IJETS@R
(ISSN: 2587-0238)

Bezek Güre, Ö. (2023). Investigating the Performance of Feature Selection Methods in Classifying Student Success, *International Journal of Education Technology and Scientific Researches, 8(24), 2695-2728.*

DOI: http://dx.doi.org/10.35826/ijetsar.668

Article Type (Makale Türü): Research Article

INVESTIGATING THE PERFORMANCE OF FEATURE SELECTION METHODS IN CLASSIFYING STUDENT SUCCESS

Özlem BEZEK GÜRE
Assist. Professor, Batman University, Batman, Türkiye, ozlem.bezekgure@batman.edu.tr
ORCID: 0000-0002-5272-4639

Received: 13.05.2023 *Accepted: 19.09.2023* *Published: 01.10.2023*

ABSTRACT

The present study investigates the optimization of machine learning algorithms, specifically the Naïve Bayes classifier, in the context of Educational Data Mining (EDM). The primary objective is to scrutinize the impact of various feature selection algorithms on the performance of the model. Six feature selection methods—Information Gain, Gain Ratio, Symmetric Uncertainty Coefficient, Relief-F, Correlation-Based Feature Selection, and One R measure—are employed for an exhaustive comparative analysis. The research utilizes the "Higher Education Students Performance Evaluation" dataset available on the UCI Machine Learning Repository. This dataset is particularly robust, comprising 145 samples and 33 features, out of which 30 have been meticulously chosen for this study. The criteria for feature selection were based on their presumed relevance and potential impact on academic performance. Upon implementing the Naïve Bayes algorithm, the study discerns that the Gain Ratio method emerges as the most proficient, boasting an accuracy rate of 60%. Interestingly, aside from the Correlation-Based Feature Selection, the semester grade point average stands out as the most significant feature affecting student success rates. According to the Gain Ratio method, additional influential variables, listed in descending order of importance, include gender, the impact of projects/activities on academic success, expected grade point average upon graduation, weekly study hours, type of scholarship received, frequency of reading non-academic literature, mother's educational level, and participation in departmental seminars/conferences as well as class attendance. The research affirms the overall effectiveness of feature selection methods, with the exception of the One R method, in enhancing the predictive accuracy of the Naïve Bayes algorithm. These findings not only validate the utility of feature selection in EDM but also provide invaluable insights for researchers and educators interested in advancing the methodologies in the field of Educational Data Mining.

Keywords: Feature selection, educational data mining, naive bayes, higher education.

INTRODUCTION

Education serves as a critical parameter reflecting the level of advancement in societies. Given this significance, investigating the factors affecting the academic success of students across various educational levels, from primary to higher education, is of utmost importance. In order to facilitate academic planning, implement solutions, and develop strategies, a multitude of studies have been undertaken recently within this field (Akçöltekin, Engin & Şevgin, 2017). Researchers are increasingly utilizing Educational Data Mining (EDM) techniques to overcome the limitations of traditional statistical methods (Baker, 2010; Fernandes et al., 2019; Şevgin & Önen, 2022; Güre, 2023). EDM is a discipline that enables the analysis of large educational datasets, integrating diverse fields such as database management, data warehousing, statistics, and machine learning (Anuradha & Velmurugan, 2016). The applicability of EDM ranges from predicting school dropout rates, establishing the relationship between university entrance exam results and academic performance, to generating highly accurate student performance models (Ramaswami & Bhaskaran, 2009; Anuradha & Velmurugan, 2016). These techniques assist decision-makers in making rational choices, thereby contributing to the enhancement of educational quality. Like other sectors, the field of education generates large and complex datasets. The elimination of irrelevant or unnecessary variables from these datasets is a critical step (Budak, 2018). Feature selection methodologies optimize the performance of prediction models by eliminating such redundant variables (Punlumjeak & Rachburee, 2015). These methodologies play a significant role in enhancing prediction accuracy and laying the foundation for educational strategies (Zaffar et al., 2020).

The present study aims to examine various feature selection methods, including Information Gain (IG), Gain Ratio (GR), Correlation Based (CB), Symmetric Uncertainty (SU), Relief-F, and One R measure, in determining factors affecting the academic success of university students. To compare the efficacy of these methods, the Naïve Bayes algorithm has been employed. Moreover, it is aimed to make predictions with minimum error and high accuracy, as well as to compare the performances of feature selection methods.The research intended to address the following questions in this direction:

1. What are the factors affecting students' success according to feature selection methods?
2. Do the prediction performance of feature selection methods differ?
3. Does the use of feature selection methods affect the performance of the Naive Bayes method?

Related Studies

Thoroughly examining existing literature makes it is clear that data mining techniques are widely used to forecast student performance. A limited number of studies have utilized the "Higher Education Students Performance Evaluation dataset," which is the focus of the current investigation. For instance, in a study conducted by Hengpraprohm, Hengpraprohm, and Sudjitjoon (2022), algorithms like K-Nearest Neighbor (KNN), Random Forest (RF), Artificial Neural Network (ANN), and Linear Regression were employed, incorporating feature selection methods such as IG, GR, CB, and Chi-Square (CS). Similarly, a study by Jabardi (2022) utilized algorithms

like RF, AdaBoost, Decision Tree (DT), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). Additionally, Phatai and Luangrungruang (2023) employed ANN and metaheuristic algorithms to classify student performance.

In a second category, other studies that have applied feature selection methods were also considered. A study by Göker, Bülbül, and Irmak (2013) utilized NB, J48, Bayes Net, and Radial Basis Function (RBF) neural network algorithms and applied feature selection methods like IG, GR, SU, One R, and CS. Punlumjeak and Rachburee (2015) evaluated feature selection methods such as IG, minimum redundancy, and maximum relevance in the context of NB, DT, KNN, and ANN algorithms, along with genetic algorithms and Support Vector Machine (SVM). A study by Estrera et al., (2017) employed DT, NB, and KNN algorithms, applying CS, IG, and GR as feature selection methods. Zaffar et al., (2018) enriched their research by incorporating additional methods like principal component analysis and the Relief method.

Considering this extensive literature, it becomes clear that there are performance disparities between various algorithms and feature selection methods. For instance, a novel feature selection method combining Chi-square and Mutual Information, proposed by Sokkhey and Okazaki (2020), demonstrated superior performance. Again, Mythili and Shanavas (2014) applied IG and GR feature selection methods in their study in which they used J48, RF, MLP, IBI, KNN and Decision Table methods. Rahman Setiawan and Permanasari (2017) evaluated the performance of Naive Bayes (NB), Decision Tree (DT), and Artificial Neural Networks (ANN) algorithms using wrapper and IG feature selection methods. Moreover, a study by Makhtar et al., (2017) applied the Best First algorithm as a feature selection method while utilizing algorithms like NB, Random Tree, KNN, Multi-class classifier, and Conjunctive Rule. Velmurugan and Anuradha (2016) for this purpose; In their studies, they used J48, NB, Bayes Net, IBk, OneR, and JRip methods; They used Best First Search, CB, Wrapper, CfsSubset, CS, IG and Relief methods. In addition; Ramaswami and Rathinasabapathy (2012) used other methods, except the One R method, among the feature selection methods used in the current study to predict students' success. Yahdin et al., (2021) used NB and KNN methods to measure the performance of the Relief feature selection method. In their study, Anuradha and Velmurugan (2016) tried to determine the effect of CS and GR feature selection methods using the NB method.

METHOD

Research Design

This research was conducted using a correlational design. A correlational design examines the relationship between multiple variables (Karasar, 2006; Güre, Kayri & Erdoğan, 2020). Using this approach provides an effective framework to assess the complex connections between the variables of interest in the current study.

