees, k-nearest neighbors, naive Bayes, random forest, and support vector machines have been employed to build the learning models. In consequence, there exist lots of indicators that impact students' academic successes, the success or failure that emerges after the measurement process can be estimated by regarding these features in advance. Such a task will enable the relationship mechanism between the educational inputs and outputs to be understandable and eliminate shortcomings concerning the education process.

Keywords: Academic performance, academic achievement, artificial intelligence, educational data mining, feature selection

Makine Öğrenmesi ile Akademik Başarının Tahmini ve Etkili Özelliklerin Seçimi

Öz

Bu çalışma öğrencilerin ölçme süreci sonrasında ortaya çıkan başarı ya da başarısızlık durumunun bazı değişkenler açısından önceden tahmin edilip edilemeyeceğinin makine öğrenmesi algoritmaları yardımıyla belirlemeyi amaçlamaktadır. Araştırmanın problem cümlesi: "Öğrencinin akademik başarı sonucuna etki eden bazı değişkenlerin makine öğrenmesi algoritmaları uygulanarak değerlendirilmesi ve seçilmesi sonucunda belirlenen değişkenlerin öğrenci başarısını tahmin etmede diğer değişkenlere göre daha fazla etkili midir?". Çalışma, nicel araştırma desenlerinden nedensel karşılaştırma araştırması ile gerçekleştirilmiştir. Bu çalışmada öğrencilerin performanslarını tahmin etmek amacıyla en temel özelliklerden oluşan bir veri setine makine öğrenmesi algoritmaları uygulanmıştır. Böylece en önemli özellikler ve en yüksek performanslı makine öğrenmesi algoritması da tespit edilmeye çalışılmıştır. Bu amaçla özellik seçim sürecinde tek değişkenli özellik seçimi, ağaç tabanlı özellik seçimi ve L1 tabanlı özellik seçimi yöntemleri kullanılmıştır. Öğrenme modellerini oluşturmak için sınıflandırma ve regresyon ağaçları, k-en yakın komşular, naive Bayes, rastgele orman ve destek vektör makineleri kullanılmıştır. Sonuç olarak öğrencilerin akademik başarılarını etkileyen pek çok gösterge mevcut olup, ölçme süreci sonrasında ortaya çıkan başarı ya da başarısızlık, bu özellikler dikkate alınarak önceden tahmin edilebilmektedir. Böyle bir görev, eğitimsel girdi ve çıktılar arasındaki ilişki mekanizmasının anlaşılmasını sağlayacak ve eğitim sürecine ilişkin eksiklikleri ortadan kaldıracaktır.

Anahtar Sözcükler: Akademik performans, akademik başarı, yapay zekâ, eğitimsel veri madenciliği, özellik seçimi

To cite this article in APA Style:

Aydın, F. (2025). Prediction of academic success and selection of effective features with machine learning. *Bartın University Journal of Faculty of Education, 14*(1), 181-195. https://doi.org/10.14686/buefad.1390209

INTRODUCTION

The intelligence concept has been defined a lot in literature. The definition by Lenat and Feigenbaum is that "intelligence is the capability to find fast in a search space an acceptable solution that is a priori for observers" (Lenat & Feigenbaum, 1991). Thus, Artificial Intelligence (AI) can be defined as a research area that uses computational models to represent intelligence and instills intelligent behaviors in automata. One of the approaches to AI is the Turing test. With respect to the Turing test approach, learning is regarded as adapting to new circumstances, finding patterns, and deducing from them (Aziz & Memon, 2023; Russell & Norvig, 2010). In this regard, Machine Learning (ML) is a study discipline that aims to reveal patterns in datasets to solve real-world problems and automata acquire abilities regarding learning. Supervised learning is one of the ML sub-research areas frequently exploited in lots of problems. Given a training set, the supervised predictors explore a function that is a member of the hypothesis space.

Applying ML algorithms and data transformation techniques to datasets is known as data mining (DM). In other words, the process of analyzing a large number of data and turning them into knowledge is defined as data mining. Educational data mining (EDM) is used for purposes such as increasing student success by analyzing student data, detecting deficiencies in education and training, and creating effective education and training environments (Özkan, 2015). EDM is a subfield of the DM area. EDM has contributed to theories of learning studied by scholars in educational psychology and the learning sciences (Ryan Shaun Baker & Inventado, 2014). Education is an indispensable factor so as to ensure the development and growth of countries. In this respect, the prediction of students' performance is significant before they take exams or courses. Thus, students' shortcomings and insufficiencies can be corrected. Further, increasing the quality of education is required throughout the current semester so as to enhance students' achievements (Yılmaz & Sekeroglu, 2020). In the education field, data mining and analysis will be used to help individuals discover learning and understand learning behavior, and will also enable a significant level of quality to be achieved in the education system outputs (Özdemir et al., 2018). Furthermore, EDM can be used to design better smart learning technologies for educational goals and to make learners and educators better informed (Ryan S. Baker, 2014).

Since advanced countries establish their governance models on the necessity of change, performance measurement, and efficient use of resources, they should monitor their performance in other countries that aim to keep up with this change and their place in the world with the support of various mechanisms (Acar, 2022). When the 2018 PISA (Programme for International Student Assessment) results and the socio-economic parameters of the countries are investigated, it is seen that the PISA success of countries with a high level of development is the same as their level of development (Yüksel, 2022). Besides, families' thoughts on participation in education vary greatly depending on their total incomes and the amount they spend on activities that will contribute to students' education in out-of-school learning environments (Yıldırım, 2020). Even though teachers who work for schools of different socioeconomic statuses agree on concepts such as academic success, a desire to learn from within the student, and student commitment to the school community, truly, this situation stems from the expectations of the environment or school nowadays (Kazak, 2021). As can be seen from the abovementioned views, there are many parameters that affect students' academic successes, the success or failure that appears at the end of the measurement process can be predicted by considering these parameters previously. This will enable the feedback mechanism between the inputs and outputs of education to be more active and to eliminate deficiencies regarding the education process.

Many works on predicting students' performance have been carried out. However, there are some differences between this study and similar works. These are as follows:

- Other works (Pallathadka et al., 2023; Salah Hashim et al., 2020; Yılmaz & Sekeroglu, 2020) focus on estimating students' final grades. This work concentrates on students' general performance. Hence, the number of categorical values of the outcome has been reduced to two from eight: 'fail' and 'pass'.