Employed dataset

The current study employs the "Higher Education Students Performance Evaluation dataset" available in the UCI database (Yılmaz & Sekeroglu, 2019). The data file was obtained from

https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation. The dataset encompasses information gathered from 145 (87 boys (60%) and 58 girls (40%)) students enrolled in faculties of Education and Engineering. It includes variables related to familial background, personal information, and study habits. Although the dataset contains 33 variables, the present study has utilized only 30 of these variables.

Data Analysis

The current study utilized the Weka data mining software to conduct the data analyses. Weka is a free, open-source platform developed at the University of Waikato that contains a collection of machine learning algorithms and data preprocessing tools (Hall et al., 2009). For this research, Weka enabled robust examination of the relationships within the dataset through its tools for statistical analysis and machine learning techniques.

Feature Selection Methods

Feature selection methodologies hold a critical stance in the realm of data mining, particularly when dealing with high-dimensional data sets. Such methods are ubiquitously employed in machine learning to select pertinent subsets from a given feature set, thereby eliminating features that are either irrelevant or superfluous. Features can also be referred to as attributes or variables. Feature selection aims to identify an optimal subset capable of representing the original data set (Budak, 2018). It serves as a data preprocessing technique (Phyu and Oo, 2016), consequently simplifying the data set by reducing the number of features, thereby facilitating a more efficient analysis. These methods have found extensive utility across diverse disciplines (Guan et al., 2014).

In recent years, an intriguing approach has been proposed that integrates feature selection methods with ensemble learning techniques. The overarching idea is to generate a multitude of feature selectors and amalgamate their outputs. Feature selection serves as one of the techniques employed for dimensionality reduction (Remeseiro & Bolon-Canedo, 2019). By discarding unrelated and superfluous features, these methods not only simplify the analysis but also enhance the quality of the data (Guan et al., 2014).

The attributes of feature selection methods are manifold, encompassing the elimination of noisy and irrelevant data, reduction in the dimensionality of the feature set, augmentation in algorithmic speed, enhancement of data quality, and improvement in the performance of the derived model (Ladha & Deepa, 2011). Given these capabilities, a plethora of methods related to feature selection has been developed in recent years (Solorio-Fernández, Carrasco-Ochoa & Martínez-Trinidad, 2020). These methods are applied across different fields for classification, clustering, and regression analyses (Jović, Brkić & Bogunović, 2015).

Feature selection methods can be broadly categorized into supervised, semi-supervised, and unsupervised methods. Supervised methods necessitate a labeled dataset to identify and select pertinent features. The label assigned to each object in the dataset could be a categorical, ordinal, or real value. Semi-supervised methods require labeling only for a subset of objects, while unsupervised methods do not necessitate a labeled dataset (Solorio-Fernández, Carrasco-Ochoa & Martínez-Trinidad, 2020).

In the current study, specific feature selection methods such as Information Gain, Gain Ratio, and Symmetric Uncertainty Coefficient, along with Relief-F, Correlation-Based Feature Selection Method, and One R measure, are discussed and evaluated.

Information Gain Attribute Evaluation

The entropy measure serves as an indicator of sample homogeneity in information t(Abe & Kudo, 2005). Ranging from 0 to 1, an entropy of zero denotes homogeneity among samples, while a value of one signifies a lack of such homogeneity (Sastry et al., 2010). Information Gain (IG) is an entropy-based feature selection technique widely deployed in machine learning. The method is instrumental in identifying attributes that are most informative regarding the dependent class variable (Win & Kham, 2019). IG quantifies the reduction in entropy in attribute Y facilitated by the utilization of attribute X (Novaković, Strbac, & Bulatović, 2011). Notably, the method is classifier-agnostic, enabling its application across various classification models (Win & Kham, 2019). However, IG is susceptible to bias towards features with higher numerical characteristics but lesser information, considered a limitation of the method (Tripathi & Trivedi, 2016).

The mathematical formulations for Information Gain are as follows:

$$H(Y) = -\sum_{y\epsilon Y} p(y) log_2 \left(p(y)\right) \quad (1)$$

$$H(Y\backslash X) = -\sum_{x\epsilon X} p(x) \sum_{y\epsilon Y} p(y\backslash x) log_2 (p(y\backslash x)) \quad (2)$$

$$Information\ Gain = H(Y) - H(X\backslash Y) \quad (3)$$

Here, H(X) represents the entropy of X, while H(X\Y) indicates the entropy of X after Y is observed (Win & Kham, 2019).

Gain Ratio Attribute Evaluation

The Gain Ratio is a non-symmetrical measure conceived to counter the inherent bias in Information Gain (Novaković, Strbac, & Bulatović, 2011; Tripathi & Trivedi, 2016; Al Janabi & Kadhim, 2018). The Gain Ratio is defined as:

$$Gain\ ratio = \frac{\text{Information Gain}}{\text{H(X)}} \quad (4)$$

The metric often outperforms basic Information Gain due to its complexity and classification accuracy (Quinlan, 1988; Rokach & Maimon, 2005).

Symmetrical Uncertainty Attribute Evaluation

The Symmetrical Uncertainty Coefficient is a metric designed to mitigate the inherent bias found in Information Gain. It is normalized to a [0,1] range, much like entropy. A coefficient of 1 signifies that information in attribute

X can fully predict that of attribute Y, whereas a value of 0 indicates no correlation between the two attributes (Hall, 1999; Al Janabi & Kadhim, 2018).

The mathematical representation of the Symmetrical Uncertainty Coefficient is as follows:

$$Symmetrical\ Unceartinity\ Coefficient = 2\frac{Information\ Gain}{H(X)+H(Y)} \quad (5)$$

Relief-F Attribute Evaluation

The Relief method has been developed as a distance-based metric and is particularly effective in the realm of two-class problems (Win & Khan, 2019). This algorithm evaluates the importance of a given feature by scrutinizing the closest samples that fall within both the same and disparate classes, employing a form of iterative sampling. Based on this evaluation, the method assigns a weight to each feature, which is directly correlated with the feature's ability to delineate between classes. While the Relief method may not identify the most minimal subset of features, it does excel in selecting features that bear statistical relevance. Subsequently, these features are ranked, and those that achieve weights meeting or exceeding a user-defined threshold are selected for further analysis (Kira & Rendell, 1992; Novakovıć, Strbac & Bulatovıć, 2011). Notably, Relief algorithms are frequently employed as a preliminary feature subset selection technique prior to the actual learning phase of the model (Kira & Rendell, 1992).

In the context of heuristic methods, feature independence is often emphasized to assess the quality of features. However, such approaches are generally not well-suited for tackling problems where features interact in complex ways. The Relief method distinguishes itself in this regard, as it operates without any such assumptions of feature independence. This unique characteristic allows the Relief method to accurately assess the quality of features, even in situations where there exists a strong interdependency among the features themselves (Marko & Igor, 2003).

Correlation Based Feature Selection

This heuristic-based method focuses on the ranking of feature subsets based on their correlation with the class variables. It prioritizes features that are highly correlated with the class yet minimally correlated with each other, thereby maximizing classification performance (Hall, 1999; Al Janabi & Kadhim, 2018).

The mathematical representation for this method is:

$$\boldsymbol{M_s = \frac{k\bar{r}_{c_i}}{\sqrt{k+k(k-1)\bar{r}_{ii}}}} \quad (6)$$

One R Attributed

The One R method is a straightforward technique initially proposed by Holte. It focuses on feature selection based on error rates, specifically by generating a rule for each feature in the training set and then selecting the one with the lowest error rate (Novaković, Strbac & Bulatović, 2011). This method is versatile in handling various

types of data, including instances with missing values, by treating such absent values as valid for computational purposes (Al Janabi & Kadhim, 2018). It is particularly effective in scenarios involving multiple properties and diverse classes, iteratively selecting a single best feature and formulating rules based solely on that chosen attribute (Tripathi & Trivedi, 2016).