- In this work, while the most successful model is a transparent model (e.g., tree-based models), the performance of black box models (e.g., instance-based learning, function learning, Bayesian learning) is higher in the other works (Ismail et al., 2021; Pallathadka et al., 2023; Salah Hashim et al., 2020; Yılmaz & Sekeroglu, 2020). Transparent models can be comprehended by human experts and a new knowledge extraction is possible in such models. However, comprehension of black box models is difficult and can be even impossible for human experts. Meanwhile, Sa et al. have used only tree-based models in their study. However, the performance of these models is low (Sa et al., 2014). Additionally, Shanmugarajeshwari and Lawrance have used a C5 model in their study (Shanmugarajeshwari & Lawrance, 2016). However, the dataset where they have used is uncomplicated compared to the dataset used in this study.

- The feature selection process is important to boost the performance of ML algorithms. In this respect, powerful feature selection methods based on a filter approach have been used in this work. However, an in-depth feature selection study was not seen in similar studies reviewed in the literature.

METHOD

In this section, the dataset used in this research has been explained, the feature selection method that is employed to select significant features, machine learning algorithms that are required to discover patterns in the dataset, and evaluation criteria to quantify the performance of the models. Finally, the Method section has been finalized by mentioning the experimental procedure. Figure 1 summarizes the overview of the whole learning process.

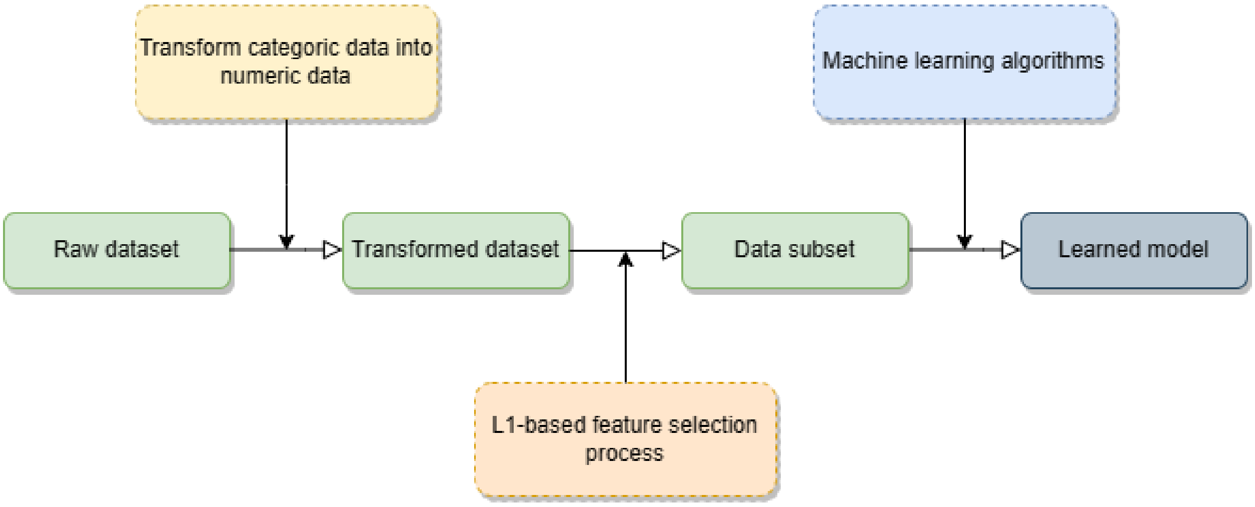
Transform categoric data into numeric data
Machine learning algorithms
Raw dataset
Transformed dataset
Data subset
Learned model
L1-based feature selection process

Figure 1. The overview of the learning process

Students Performance Dataset

The Students' Performance dataset (Yılmaz & Sekeroglu, 2020) from the Kaggle[1] has been used in this research. This dataset consists of questions that 145 students are asked. Each column (i.e., features) of the dataset represents those questions. The first 10 features are personal questions. Features 11-16 include questions regarding family. Finally, the rest of the features are on self-education habits. Table 1 shows the detailed information concerning this dataset. In addition, Table 2 presents the descriptive statistics related to the dataset. In this study, each feature takes categorical values. But categorical values have been transformed into numerical values. The reason is that the ML algorithms used can process datasets that consist of numerical data. The outcome class truly consists of 8 categorical values (i.e., AA, BA, BB, CB, CC, DC, DD, FF). The number of these categorical values has been reduced to 2 categorical values: ‘fail’ and ‘pass’ (i.e., FF, DD, and DC are regarded as ‘fail’. The others are counted as ‘pass’). This is because, at many universities in Türkiye, DC, and DD grades are among the conditionally valid letter grades. In other words, if the student's Cumulative Grade Point Average (CGPA) is above 2.00, the student is considered to have passed these courses. However, if the CGPA is below 2.00, these courses can be taken again and the CGPA can be increased. Besides, the aim of this study is to forecast students' general performance rather than students' final grades.

[1] https://www.kaggle.com/datasets/joebeachcapital/students-performance/data

Understanding the nature of the categorical variables allows us to determine appropriate transformation techniques. Categorical variables can be classified into two groups: ordinal and nominal. Nominal variables lack natural order or hierarchy, but ordinal variables do not. Categorical variables particularly present a variety of challenges in the ML area and DM process (Lee & Kim, 2010; Nisbet et al., 2018). These are as follows:

- Most machine learning algorithms take numerical values as input.
- In the case of assigning numbers in ascending order for categorical variables some ML algorithms may misinterpret these values.
- In the case of high cardinality, encoding the performance of a model on a test set can be low in the case of the fact that a categorical variable occurs in a small number in the dataset. Such variables are computationally expensive for the ML algorithms.

In this study, the whole categories within a categorical variable have been first detected. Afterward, each category is assigned an integer value based on the order of appearance in the data if it is nominal, otherwise, the assignment is carried out according to its inherent order or hierarchy. Categorical values in the original dataset are substituted with their corresponding numerical labels later on. Such an encoding process seems more appropriate for ordinal categorical variables since numerical labels accurately reflect the natural order between the categories.

Table 1. The detailed information about the features of the dataset (data counts for each feature are given in parentheses)

#	Features	Categorical values
F1	Student age	18-21 (Count: 65), 22-25 (Count: 70), above 26 (Count: 10)
F2	Gender	Female (Count: 58), male (Count: 87)
F3	Graduated high-school type	Private (Count: 25), state (Count: 103), other (Count:17)
F4	Scholarship type	None (Count: 1), 25% (Count: 3), 50% (Count: 76), 75% (Count: 42), full (Count: 23)
F5	Additional work	Yes (Count: 49), no (Count: 96)
F6	Regular artistic or sports activity	Yes (Count: 58), no (Count: 87)
F7	Do you have a partner?	Yes (Count: 61), no (Count: 84)
F8	Total salary ($)	135-200 (Count: 93), 201-270 (Count: 27), 271-340 (Count: 16), 341-410 (Count: 4), above 410 (Count: 5)
F9	Transportation to the university	Bus (Count: 98), private car/taxi (Count: 25), bicycle (Count: 1), other (Count: 21)
F10	Accommodation type in Cyprus	Rental (Count: 68), dormitory (Count: 49), family (Count: 27), other (Count: 1)
F11	Mother's education	Primary school (Count: 54), secondary school (Count: 27), high school (Count: 39), bachelor (Count: 21), MSc. (Count: 2), Ph.D. (Count: 2)
F12	Father's education	Primary school (Count: 29), secondary school (Count: 36), high school (Count: 46), bachelor (Count: 28), MSc. (Count: 5), Ph.D. (Count: 1)
F13	Number of siblings	1 (Count: 27), 2 (Count: 45), 3 (Count: 26), 4 (Count: 23), 5 or above (Count: 24)
F14	Parents' marital status	Married (Count: 127), divorced (Count: 11), died - one of them or both (Count: 7)
F15	Mother's occupation	Retired (Count: 6), housewife (Count: 103), government officer (Count: 16), private sector employee (Count: 18), self-employment (Count: 2)
F16	Father's occupation	Retired (Count: 36), government officer (Count: 22), private sector employee (Count: 35), self-employment (Count: 38), other (Count: 14)
F17	Weekly study hours	None (Count: 29), <5 hours (Count: 74), 6-10 hours (Count: 30), 11-20 hours (Count: 8), more than 20 hours (Count: 4)

F18	Reading frequency (non-scientific books/journals)	None (Count: 27), sometimes (Count: 99), often (Count: 19)
F19	Reading frequency (scientific books/journals)	None (Count: 20), sometimes (Count: 103), often (Count: 22)
F20	Attendance to the seminars/conferences related to the department	Yes (Count: 114), no (Count: 31)
F21	Impact of your projects/activities on your success	Positive (Count: 128), negative (Count: 4), neutral (Count: 13)
F22	Attendance to classes	Always (Count: 110), sometimes (Count: 35)
F23	Preparation to midterm exams 1	Alone (Count: 107), friends (Count: 27), not applicable (Count: 11)
F24	Preparation to midterm exams 2	Closest date to the exam (Count: 123), regularly during the semester (Count: 20), never (Count: 2)
F25	Taking notes in classes	Never (Count: 5), sometimes (Count: 56), always (Count: 84)
F26	Listening in classes	Never (Count: 29), sometimes (Count: 79), always (Count: 37)
F27	Discussion improves my interest and success in the course	Never (Count: 9), sometimes (Count: 70), always (Count: 66)
F28	Flipped classroom	Not useful (Count: 64), useful (Count: 45), not applicable (Count: 36)
F29	Cumulative grade point average in the last semester (/4.00)	<2.00 (Count: 17), 2.00-2.49 (Count: 38), 2.50-2.99 (Count: 25), 3.00-3.49 (Count: 40), above 3.49 (Count: 25)
F30	Expected cumulative grade point average in the graduation (/4.00)	<2.00 (Count: 16), 2.00-2.49 (Count: 38), 2.50-2.99 (Count: 61), 3.00-3.49 (Count: 30), above 3.49 (Count: 0)
Class	Outcome	fail (Count: 67), pass (Count: 78) ‘fail’ includes FF, DD, and DC scores ‘pass’ includes CC, CB, BB, BA, and AA scores Raw class information: FF (Count: 8), DD (Count: 35), DC (Count: 24), CC (Count: 21), CB (Count: 10), BB (Count: 17), BA (Count: 13), AA (Count: 17)

Table 2. The descriptive statistics concerning the transformed dataset

#	Min	Max	Mean	Median	Mode	Variance	Kurtosis	Skewness
F1	1	3	1.62	2	2	0.38	-0.64	0.44
F2	1	2	1.60	2	2	0.24	-1.86	-0.41
F3	1	3	1.94	2	2	0.29	0.50	-0.05
F4	1	5	3.57	3	3	0.65	-0.19	0.37
F5	1	2	1.66	2	2	0.23	-1.54	-0.69
F6	1	2	1.60	2	2	0.24	-1.86	-0.41
F7	1	2	1.58	2	2	0.25	-1.92	-0.32
F8	1	5	1.63	1	1	1.04	2.58	1.76
F9	1	4	1.62	1	1	1.13	0.89	1.55
F10	1	4	1.73	2	1	0.61	-0.84	0.60
F11	1	6	2.28	2	1	1.50	-0.38	0.57
F12	1	6	2.63	3	3	1.32	-0.58	0.20
F13	1	5	2.81	3	2	1.85	-1.14	0.31
F14	1	3	1.17	1	1	0.24	7.34	2.88
F15	1	5	2.36	2	2	0.65	1.29	1.36
F16	1	5	2.81	3	4	1.77	-1.22	-0.03
F17	1	5	2.20	2	2	0.84	1.05	0.90
F18	1	3	1.94	2	2	0.32	0.20	-0.02
F19	1	3	2.01	2	2	0.29	0.51	0.01
F20	1	2	1.21	1	1	0.17	-0.01	1.41
F21	1	3	1.21	1	1	0.35	5.18	2.63
F22	1	2	1.24	1	1	0.18	-0.52	1.22

F23	1	3	1.34	1	1	0.38	1.52	1.64
F24	1	3	1.17	1	1	0.17	5.37	2.42
F25	1	3	2.54	3	3	0.32	-0.43	-0.76
F26	1	3	2.06	2	2	0.46	-0.78	-0.07
F27	1	3	2.39	2	2	0.37	-0.65	-0.44
F28	1	3	1.81	2	1	0.66	-1.38	0.37
F29	1	5	3.12	3	4	1.69	-1.19	-0.08
F30	1	4	2.72	3	3	0.84	-0.69	-0.30
Class	0	1	0.54	1	1	0.25	-2.00	-0.15

Feature Selection

Feature selection is a sub-field of the machine learning study area. Feature selection is the task of choosing a subset of the attributes employed in the model-building process and it is preferred due to three reasons: facilitating the interpretation of models readily (James et al., 2013), decreasing the classifier training time, and preventing overfitting based errors (Bermingham et al., 2015). Feature selection is categorized into four strategies: the filter methods, the wrapper methods, the embedded methods, and the hybrid methods. The filter methods utilize the statistical aspects of the dataset independently of machine learning models (Bolón-Canedo et al., 2013). The wrapper methods search for the most suitable feature subset concerted to a machine learning algorithm by considering the relevance of features to each other (Kohavi & John, 1997). The embedded methods concern the classifiers that conduct good learning about which of the chosen feature subsets. Those methods look like the wrapper method. Whereas in an embedded approach, the learning stage affects the search process. Thereby, this reduces the computational cost, as well. The last approach is a fuse of all the above-mentioned strategies.