Naive Bayes

The Naive Bayes (NB) method employs Bayes' theorem for classification purposes. It offers a simple, rapid, and effective probability-based classification approach conducive to the expeditious creation of accurate machine learning models (Rish, 2001; Jabardi, 2022; Wickramasinghe & Kalutarage, 2021). The method possesses distinct features such as computational efficiency, low variance, incremental learning, direct posterior probability estimation, and robustness to noisy and missing data (Sammut & Webb, 2017). The NB algorithm does not necessitate complex iterative parameter estimations, thereby simplifying its construction and interpretation (Wu et al., 2008). It has been particularly noted for its strong performance on large datasets (Nikam, 2015; Jabardi, 2022). The NB method calculates a conditional probability for each relationship and utilizes this probability to analyze the association between the independent and dependent variables (Muda et al., 2011).

Mathematically, the NB classifier operates based on Bayes' rule:

$$\mathrm{P}(\mathrm{y}\backslash\mathbf{x}) = (\mathrm{P}(\mathrm{y})\mathrm{P}(\mathbf{x}\backslash\mathrm{y}))/\mathrm{P}(\mathbf{x}) \quad (7)$$

Assuming conditional independence of the features given the class, this premise affords:

$$P(x\backslash y) = \prod_{i=1}^{n} P(x_i\backslash y) \quad (8)$$

Here, x_i represents the value of the ith feature in x, and n denotes the number of features.

$$P(\boldsymbol{x}) = \prod_{i=1}^{k} P(c_i)P(\boldsymbol{x}\backslash c_i) \quad (9)$$

Here, c_i denotes the ith class, and k signifies the number of classes. Upon normalization of the right-hand side of the equation, Equation (7) is obtained (Webb, Keog, & Miikkulainen, 2010).

FINDINGS

Evaluation metric

Accuracy, precision, recall and F1 score were used as performance criteria in the study. The equations about performance criteria are given below:

Table 1. Confusion Matrix

		Estimated Class		
		No	Yes	Total
Real Class	No	TN	FP	TN+FP
	Yes	FN	TP	FN+TP
	Total	TN+FN	FP+TP	TN+FN+FP+TP

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F1\ score = 2x\ \frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (13)$$

The most important 10 variables obtained by using feature selection methods are given in Table 2.

Table 2. Variables Considered Important According to Different Feature Selection Methods

IG	GR	SU	Relief F	Correlation based	One R
Cumulative grade point average in the last semester	Cumulative grade point average in the last semester	Cumulative grade point average in the last semester	Cumulative grade point average in the last semester	Gender	Cumulative grade point average in the last semester
Expected Cumulative grade point average in the graduation	Gender	Gender	Scholarship type	Cumulative grade point average in the last semester	Expected Cumulative grade point average in the graduation
Weekly study hours	Impact of your projects/activities on your success	Expected Cumulative grade point average in the graduation	Gender	Attendance to the seminars/conferences related to the department	Number of sisters/brothers (if available):
Mother's educational level	Expected grade point average in the graduation	Weekly study hours	Student Age	Attendance to classes	Weekly study hours:
Gender	Weekly study hours:	Impact of your projects/activities on your success:	Mother's educational level	Total salary if available	Listening in classes
Scholarship type	Scholarship type	Mother's educational level	Accommodation type	Impact of your projects/activities on your success	Regular artistic or sports activity: (
Father’s occupation	Reading frequency (non-scientific books/journals)	Scholarship type	Weekly study hours:	Mother's educational level	Graduated high-school type
Reading frequency (non-scientific books/journals)	mother's educational level	Reading frequency (non-scientific books/journals)	Graduated high-school type	Flip-classroom	Transportation to the university
Impact of your projects/activities on your success	Attendance to the seminars/conferences related to the department	Mother's occupation	Reading frequency (non-scientific books/journals)	Scholarship type	Reading frequency (scientific books/journals)
Father’s educational level	Attendance to classes	Total salary if available	Flip-classroom	Expected Cumulative grade point average in the graduation	Attendance to the seminars/conferences related to the department

Subsequently, using the selected features, analyses were performed with the Naïve Bayes method in the Weka program. The analysis results are shown below.

Table 3. Performance of Feature Selection Methods Utilizing the Naïve Bayes Method

Method	Feature selection methods	Accuracy	Precision	Recall	F1-score
Naïve Bayes	All Features	49.65	0.48	0.49	0.48
	IG	56.55	0.561	0.566	0.562
	GR	60.00	0.595	0.600	0.597
	SU	56.55	0.552	0.566	0.554
	Relief-F	58.62	0.586	0.586	0.580
	CB	56.55	0.548	0.566	0.553
	One R	49.65	0.442	0.497	0.465

As can be discerned from Table 3, Gain Ratio measures emerge as the most efficacious feature selection methodology for pinpointing factors that influence student performance when utilizing the Naïve Bayes algorithm. It can be stated that, with the exception of the One R method, all other techniques have contributed to enhancing the performance of the Naïve Bayes approach.

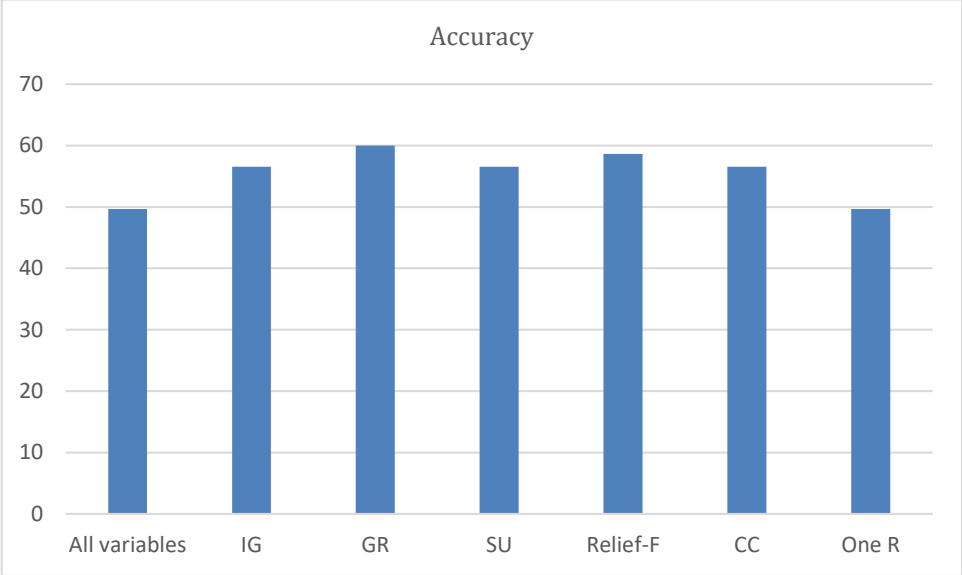


Figure 1. Distribution of feature selection methods according to accuracy

Upon examining Figure 1, it is observed that the GR method performs better in terms of accuracy. The One R method has a similar accuracy rate to the results of the analysis where all features are used. It is seen that other methods enhance the classification performance of the Naive Bayes method.