In this study, L1-based feature selection has been used, which is a sort of wrapper approach for the feature selection process. L1-based feature selection is a meta-strategy that is used with any classifier that nominates importance to each feature by a specific quantity. The features are supposed trivial if the importance values of the features are underneath a certain threshold value. SVM classifiers (Vapnik & Lerner, 1963) using the linear kernel and L1 norm deliver sparse solutions. Thus, the dimensionality of the data is decreased to choose the non-zero factors. Briefly, a beneficial sparse classifier for this objective is LinearSVC. Finally, the regularization parameter in SVM adjusts the sparsity. The fewer features can be selected with small regularization parameters.

In spite of the popularity of the Lasso as a variable-choosing strategy, the issue of constructing well-founded inferences for a model selected by the Lasso is not greatly resolved. Cai and Yuan have proposed a method called covariance test statistics to test the significance of the predictor variables chosen by Lasso. The researchers argue that the chi-squared test fails when applying to the forward stepwise regression or the Lasso in a vanilla fashion (Lockhart et al., 2014).

Machine Learning Algorithms

Table 3 shows the ML classifiers used in the experiments and some important parameters of them. Accordingly, these algorithms in this sub-section have shortly been mentioned.

Classification and regression trees (CART) are a tree-based machine learning classifier (Breiman et al., 1984). CART builds straightforward but effective models by recursively dividing the data space to extract a hypothesis from the data and commits a greedy technique in which decision trees are created in the form of a top-down recursively to be separated and conducted.

The k-Nearest Neighbors (KNN) classifier is a lazy learning algorithm known to all. The KNN classifiers detect k data points that are nearest to a data point in question in a data space (Beyer et al., 1999). The bias-variance balance of a KNN model can be adjusted by the parameter k (Manning & Raghavan, 2009). Foremost, a distance metric is required to quantify proximity between points in a data space (Han & Kamber, 2006). In this regard, there exist a few conventional distance metrics: Euclidean distance, the City Block distance, etc. In addition to detecting the optimal k value, choosing a fitting distance metric is significant in terms of high classification accuracy (Hechenbichler & Schliep, 2004).

The naive Bayes (NB) is one of the most commonly used classifiers in the ML area. The NB classifier is frequently used in cases where the independent variables are not actually conditionally independent given the dependent variable, as well. NB can perform unexpectedly well, even when the conditional independence assumption is mostly incorrect. In other words, although NB assumes the conditional independence of all input variables, given a single outcome variable, it turns out to accomplish unusually well in numerous applications (Russell & Norvig, 2010).

The random forest (RF) algorithm is a meta-classifier that fits a series of decision tree algorithms on the division of the dataset and uses averaging to increase classification accuracy and prevent high variance problems (Ghosh & Cabrera, 2022). When the number of sub-samples is not controlled with the corresponding hyperparameter, the entire dataset is employed to form each tree.

Support Vector Machines (SVM) are a number of supervised learning models employed for various ML tasks. SVM can run effectively on high-dimensional datasets. Additionally, it is also partly influential on datasets where the number of features is larger than the number of instances. SVM efficiently uses memory by means of support vectors. For linearly inseparable data points, it allows the use of kernel functions. However, despite these advantages of SVM, it has some crucial disadvantages. In cases where the number of dimensions is much larger than the number of data points, the high variance problem should be paid attention. Lastly, SVM does not directly deliver probabilistic estimations (Baudat & Anouar, 2000; Burges, 1998; Cristianini & Shawe-Taylor, 2000; Lin et al., 2007).

Table 3. The machine learning algorithms that are used in the experiments.

Algorithm	Parameter values
Classification and Regression Trees (CART)	Criterion = “gini”, splitter = “best”
K-Nearest Neighbors (KNN)	N_neighbors = 8, weights = “distance”, metric = “euclidean”
Naïve Bayes (NB)	Force_alpha = True
Random Forest (RF)	Bootstrap = True
Support Vector Machines (SVM)	Kernel = “linear”, C = 1

Evaluation Criteria

The classification accuracy rate in Equation (1) and F1-score in Equation (5) to evaluate the performance of the algorithms have been used. ACC specifies the number of instances classified correctly within all the instances.

$$Accuracy = \frac{1}{m}\sum_{i=1}^{m}\delta(a_i, p_i) \tag{1}$$

$$\delta(x, y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases} \tag{2}$$

where a denotes the true class labels and p denotes the predictions. In Equation (3), the precision value specifies the rate of positive instances classified correctly within instances predicted as positive. The recall value in Equation (4) specifies the rate of positive instances classified correctly within instances whose actual classes are positive. F1-score is the harmonic mean of precision and recall values as seen in Equation (5). If the F1-score is greater than or equal to 0.5, it can be said the model learns classes well for each of them.

$$Precision = TP/(TP + FP) \tag{3}$$

$$Recall = TP/(TP + FN) \tag{4}$$

$$F1 - score = 2 \times Precision \times Recall/(Precision + Recall) \tag{5}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. TP indicates instances classified correctly as positive. FP is instances classified incorrectly as positive. FN points instances classified incorrectly as negative.

Area Under the Curve (AUC) is a performance criterion for classification problems at a variety of threshold values and it indicates the measurement of separability. AUC describes how much the model is qualified for differentiating between classes. The higher the AUC, the better the model distinguishes classes better.

Experimental Procedure

In this research, Google Colab as a programming and numeric computing platform has been used and all the experiments have been performed under 10-fold cross-validation (CV). Thus, the classifiers have been trained on the same training and test sets each time. Subsequently, the performance of each classifier has been measured. CV provides a method of separating data into almost equal parts so as to estimate a classifier's performance on a test set.

FINDINGS

This section contains the findings that are needed to reinforce the conclusions of this research. Table 4 shows the comparative results of the classifiers in terms of accuracy, F1-score, precision, recall, and AUC before the feature selection process. In terms of classification accuracy, CART, KNN, RF, and SVM correspond to 0.5929, 0.5590, 0.6367, and 0.6076, respectively. Accordingly, the RF classifier delivered the highest average accuracy rate before the feature selection process. The SVM classifier ranks second after RF. In terms of F1-score, CART, KNN, RF, and SVM correspond to 0.6170, 0.5893, 0.5859, and 0.6145, respectively. Accordingly, the CART classifier yielded the highest average F1-score value before the feature selection process. The SVM classifier ranks second after CART. In terms of the weighted average of the accuracy rate and F1-score values, the RF classifier ranks first by 0.6113. Whereas the SVM classifier ranks second by 0.6111. As a result, it can be said that the RF classifier is successful on the original dataset compared to the other classifiers. Additionally, the RF classifier builds transparent models that can be understandable by human experts. In terms of training time, the RF classifier takes 3.6 seconds. But this time can be reduced by adjusting some hyperparameters of the RF classifier. In this case, the performance of the model may fall. However, this dataset does not have a large amount of data. On average, CART, SVM, KNN, and NB take 118, 149, 81, and 94 milliseconds, respectively.

Table 4. The comparison of the classifiers in terms of accuracy, F1-score, precision, recall, and AUC before the feature selection process

Algorithm	Accuracy	Precision	Recall	F1 score	AUC
CART	0.5929	0.6332	0.6143	**0.6170**	0.5929
KNN	0.5590	0.5800	0.6179	0.5893	0.6267
RF	**0.6367**	**0.6819**	**0.7054**	0.5859	**0.6618**
SVM	0.6076	0.6588	0.6161	0.6145	0.6380

Table 5 shows the comparative results of the classifiers in terms of accuracy, F1-score, precision, recall, and AUC after the feature selection process. The features F4, F8, F9, and F29 have been selected by the L1-based feature selection method after the feature selection process. Meanwhile, the feature F29 (i.e., Cumulative grade point average in the last semester) is important to estimate students' general performance, not final grades. Therefore, the fact that this feature was selected during the Feature selection process is also compatible with our intuition. In terms of classification accuracy, CART, KNN, NB, RF, and SVM correspond to 0.7700, 0.7781, 0.7100, 0.7295, and 0.7033, respectively. Accordingly, the KNN classifier delivered the highest average accuracy rate before the feature selection process. The CART classifier ranks second after KNN. In terms of F1-score, CART, KNN, NB, RF, and SVM correspond to 0.7888, 0.7848, 0.7212, 0.7190, and 0.7256, respectively. Accordingly, the CART classifier yielded the highest average F1-score value before the feature selection process. The KNN classifier ranks second after CART. In terms of the weighted average of the accuracy rate and F1-score values, the KNN classifier ranks first by 0.7815. Whereas the CART classifier ranks second by 0.7794. As a result, it can be said that the KNN classifier is successful on the transformed dataset compared to the other classifiers. However, the results of the KNN and CART classifiers are close to each other. Besides, while the CART classifiers generate transparent models, the KNN classifiers build black box models. In other words, the tree-based models are comprehensible to human experts. Therefore, the tree-based model has been preferred in this study.

Table 5. The comparison of the classifiers in terms of accuracy, F1-score, precision, recall, and AUC after the feature selection process

Algorithm	Accuracy	Precision	Recall	F1 score	AUC
CART	0.7700	0.8071	**0.7911**	**0.7888**	0.7260
KNN	**0.7781**	**0.8227**	0.7643	0.7848	0.7533

NB	0.7100	0.7566	0.7161	0.7212	0.7368
RF	0.7295	0.7698	0.7393	0.7190	0.7440
SVM	0.7033	0.7230	0.7554	0.7256	**0.7536**

For a suitable selection of the regularization parameter (i.e., alpha), the Lasso can entirely rescue the exact set of non-zero variables employing only a few observations, providing that typical conditions are encountered. Especially, the number of instances should be adequately big, otherwise, L1 models will randomly run due to reasons such as the number of features, and the amount of noise. There does not exist a widespread rule to choose an appropriate alpha value for the rescue of non-zero coefficients. This selection can be performed by LassoCV or LassoLarsCV. However, this strategy might cause under-penalized models. Unlike those, LassoLarsIC is disposed to determine high values of alpha. Figure 2 explains the changes in the accuracy and F1-score values of the CART classifier according to the regularization parameter during the feature selection process. Accordingly, the highest accuracy and F1-score values have been obtained while the value of the regularization parameter is 0.02. While the value of the regularization parameter increases, the accuracy rate and F1-score values of the CART classifier decrease, as well. On the contrary, the number of features of the model increases. Consequently, the L1-based feature selection method has selected the features F4, F8, F9, and F29 during the feature selection process.

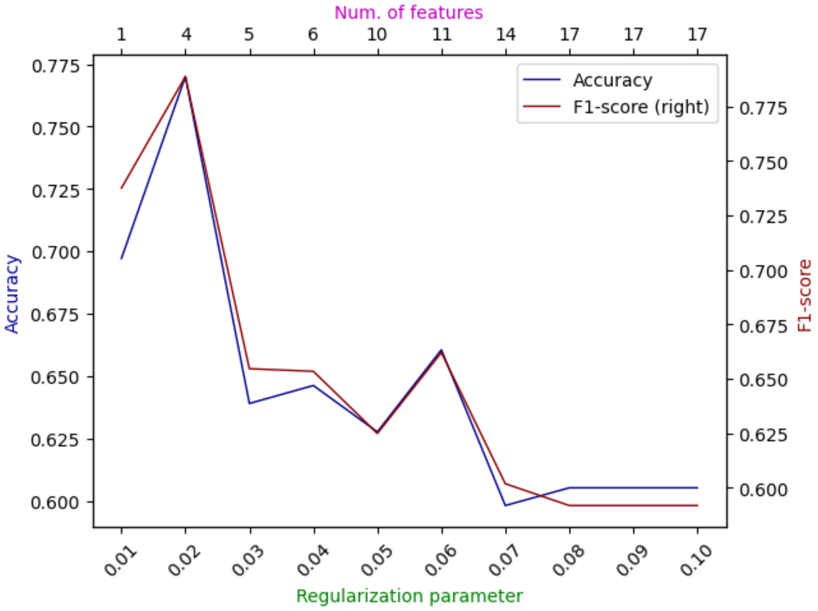

Figure 2. The changes in the accuracy and F1-score values of the CART classifier according to the regularization parameter during the feature selection process.

The CART classifier provides a strategy named cost complexity pruning to prevent a tree from overfitting or from another perspective, to control the size of a tree. This pruning approach is parameterized by alpha. Larger values of alpha increase the number of nodes pruned. Figure 3 shows the change in the accuracy rate of the CART classifier on training and testing sets according to the change in the hyperparameter alpha. Accordingly, the highest alpha value is selected by keeping on high accuracy rate on the test set. Thus, the eleventh model has been selected. The accuracy rate of this model is 0.7919. Namely, the accuracy rate of the model has increased a little. Also, the decision tree model is simplified.