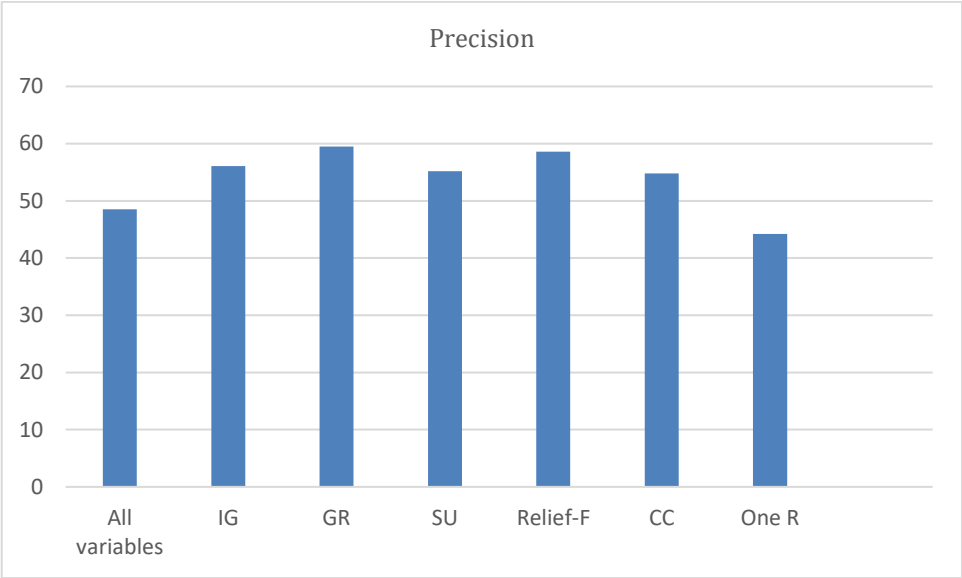


Figure 2. Distribution of feature selection methods according to precision

Figure 2 illustrates that the GR method outperforms others in terms of precision, whereas the One R method registers the lowest rate.

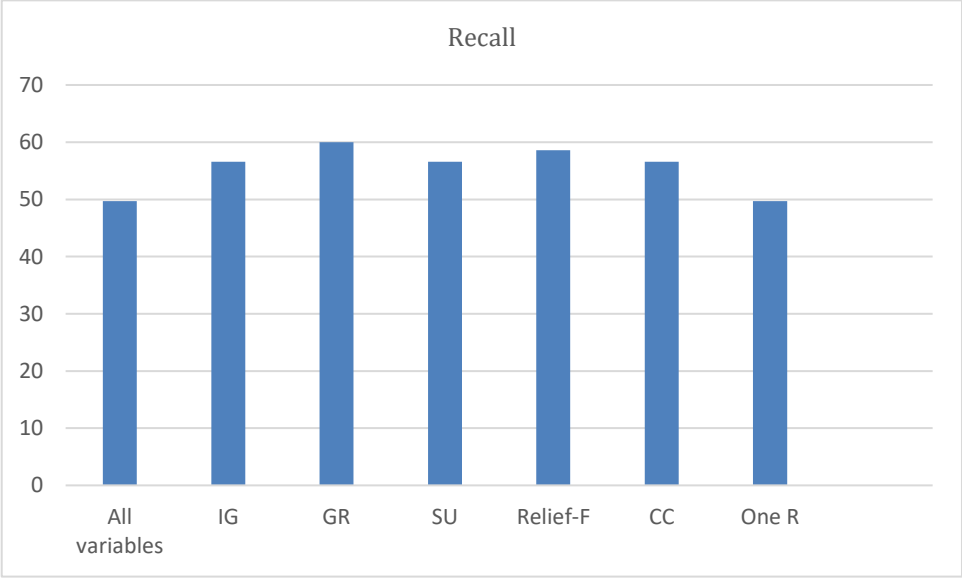


Figure 3. Distribution of Feature Selection Methods According to Recall

It is observed in Figure 3 that the GR method has a higher rate in terms of recall.

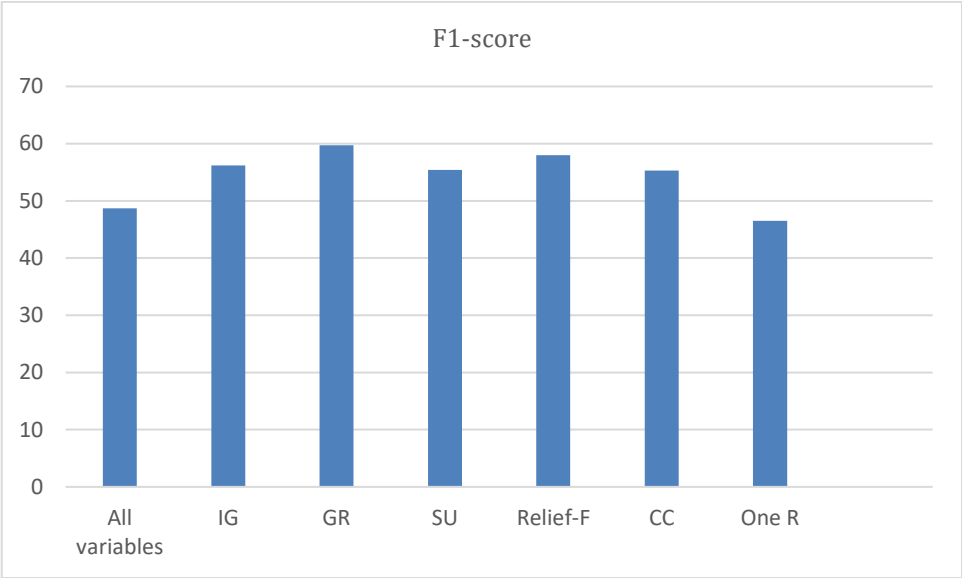


Figure 4. Distribution of Feature Selection Methods According to F1-score

Upon examining Figure 4, it is evident that the Gain Ratio (GR) method achieves the highest rate in terms of F-measure, while the One R method registers the lowest rate.

CONCLUSION and DISCUSSION

The aim of this study is to identify the factors affecting the academic performance of university students by employing feature selection methods including IG, GR, SU, CB, Relief-F, and One R Measure. To compare the efficacy of these feature selection methods, the Naïve Bayes classifier has been applied. According to the results obtained via the NB classifier, Gain Ratio have emerged as the most effective feature selection methods, both with an accuracy rate of 60%. The analysis further reveals that, with the exception of the CB method, the most significant factor influencing student performance across all other methods is the student's grade point average for the final semester. According to the GR method, the subsequent variables of importance, in descending order, are: gender, impact of your projects/activities on your success expected grade point average in the graduation, weekly study hours, scholarship type, frequency of reading non-academic books. mother's educational level, attendance to the seminars/conferences related to the department, and attendance to classes.

In research closely related to our study, Hengpraprohm, Hengpraprohm, and Sudjitjoon (2022) employed KNN, RF, ANN, and Linear Regression methods, incorporating IG, GR, CB, and CS for feature selection. They identified IG and GR as the most efficacious methods. Specifically, when utilizing the GR method, the accuracy rates for KNN and RF were observed to be 48.53% and 76.07%, respectively. In contrast, when employing the IG method, ANN yielded an accuracy of 40.0%, and Linear Regression resulted in 54.37%. Additionally, their study discerned several pivotal variables, including the student's grade point average in the final semester, academic achievement expectations, frequency of reading non-academic books, and the presence of a divorced or deceased parent.

Moreover, Jabardi (2022) examined the performance of RF, AdaBoost, DT, NB, and MLP for the same objective. The observed accuracy rates were 78.62%, 84.82%, 74.48%, 66.20%, and 73.10%, respectively. Notably, they underscored the superior performance of the AdaBoost method in comparison to the other techniques.

Punlumjeak and Rachburee (2015) explored an array of methods—NB, DT, KNN, and ANN—in conjunction with feature selection techniques such as Genetic Algorithms, SVM, IG, Minimum Redundancy, and Maximum Relevance. Their study found that the most effective performance was achieved when using the Minimum Redundancy and Maximum Relevance feature selection method with KNN. Specifically, when employing the Naïve Bayes method, the highest accuracy rate of 83.87% was achieved under SVM feature selection.