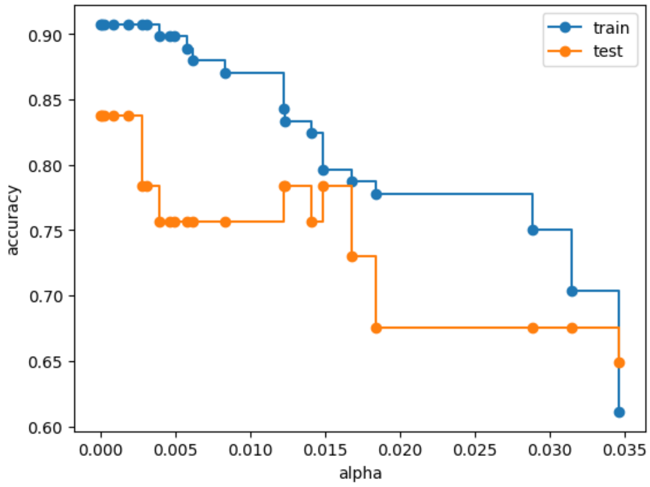

Figure 3. Accuracy vs alpha for training and test sets in terms of the CART model

Figure 4 shows the confusion matrix of the model that is constructed by using the CART classifier. Accordingly, the number of instances that are predicted correctly is 62 for the class 'pass'. For the class 'fail', the number of instances that are predicted correctly is 50. Thus, the number of instances that are predicted correctly is 112 in total. The 33 instances are in total predicted incorrectly. For both 'fail' and 'pass' classes, the number of instances predicted incorrectly is almost the same.

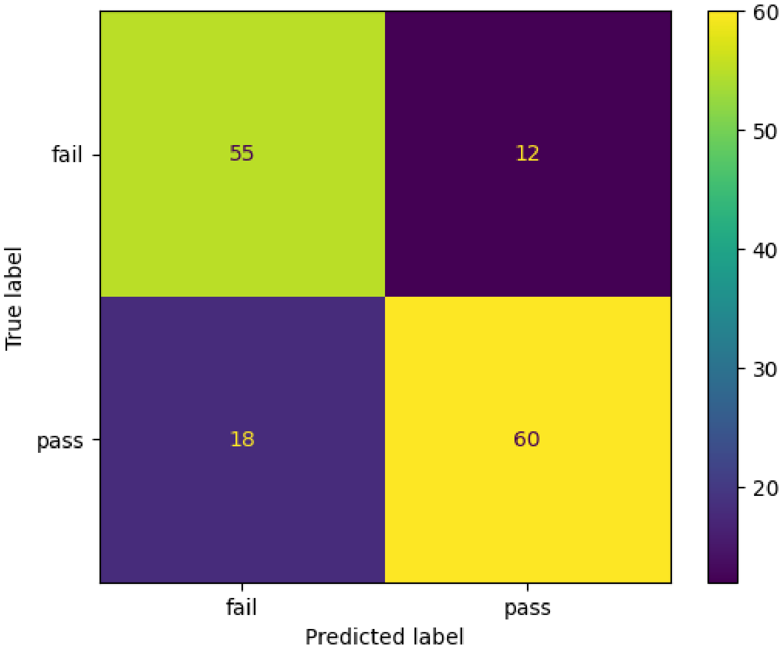

Figure 4. The confusion matrix for the CART classifier after post pruning

Figure 5 shows the decision tree model of the CART classifier. The first row in each node represents the threshold value of the feature of interest (e.g., F29 <= 2.5 for the root). The second row corresponds to Gini index value calculated for situation of interest. Speaking of which, the lower the Gini index, the more homogeneous the node is. The third row denotes the number of instances used in the calculation of the Gini index. As for the fourth row, for instance, the root node has value = [67, 78] which points there are 67 samples of class 'fail' and 78 samples of class 'pass' at the root node. Traversing the tree, the instances are divided, and hereby, the value array reaching each node changes. The last row indicates the majority class after the samples are split.

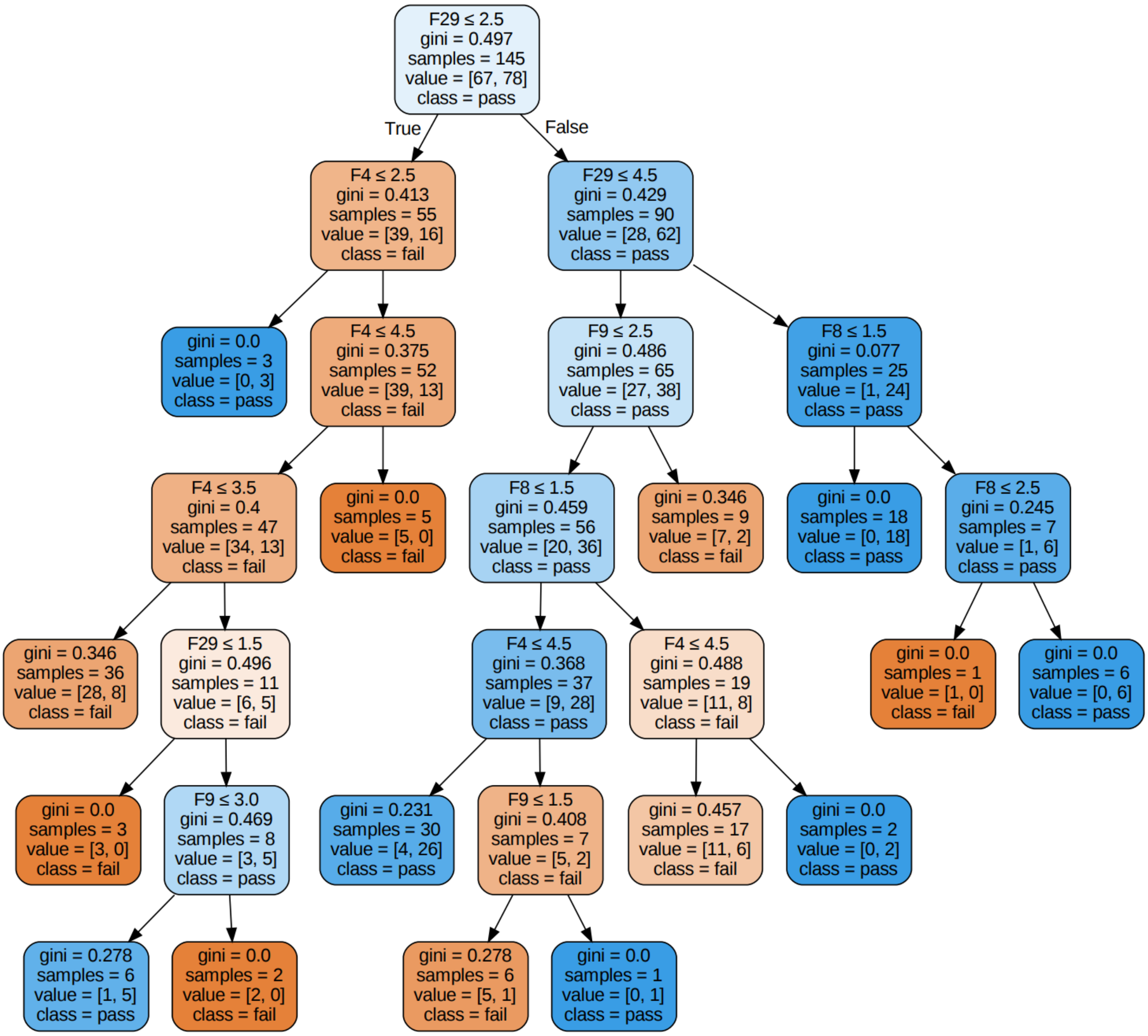

Figure 5. The decision tree model of the CART classifier after post pruning

Figure 6 shows the decision surfaces of the CART classifier trained on pairs of features. Here, for each pair of features, the decision tree learns decision boundaries that are built combinations of uncomplicated decision rules inferred from the training set. As seen Figure 6, decision tree models yield simple rules, but they are non-linear. In this regard, they are effectively used on non-linear datasets. Furthermore, implementing and deploying decision tree models is easy in terms of programming.

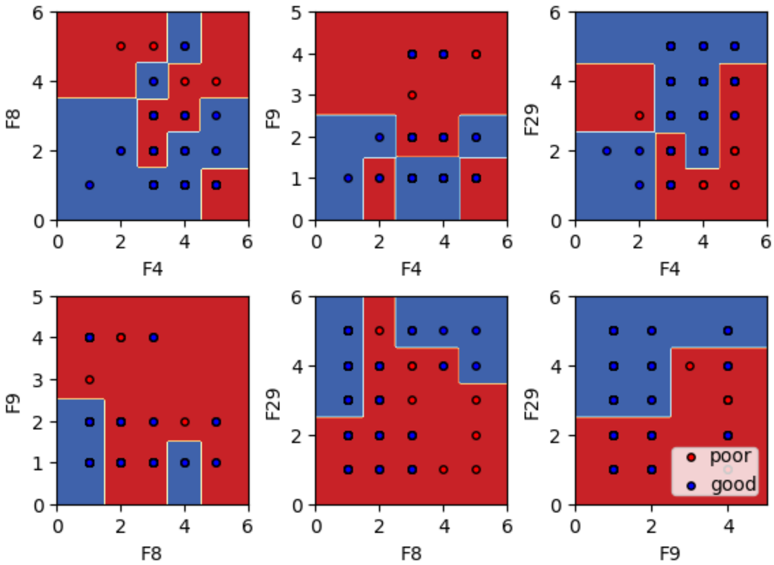

Figure 6. The decision surfaces of the CART classifier trained on pairs of features.

To sum up, throughout the model creation process, it has been observed that the performance of the models built without the feature selection process is low. To this end, the feature selection algorithms (i.e., Univariate feature selection, Tree-based feature selection, and L1-based feature selection) have been applied to the dataset. The performance of Univariate and Tree-based feature selection methods was not as good as the L1-based feature selection method. The L1-based feature selection method delivers good performance because it adopts the wrapper approach. The features that yield the best performance by the L1-based feature selection method have been selected. These features are: F4, F8, F9, and F29. To put it more explicitly, these are scholarship type, total salary, transportation to the university, and cumulative grade point average in the last semester. Subsequently, machine learning algorithms have been applied to the new dataset with 4 features. CART classifier delivered more performance compared to the other algorithms on this new dataset. When the tree-based model has been examined, the feature F29 has been placed at the root node. In other words, the Gini index value calculated for the feature F29 is lower than the other features. The smaller the Gini index the lower the uncertainty because the Gini index value is used to measure the impurity or purity of features. Accordingly, the cumulative grade point average in the last semester is the most important feature to predict the students' performance. After feature F29 at the root node, the purity of feature F4 is higher compared to features F8 and F9 for the left branch of the tree. Subsequent to feature F29 at the root node, the purity of feature F9 is higher compared to features F4 and F8 for the right branch of the tree. Finally, as traversing root to down, note that the purity values of these features will change.

DISCUSSION AND CONCLUSION

When the other works are explored to glimpse the importance of the attributes obtained across the feature selection process, Aslanargun et al. have reached the conclusion that the academic success of children whose families do not have economic troubles is high because they can easily attain adequate opportunity and support and their self-confidence and self-esteem are high (Aslanargun et al., 2016). In this respect, their research supports F4, F8, and F9 are proper to model the students' performance. In addition to Aslanargun's work, Sarıer's study has also shown that one of the variables that positively affect students' mathematics success is socio-economic status in TIMSS applications (Sarıer, 2020). Namely, these two papers overlap with the results obtained in this study. In the study (Şahin & Demirtaş, 2014) conducted on the academic success of foreign students, it was concluded that students who use public transport have difficulty in paying transport fees and also that the academic success of these students is low. As a result, the tie between students' academic performance and the ways they commute to school in their study overlaps with the results of this study. In the study (Akdamar & Kızılkaya, 2022) conducted on university students' academic procrastination tendencies, it was observed that students with low-grade point averages showed academic procrastination tendencies such as not completing homework on time and studying late for exams. This explains why students whose academic grade point average was low in the last semester will be able to have low-grade points in the next semester.

Numerous studies on forecasting students' performance have been realized. However, there exist some distinctions between this study and similar works. For instance, while studies done by Pallathadka, Salah Hashim, and Yılmaz concentrate on predicting students' final grades, this work focuses on students' general performance.

Therefore, the number of categorical values of the class has been reduced from to two: ‘fail’ and ‘pass’. Besides, in this paper, while the highest-performance model is transparent-based (e.g., a CART model), the performance of black box models in the studies done by Ismail, Pallathadka, Salah Hashim, and Yılmaz is higher. Transparent models can be understood by human experts and a new knowledge extraction is achievable in such models. However, understanding black box models (e.g., SVM, NB, Radial-based Neural Networks) is challenging and can even be incomprehensible for human experts. Further, Sa et al. have utilized solely tree-based models in their investigation. However, the performance of these models is rather low. More addition, Shanmugarajeshwari and Lawrance have employed a C5 model in their work. However, the dataset from which they benefit is small in comparison to the dataset employed in this paper. Considering the works that are similar to the outcomes of this study, the features such as scholarship type, total salary, transportation to the university, and cumulative grade point average in the last semester affect students’ performance more in comparison with the other features. Moreover, it is possible to model Students' performance by using only these four features. Besides, it is easy for human experts to comprehend the tree-based model, as well. This leads to obtaining new knowledge. All in all, such models formed by AI systems will hereafter be a pioneer in increasing students' performance.

This study was conducted on a certain data set. Different results can be obtained by applying it to different data sets. Additionally, different results can be obtained by applying different algorithms and methods to data sets. Naturally, the performance of models will be decisive in this situation. In a nutshell, training data, feature selection methods, machine learning classifiers, and data transformation methods all determine the final model. Therefore, many remarkable new investigations on EDM will be able to be carried out in the near future with the development of new approaches and algorithms for AI.

Statements of Publication Ethics

All authors declare that they obey the principles of publication ethics.

Conflict of Interest

This study has no conflict of interest.