Similarly, Göker, Bülbül, and Irmak (2013) used a variety of methods like NB, J48, Bayes Net, and RBF while applying feature selection methods including IG, GR, SU, One R, and CS. They identified SU as the best feature selection method with an average accuracy of 83.87%. Additionally, Rahman Setiawan and Permanasari (2017) employed NB, DT, and ANN methods in their study, utilizing Wrapper and IG for feature selection. The accuracy rate for NB using the Wrapper feature selection method was determined to be 75.41%, highlighting that feature selection techniques enhance the performance of the Naïve Bayes method.

Furthermore, Velmurugan and Anuradha (2016) conducted a study using multiple algorithms such as J48, NB, Bayes Net, IBk, OneR, and JRip, while incorporating feature selection methods like Best First Search, Wrapper, CfsSubset, CS, IG, and Relief. Their results showed that the highest performance was obtained using the CfsSubset feature selection method, with Naïve Bayes registering an accuracy of 98.56%. On the other hand, Anuradha and Velmurugan (2016) focused on evaluating the effectiveness of the CB and GR feature selection methods using the NB algorithm and found that CB outperformed GR, achieving an accuracy rate of 84%. Their findings corroborate that feature selection methods enhance classification performance. Similar outcomes were also reported in studies by Makhtar et al. (2017), Priyasadie and Isa (2021), and Yahdin et al. (2021).

In their seminal work, Ramaswami and Rathinasabapathy (2012) employed various feature selection methods, with the exception of the One R method, that are also utilized in the current study, aimed at predicting student performance. Their findings underscore not only the enhancement of predictive accuracy through the use of feature selection methods but also a reduction in computational time required by the algorithms.

Our analysis indicates that, with the exception of the One R method, the implementation of feature selection methods effectively enhances the classification performance of the Naïve Bayes (NB) algorithm. This observation aligns well with the results from existing literature in the field, further substantiating the contributions of our study.

SUGGESTIONS

For forthcoming scholarly inquiries, it is advised to investigate the application of feature selection methods to augment the effectiveness of data mining techniques. Additionally, the performance of various data mining

methodologies could be assessed through empirical evaluations. Further investigations could also be conducted on different datasets to broaden the scope and applicability of the findings. The current study utilized the Weka software for evaluating feature selection methods; future research may benefit from exploring alternative software solutions for a more comprehensive understanding.

ETHICAL TEXT

"In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. The responsibility belongs to the author (s) for any violations that may arise regarding the article. The dataset used in the present study was sourced from the publicly accessible UCI database; therefore, the research did not necessitate ethical committee approval."

Author(s) Contribution Rate: In this study, the contribution rate of the first author is 100%.

REFERENCES

Abe, N. & Kudo, M., (2005, September 14-16,). "Entropy criterion for classifier-independent feature selection" [Oral presentation]. 9th International Conference, KES 2005, Melbourne, Australia.

Akcoltekin, A., Engin, A. O., & Sevgin, H. (2017). Attitudes of high school teachers to educational research using classification-tree method. Eurasian Journal of Educational Research, 17(68), 19-47. https://dergipark.org.tr/en/pub/ejer/issue/42457/511275

Al Janabi, K. B., & Kadhim, R. (2018). Data reduction techniques: a comparative study for attribute selection methods. International Journal of Advanced Computer Science and Technology, 8(1), 1-13.

Anuradha, C., & Velmurugan, T. (2016, January, 19-21). Feature selection techniques to analyse student acadamic performance using Naïve Bayes classifier [Oral presentation]. In The 3rd international conference on small & medium business. Hochiminh, Vietnam

Baker, R. S. J. D. (2010). Data mining for education. International encyclopedia of education. Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition) (pp.112-118.. Oxford, UK: Elsevier.

Bezek Güre, Ö. (2023). Investigation of ensemble methods in terms of statistics: TIMMS 2019 example. Neural Computing and Applications, 1-14. https://doi.org/10.1007/s00521-023-08969-0

Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi. 22, 21-31. https://doi.org/10.19113/sdufbed.01653

Estrera, P. J. M., Natan, P. E., Rivera, B. G. T., & Colarte, F. B. (2017). Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School Abstract. International Journal of Engineering and Technology, 3(5), 147-153. http://www.ijetjournal.org/

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of business research, 94, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

Göker, H., Bülbül, H. I., & Irmak, E. (2013, December, 04-07). The estimation of students' academic success by data mining methods. In 2013 12th International Conference on Machine Learning and Applications (Vol. 2, pp. 535-539). IEEE. Miami, FL, USA

Guan. D.. Yuan. W.. Lee. Y. K.. Najeebullah. K.& Rasel. M. K. (2014). A review of ensemble learning based feature selection. IETE Technical Review. 31(3). 190-198. https://doi.org/10.1080/02564602.2014.906859

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of Factors Effecting PISA 2015 Mathematics Literacy via Educational Data Mining. Education & Science/Egitim ve Bilim, 45(202). http://dx.doi.org/10.15390/EB.2020.8477

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD explorations newsletter, 11(1), 10-18. https://dl.acm.org/doi/abs/10.1145/1656274.1656278

Hall. M. (1999). Correlation-based Feature Selection for Machine Learning. [PhD Thesis]. The University of Waikato.

Hengpraprohm, K., Hengpraprohm, S., & Sudjitjoon, W. (2022). A Study of Factors Affecting Learning Efficiency on Higher Education Student Performance Evaluation Dataset Using Feature Selection Techniques. Information Technology Journal, 18(2), 34-43. https://ph01.tci-thaijo.org/index.php/IT_Journal/article/view/251051

Jabardi, M. H. (2022). Machine learning techniques for assessing students'environments'impact factors on their academic performance. International Journal of Advanced Research in Computer Science, 13(2). http://dx.doi.org/10.26483/ijarcs.v13i2.6813

Jović, A., Brkić, K., & Bogunović, N. (2015, May, 25-29). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). IEEE. Opatija, Croatia

Karasar, N. (2009). Bilimsel araştırma yöntemi (23. bs.). Ankara: Nobel Yayınlar

Kira, K..& Rendell, L.A. (1992). A practical approach to feature selection. In D. Sleeman. & P. Edwards (Eds.). Machine Learning: Proceedings of International Conference (ICML'92) (pp. 249–256). Morgan Kaufmann

Ladha. L. & Deepa. T. (2011). Feature Selection Methods And Algorithms. International Journal on Computer Science and Engineering. 3(5). 1787-1797

Makhtar, M., Nawang, H., & Wan Shamsuddin, s. N. (2017). Analysis on students performance using naïve bayes classifier. Journal of Theoretical & Applied Information Technology, 95(16). https://www.researchgate.net/profile/Hasnah-Nawang/publication/319955477_Analysis_on_students_performance_using_naIve_Bayes_classifier/links/60ebca8cb8c0d5588cee6bfa/Analysis-on-students-performance-using-naIve-Bayes-classifier.pdf

Marko. R.S., & Igor. K. (2003). "Theoretical and empirical analysis of relief and rreliefF". Machine Learning Journal. 53 23–69. https://doi.org/10.1023/A:1025667309714

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011, July, 12-13). Intrusion detection based on K-Means clustering and Naïve Bayes classification. In 2011 7th international conference on information technology in Asia (pp. 1-6). IEEE. Sarawak, Malaysia

Mythili, M. S., & Shanavas, A. M. (2014). An Analysis of students' performance using classification algorithms. IOSR Journal of Computer Engineering, 16(1), 63-69. https://www.researchgate.net/profile/Mohamed-Shanavas/publication/314445897_An_Analysis_of_students'_performance_using_classification_algorithms/links/58d5efff92851c44d461e5af/An-Analysis-of-students-performance-using-classification-algorithms.pdf

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, 8(1), 13-19. http://www.computerscijournal.org/?p=1592

Novakavic. J.. Strbac. P.. Bulatovic. D. (2011). Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms. Yugoslav Journal of Operations Research. 21(1). 119-135. http://www.yujor.fon.bg.ac.rs/index.php/yujor/article/download/364/255

Phatai, G., & Luangrungruang, T. (2023, March, 18-20). A Comparative Study of Hybrid Neural Network with Metaheuristics for Student Performance Classification. In 2023 11th International Conference on Information and Education Technology (ICIET) (pp. 448-452). IEEE. Fujisawa, Japan

Phyu, T. Z., & Oo, N. N. (2016). Performance comparison of feature selection methods. In MATEC web of conferences (Vol. 42, p. 06002). EDP Sciences. https://doi.org/10.1051/matecconf/20164206002

Priyasadie, N., & Sani, M. I. (2021). Educational Data Mining in Predicting Student Final Grades on Standardized Indonesia Data Pokok Pendidikan Data Set. International Journal of Advanced Computer Science and Applications, 12(12). https://doi.org/10.14569/IJACSA.2021.0121227

Punlumjeak, W., & Rachburee, N. (2015, October, 29-30). A comparative study of feature selection techniques for classify student performance. In 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 425-429). IEEE. Chiang Mai, Thailand

Quinlan. J.R.. (1988). Decision trees and multivalued attributes. J. Richards. ed.. Machine Intelligence. 11: 305-318. Oxford University Press

Rahman, L., Setiawan, N. A., & Permanasari, A. E. (2017, November, 01-02). Feature selection methods in improving accuracy of classifying students' academic performance. In 2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE) (pp. 267-271). IEEE. Yogyakarta, Indonesia

Ramaswami, M., & Rathinasabapathy, R. (2012). Student performance prediction. International Journal of Computational Intelligence and Informatics, 1(4), 231-235. https://www.academia.edu/download/34746728/IJCII_1-4-38_-_Ramasamy.pdf

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

Rokach. L.. Maimon. O.. (2005). Data Mining and Knowledge Discovery Handbook. Springer New York Dordrecht Heidelberg London.1285.

Sachin, R. B., & Vijay, M. S. (2012, January, 07-08). A survey and future vision of data mining in educational field. In 2012 second international conference on advanced computing & communication Technologies, IEEE, 96-100. Rohtak, India

Sastry PM. Krishnan R. Ram .B.V.S.. (2010). Classification and identification of teluguh and written characters extracted from palm leavesusing decision tree approach. ARPN Journal of Engineering and Applied Sciences. 5 (3): 22-32. https://www.researchgate.net/profile/Narahari-Sastry/publication/242591979_Classification_and_identification_of_Telugu_handwritten_characters_extracted_from_palm_leaves_using_decision_tree_approach/links/56152f4508aed47facefb7bd/Classification-and-identification-of-Telugu-handwritten-characters-extracted-from-palm-leaves-using-decision-tree-approach.pdf

Şevgin, H. & Önen, E. (2022). Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE- 2016 Case. Education & Science/Egitim ve Bilim, , 47(211). http://dx.doi.org/10.15390/EB.2022.10575

Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. International Journal of Advanced Computer Science and Applications, 11(8), 492-502. https://www.academia.edu/download/64408036/Paper_62-Study_on_Dominant_Factor_for_Academic_Performance.pdf

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. Artificial Intelligence Review, 53(2), 907-948. https://doi.org/10.1007/s10462-019-09682-y

Tripathi. A..& Trivedi. S. K. (2016, 24-24 October). Sentiment analyis of Indian movie review with various feature selection techniques. In 2016 IEEE international conference on advances in computer applications (ICACA) (pp. 181-185). IEEE. Coimbatore

UCI Machine Learning Repository: Higher Education Students Performance Evaluation Dataset Data Set. https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset#

Velmurugan, T., & Anuradha, C. (2016). Performance evaluation of feature selection algorithms in educational data mining. Performance Evaluation, 5(02). https://www.researchgate.net/profile/Velmurugan-Thambusamy/publication/311773948_Performance_Evaluation_of_Feature_Selection_Algorithms_in_Educational_Data_Mining/links/585a381108ae3852d256dfb0/Performance-Evaluation-of-Feature-Selection-Algorithms-in-Educational-Data-Mining.pdf

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714.

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293. https://doi.org/10.1007/s00500-020-05297-6

Win. T. Z..& Kham. N. S. M. (2019). Information gain measured feature selection to reduce high dimensional data. In Seventeenth International Conference on Computer Applications (ICCA 2019) (Vol. 68. No. 73. pp. 1-5). https://meral.edu.mm/record/3413/files/ICCA%202019%20Proceedings%20Book-pages-79-84.pdf

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14, 1-37. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/content/pdf/10.1007/s10115-007-0114-2.pdf&casa_token=Nx6o3BNiK_kAAAAA:6VTMnyDyNk9J_-wfg09CuJHptY6ERpasiIJfKIJ3weOKbWQNi5mWeJrC1_Yxcs0C3x6XMDx8OIi0b-i6sw

Yahdin, S., Desiani, A., Gofar, N., & Agustin, K. (2021). Application of the relief-f algorithm for feature selection in the prediction of the relevance education background with the graduate employment of the universitas sriwijaya. Computer Engineering and Applications Journal, 10(2), 71-80. https://doi.org/10.18495/comengapp.v10i2.369

Yılmaz, N., & Sekeroglu, B. (2019, August). Student performance classification using artificial intelligence techniques. In International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (pp. 596-603). Cham: Springer International Publishing.

Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. International Journal of Advanced Computer Science and Applications, 9(5). https://www.researchgate.net/profile/Maryam-Zaffar/publication/325574028_A_Study_of_Feature_Selection_Algorithms_for_Predicting_Students_Academic_Performance/links/5b28ac4ba6fdcca0f09c62fa/A-Study-of-Feature-Selection-Algorithms-for-Predicting-Students-Academic-Performance.pdf

Zaffar, M., Hashmani, M. A., Savita, K. S., Rizvi, S. S. H., & Rehman, M. (2020). Role of FCBF feature selection in educational data mining. Mehran University Research Journal Of Engineering & Technology, 39(4), 772-778. https://search.informit.org/doi/pdf/10.3316/informit.459135063399758