REFERENCES

Acar, E. (2022). Comparison of the Performances of OECD Countries in the Perspective of Socio-Economic Global Indices: CRITIC-Based Cocoso Method. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, *73*, 256–277. https://doi.org/10.51290/dpusbe.1122650

Akdamar, E., & Kızılkaya, Y. M. (2022). Üniversite Öğrencilerinin Akademik Erteleme Eğilimleri ile Umutsuzluk Seviyeleri ve Akademik Başarıları Arasındaki İlişkinin İncelenmesi. *Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Dergisi*, *19*(1), 212–221. https://doi.org/10.33437/ksusbd.844605

Aslanargun, E., Bozkurt, S., & Sarıoğlu, S. (2016). Sosyo Ekonomik Değişkenlerin Öğrencilerin Akademik Başarısı Üzerine Etkileri. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, *9*(27/3), 201–234.

Aziz, Y., & Memon, K. H. (2023). Fast geometrical extraction of nearest neighbors from multi-dimensional data. *Pattern Recognition*, *136*, 109183. https://doi.org/10.1016/j.patcog.2022.109183

Baker, Ryan S. (2014). Educational Data Mining: An Advance for Intelligent Systems in Education. *IEEE Intelligent Systems*, *29*(3), 78–82. https://doi.org/10.1109/MIS.2014.42

Baker, Ryan Shaun, & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In *Learning Analytics* (pp. 61–75). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4

Baudat, G., & Anouar, F. (2000). Generalized Discriminant Analysis Using a Kernel Approach. *Neural Computation*, *12*(10), 2385–2404. https://doi.org/10.1162/089976600300014980

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., & Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*(1), 10312. https://doi.org/10.1038/srep10312

Beyer, K. S., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When Is “‘Nearest Neighbor’” Meaningful? *ICDT ’99 Proceedings of the 7th International Conference on Database Theory*, 217–235.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*(3), 483–519. https://doi.org/10.1007/s10115-012-0487-8

Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees* (1st ed.). Chapman and Hall/CRC.

Burges, C. C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel based Learning Methods*. Cambridge University Press.

Ghosh, D., & Cabrera, J. (2022). Enriched Random Forest for High Dimensional Genomic Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(5), 2817–2828. https://doi.org/10.1109/TCBB.2021.3089417

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann.

Hechenbichler, K., & Schliep, K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. *Collaborative Research Center 386*, *399*. https://doi.org/10.5282/ubm/epub.1769

Ismail, L., Materwala, H., & Hennebelle, A. (2021). Comparative Analysis of Machine Learning Models for Students’ Performance Prediction. In *Advances in Intelligent Systems and Computing* (pp. 149–160). https://doi.org/10.1007/978-3-030-71782-7_14

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kazak, E. (2021). Farklı Sosyo Ekonomik Çevrelerde Bulunan Okulların Etkililiğine İlişkin Öğretmenlerin Görüşleri. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *21*(1), 139–161. https://doi.org/10.17240/aibuefd.2021.21.60703-829153

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Lee, N., & Kim, J.-M. (2010). Conversion of categorical variables into numerical variables via Bayesian network classifiers for binary classifications. *Computational Statistics & Data Analysis*, *54*(5), 1247–1265. https://doi.org/10.1016/j.csda.2009.11.003

Lenat, D. B., & Feigenbaum, E. A. (1991). On the thresholds of knowledge. *Artificial Intelligence*, *47*(1–3), 185–250. https://doi.org/10.1016/0004-3702(91)90055-O

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on Platt’s probabilistic outputs for support vector machines. *Machine Learning*, *68*(3), 267–276. https://doi.org/10.1007/s10994-007-5018-6

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*(2). https://doi.org/10.1214/13-AOS1175

Manning, C. D., & Raghavan, P. (2009). An Introduction to Information Retrieval. In *Online* (p. 1). https://doi.org/10.1109/LPT.2009.2020494

Nisbet, R., Miner, G., & Yale, K. (2018). Data Understanding and Preparation. In *Handbook of Statistical Analysis and Data Mining Applications* (pp. 55–82). Elsevier. https://doi.org/10.1016/B978-0-12-416632-5.00004-9

Özdemir, A., Saylam, R., & Bilen, B. B. (2018). Eğitim Sisteminde Veri Madenciliği Uygulamaları Ve Farkındalık Üzerine Bir Durum Çalışması. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *22*(Özel Sayı 2), 2159–2172.

Özkan, Ö. (2015). Veri Madenciliği Kavramı ve Eğitimde Veri Madenciliği Uygulamaları. *Uluslararası Eğitim Bilimleri Dergisi*, *5*, 262–272.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, *80*, 3782–3785. https://doi.org/10.1016/j.matpr.2021.07.382

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall.

Sa, C. L., Abang Ibrahim, D. H. bt., Dahliana Hossain, E., & bin Hossin, M. (2014). Student performance analysis system (SPAS). *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 1–6. https://doi.org/10.1109/ICT4M.2014.7020662

Şahin, M., & Demirtaş, H. (2014). Üniversitelerde Yabancı Uyruklu Öğrencilerin Akademik Başarı Düzeyleri, Yaşadıkları Sorunlar ve Çözüm Önerileri. *Milli Eğitim Dergisi*, *44*(204), 88–113.

Salah Hashim, A., Akeel Awadh, W., & Khalaf Hamoud, A. (2020). Student Performance Prediction Model based on Supervised Machine Learning Algorithms. *IOP Conference Series: Materials Science and Engineering*, *928*(3), 032019. https://doi.org/10.1088/1757-899X/928/3/032019

Sarıer, Y. (2020). TIMSS Uygulamalarında Türkiye'nin Performansı ve Akademik Başarıyı Yordayan Değişkenler. *Temel Eğitim*, *2*(2), 6–27.

Shanmugarajeshwari, V., & Lawrance, R. (2016). Analysis of students' performance evaluation using classification techniques. *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, 1–7. https://doi.org/10.1109/ICCTIDE.2016.7725375

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774–780.

Yıldırım, H. İ. (2020). The Effect of Using Out-of-School Learning Environments in Science Teaching on Motivation for Learning Science. *Participatory Educational Research*, *7*(1), 143–161. https://doi.org/10.17275/per.20.9.7.1

Yılmaz, N., & Sekeroglu, B. (2020). Student Performance Classification Using Artificial Intelligence Techniques. In R. A. Aliev, J. Kacprzyk, W. Pedrycz, M. Jamshidi, M. B. Babanli, & F. M. Sadikoglu (Eds.), *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019* (pp. 596–603). Springer, Cham. https://doi.org/10.1007/978-3-030-35249-3_76

Yüksel, M. (2022). PISA 2018 Araştırma Sonuçlarına Göre Ülkelerin Bileşik PISA Performans Sıralaması. *Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 788–821. https://doi.org/10.21666/muefd.1093574