ÖĞRENCİLERİN BAŞARILARINI SINIFLANDIRMADA ÖZELLİK SEÇİM YÖNTEMLERİNİN PERFORMANSLARININ İNCELENMESİ

ÖZ

Bu çalışma, Eğitsel Veri Madenciliği (EDM) bağlamında makine öğrenme algoritmalarının, özellikle de Naïve Bayes sınıflandırıcısının optimizasyonunu araştırmaktadır. Birincil amaç, çeşitli özellik seçme algoritmalarının modelin performansı üzerindeki etkisini incelemektir. Bu çalışmada; özellik seçim yöntemlerinden Bilgi Kazancı, Kazanç Oranı, Simetrik Belirsizlik katsayısı, Relief-F, Korelasyon tabanlı ve One R özellik seçim yöntemi kullanılarak, yüksek öğrenimde öğrenim gören öğrencilerin başarılarını etkileyen faktörleri belirlemek amaçlanmaktadır. Özellik seçim yöntemlerinin etkisini karşılaştırmak amacıyla Naïve Bayes yöntemi uygulanmıştır. Bu amaçla, UCI Machine Learning Repository veri tabanında yer alan "Higher Education Students Performance Evaluation" veri seti kullanılmıştır. Veri seti, 33 değişken ve 145 örnekten meydana gelmektedir. Mevcut çalışmada; 30 değişken kullanılmıştır. Naïve Bayes algoritmasının uygulanmasının ardından çalışmada, Kazanç Oranı yöntemi %60'lık bir doğru sınıflama oranıyla en başarılı yöntem olarak belirlenmiştir. Kazanç Oranı kullanılarak seçilen özelliklere göre Naïve Bayes yöntemine ait doğru sınıflama oranı %60 ile, en iyi performans gösteren özellik seçim yöntemi olarak belirlenmiştir. Korelasyon tabanlı özellik seçim yöntemi hariç diğer tüm yöntemlerde öğrenci başarısını etkileyen en önemli faktör, öğrencinin son yarıyılda aldığı genel not ortalaması olarak tespit edilmiştir. Gain Ratio yöntemine göre diğer önemli değişkenler sırasıyla; cinsiyet, projelerin ve faaliyetlerin akademik başarıya etkisi, mezuniyet sonrası beklenen not ortalaması, öğrencinin haftalık ders çalışma saatleri, öğrencinin aldığı burs türü, akademik olmayan kitap ve dergi okuma sıklığı, annenin eğitim düzeyi, alanıyla ilgili yapılan seminer ve konferanslara katılım ile öğrencinin derslere katılım'dır. Diğer taraftan; One R yöntemi haricinde kullanılan diğer özellik seçim yöntemlerinin Naïve Bayes yönteminin performasını artırdığı görülmektedir. Özellik seçim yöntemlerinin veri madenciliği yöntemlerin verimliliğini artırmak amacıyla kullanılması önerilmektedir. Araştırma sonuçları, yalnızca Eğitsel Veri Madenciliğinde özellik seçiminin faydasını göstermekle kalmayıp aynı zamanda Eğitsel Veri Madenciliği alanındaki metodolojileri geliştirmekle ilgilenen araştırmacılar ve eğitimciler için değerli bilgiler sunmaktadır.

Anahtar kelimeler: Özellik seçimi, eğitsel veri madenciliği, naive bayes, yüksek öğrenim.

GİRİŞ

Eğitim, toplumlarınun gelişmişlik düzeyini gösteren en önemli öğelerden biridir. Bu nedenle ilkokuldan itibaren yüksek öğretime kadar bulunan tüm basamaklarda öğrenim gören öğrencilerin akademik başarılarını etkileyen faktörleri belirlemek önemlidir. Akademik planlamaya yardımcı olmak, çözümleri uygulamak ve stratejiler geliştirmek amacıyla birçok çalışma yapılmaktadır (Akçöltekin, Engin ve Şevgin, 2017). Araştırmacılar, öğrencilerin akademik başarısının yanı sıra öğrenme stilleri ve bunlarla ilişkili olabilecek çeşitli konuları daha iyi anlamak için geleneksel istatistiksel yöntemlerin yetersiz kaldığı durumlarda Eğitsel veri madenciliği (EDM) yöntemlerini kullanmaktadırlar (Baker, 2010; Fernandes ve ark., 2019; Şevgin ve Önen, 2022; Güre, 2023). EDM, eğitim alanındaki devasa büyüklükteki veri yığınlarını keşfetmek için kullanılan yöntemler topluluğudur (Sachin ve Vijay, 2012). EDM, veri tabanı sistemleri, veri ambarı, istatistik, makine öğrenimi gibi farklı alanları birleştiren bir disiplindir (Anuradha ve Velmurugan, 2016). EDM ile okulu bırakan öğrenci sayılarını tahmin etmek, öğrencinin üniversiteye giriş sınavı sonuçları ile başarısı arasındaki ilişkiyi belirlemek, öğrencinin akademik performansını yüksek doğrulukla tahmin edecek öğrenci modellerini geliştirmek mümkündür (Ramaswami ve Bhaskaran, 2009; Anuradha ve Velmurugan, 2016). Yöntemlerin kullanımı, karar vericilerin rasyonel karar alma süreçlerini desteklemeye ve eğitimin kalitesini artırmaya yardımcı olacaktır. Eğitim alanında diğer alanlarda olduğu gibi çok büyük miktarlarda ve çok sayıda değişkenden oluşan veri setleri meydana gelmektedir. Bu veri setleri içinden ilgisiz veya gereksiz değişkenlerin çıkarılması önemlidir (Budak, 2018). Özellik seçim yöntemleri kullanılarak, önemli verilerde herhangi bir kayıp olmadan fazla olan değişken sayıları en aza indirilir (Punlumjeak ve Rachburee, 2015). Yöntemler, bir tahmin modelinin performansını geliştirmede önemli rol oynayabilir. Seçilen özellikler hem tahmin doğruluğunu arttırmada hem de eğitim ortamına yönelik stratejik planların temelini oluşturmada etkili olabilir (Zaffar ve ark., 2020).

Bu çalışmada, üniversite öğrencilerinin başarılarını etkileyen faktörleri belirlemek amacıyla; özellik seçim yöntemlerinden Bilgi Kazancı (Information Gain-IG), Kazanç Oranı (Gain Ratio -GR), Korelasyon Tabanlı (Correlation Based- CB), Simetrik Belirsizlik (Symmetric Uncertainty- SU), Relief-F ve One R measure kullanılmıştır. Kullanılan özellik seçim yöntemlerinin performansını karşılaştırmak amacıyla Naïve Bayes yöntemi uygulanmıştır. Ayrıca; minimum hata ve yüksek doğruluk oranı ile tahminlemeler yapmanın yanı sıra özellik seçim yöntemlerinin performanslarını karşılaştırmak amaçlanmaktadır. Bu amaç doğrultusunda araştırmada aşağıdaki sorulara yanıt aranmıştır:

1. Özellik seçim yöntemlerine göre öğrencilerin başarısını etkileyen faktörler nelerdir?

2. Özellik seçim yöntemlerinin tahminleme performansı farklılaşmakta mıdır?

3. Özellik seçim yöntemlerinin kullanımı Naive Bayes yönteminin performansını etkilemekte midir?

İlişkili çalışmalar

Alan yazın incelendiğinde öğrenci başarısını tahminlemek amacıyla veri madenciliği yöntemlerinin kullanıldığı birçok çalışmanın olduğu görülmektedir. Bu çalışmalar arasında mevcut çalışmada kullanılan “Higher Education

Students Performance Evaluation" veri setini kullanan sınırlı sayıda çalışma bulunmaktadır. Hengpraprohm, Hengpraprohm ve Sudjitjoon, (2022), K-En Yakın Komşu (K-Nearest Neighbor-KNN), Rastgele Orman (Random Forest -RF), Yapay Sinir Ağları (Artificial Neural Network-ANN) ve Lineer Regresyon yöntemlerini kullandıkları çalışmalarında IG, GR, CB ve CS özellik seçim yöntemlerini uygulamışlardır. Yine, Jabardi (2022) bu amaçla RF, AdaBoost, Karar Ağaçları (Decision Tree-DT), Naive Bayes (NB) ve Çok Katmanlı Algılayıcı (Multi-Layer perceptron-MLP) yapay sinir ağları yöntemlerini kullanmıştır. Diğer taraftan; Phatai ve Luangrungruang (2023) ise öğrenci başarısını sınıflamak amacıyla ANN ile metaheuristic algorithms kullanmışlardır.

Alan yazın incelendiğinde, öğrenci başarısını tahminlemek amacıyla özellik seçim yöntemlerini kullanıldığı başka çalışmalar da bulunmaktadır. Göker, Bülbül ve Irmak (2013), NB, J48, Bayes Net ve Radyal Tabanlı fonksiyon (Radial Basis Function-RBF) yapay sinir ağları yöntemlerini kullandıkları çalışmalarında, IG, GR, SU, One R ve CS özellik seçim yöntemlerini kullanmışlardır. Yine Punlumjeak ve Rachburee (2015), NB, DT, KNN ve ANN yöntemlerine ek olarak genetik algoritmalar, Destek Vektör Makineleri (Support Vector Machine-SVM), IG, minimum redundancy ve maximum relevance gibi özellik seçim yöntemlerini kullanmışlardır. Benzer bir şekilde; Estrera ve ark., (2017) tarafından yapılan çalışmada; DT, NB ve KNN yöntemleri ile birlikte CS, IG ve GR özellik seçim yöntemlerini kullanmışlardır. Zaffar ve ark. (2018) ise bu yöntemlere ek olarak temel bileşenler analizi (principal component analysis) ve Relief yöntemini kullanmışlardır.

Bu kapsamlı literatür incelendiğinde, çeşitli algoritmalar ve özellik seçme yöntemleri arasında performans farklılıkları olduğu açıkça ortaya çıkmaktadır. Örneğin, Sokkhey ve Okazaki (2020) çalışmalarında, bu amaçla KNN, C5.0, RF ve Improved Deep Belief Networks- IDBN yöntemlerini kullanmışlardır. Özellik seçim yöntemleri olarak IG, SU, CS, Mutual information (MI)'nin yanı sıra Ki-kare (Chi-square) ve MI birleşiminden oluşan yeni bir özellik seçim yöntemini önermişlerdir. Önerdikleri yöntemin daha iyi performans gösterdiklerini belirtmektedirler. Yine Mythili ve Shanavas, (2014) ise J48, RF, MLP, IBI, KNN ve Decision Table yöntemlerini kullandıkları çalışmalarında IG ve GR özellik seçim yöntemlerini uygulamışlardır. Rahman Setiawan ve Permanasari, (2017), NB, DT ve ANN yöntemini kullandıkları çalışmalarında özellik seçim yöntemi olarak wrapper ve IG özellik seçim yöntemlerini kullanmışlardır. Bununla birlikte; Makhtar ve ark., (2017), öğrenci başarısını sınıflamak amacıyla NB, Random Tree, KNN, Multi class sınıflandırıcı ve Conjunctive Rule yöntemlerini kullandıkları çalışmalarında, özellik seçim yöntemi olarak Best First algorithm uygulamışlardır. Velmurugan ve Anuradha (2016) ise bu amaçla; J48, NB, Bayes Net, IBk, OneR, ve JRip yöntemlerini kullandıkları çalışmalarında; Best First Search, CB, Wrapper, CfsSubset, CS, IG ve Relief yöntemlerini kullanmışlardır. Buna ek olarak; Ramaswami ve Rathinasabapathy (2012), öğrencilerin başarısını tahminlemek amacıyla mevcut çalışmada kullanılan özellik seçim yöntemleri arasından One R yöntemi hariç diğer yöntemleri kullanmışlardır. Yahdin ve ark., (2021), Relief özellik seçim yönteminin performansını ölçmek amacıyla NB ve KNN yöntemlerini kullanmışlardır. Anuradha ve Velmurugan (2016) çalışmalarında CS ve GR özellik seçim yöntemlerinin NB yöntemini kullanarak etkisini belirlemeye çalışmışlardır.

YÖNTEM

Araştırmanın Modeli

Bu araştırma ilişkisel desen kullanılarak gerçekleştirilmiştir. Korelasyonel desen, birden fazla değişken arasındaki ilişkiyi inceler (Karasar, 2006; Güre, Kayri ve Erdoğan, 2020). Bu yaklaşımın kullanılması, mevcut çalışmada ilgilenilen değişkenler arasındaki karmaşık bağlantıları değerlendirmek için etkili bir çerçeve sağlamaktadır.

Kullanılan veri kümesi

Mevcut çalışmada, UCI veri tabanında bulunan "Yükseköğretim Öğrencileri Performans Değerlendirme veri seti" kullanılmaktadır (Yılmaz ve Şekeroğlu, 2019). Veri dosyası https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation adresinden elde edilmiştir. Veri seti, Eğitim ve Mühendislik fakültelerinde kayıtlı 145 (87 erkek (%60) ve 58 kız (%40)) öğrenciden toplanan bilgilerden oluşmaktadır. Bu bilgiler, ailesel, kişisel bilgiler ve çalışma alışkanlıklarıyla ilgili değişkenleri içermektedir. Veri seti, 33 değişken içermesine rağmen bu çalışmada değişkenlerden yalnızca 30'u kullanılmıştır.

Veri Analizi

Bu çalışmada veri analizlerini gerçekleştirmek için Weka veri madenciliği programı kullanılmıştır. Weka, Waikato Üniversitesi'nde geliştirilen, makine öğrenimi algoritmaları ve veri ön işleme araçlarından oluşan bir koleksiyon içeren ücretsiz, açık kaynaklı bir platformdur (Hall ve diğerleri, 2009). Bu araştırma için Weka, istatistiksel analiz ve makine öğrenimi tekniklerine yönelik araçları aracılığıyla veri kümesi içindeki ilişkilerin sağlam bir şekilde incelenmesini sağlamaktadır.

Özellik Seçim Yöntemleri

Özellik seçme yöntemleri veri madenciliği alanında özellikle yüksek boyutlu veri setlerinde önemli bir konu olarak görülmektedir. Bu yöntemler makine öğreniminde yaygın olarak kullanılmaktadır. Özellik seçme yöntemleri, ilgisiz ve gereksiz özellikleri silerek özellik kümesinden alt kümeleri seçer. Özellik, öznitelik veya değişken olarak adlandırılmaktadır. Özel seçim, orijinal veri setini temsil etme özelliğine sahip ve en iyi alternatif seçim olarak kullanılabilir (Budak, 2018). Özel seçim veri önişleme tekniğidir (Phyu ve Oo, 2016). Böylece daha az özellik elde edilerek sonuçların araştırılması kolaylaşır. Yöntemler farklı alanlarda yaygın olarak kullanılmaktadır (Guan vd., 2014). Son yıllarda özellik seçim yöntemleri topluluk öğrenmesi ile entegre edilerek, topluluk öğrenme tabanlı öznitelik seçimi yaklaşımı önerilmiş ve çalışılmıştır. Genel fikir, çok çeşitli özellik seçicileri oluşturmak ve bunların çıktılarını birleştirmektir. Özellik seçim yöntemleri, boyut indirgeme amacıyla kullanılan tekniklerden biridir (Remeseiro ve Bolon-Canedo, 2019). Yöntemler, ilişkili olmayan ve gereksiz olan özellikleri silerek özellik kümesinden alt kümeler seçmektedir. Böylelikle; daha az özellik elde edilerek sonuçların araştırılması kolaylaşmaktadır (Guan ve ark.,2014).

Özellik seçim yöntemleri, gürültülü ve ilişkisiz verileri silme, özellik kümesinin boyutunu düşürme, algoritmanın hızını arttırma, verilerin kalitesini arttırma, elde edilen modelin performansını arttırma, özellik setini azaltma,

tahminleme doğruluğunu arttırma gibi özelliklere sahiptir (Ladha ve Deepa, 2011). Bu özelliklerinden dolayı, son yıllarda özellik seçimiyle ilgili çeşitli yöntemler geliştirilmiştir (Solorio-Fernández, Carrasco-Ochoa ve Martínez-Trinidad, 2020). Yöntemler farklı alanlarda sınıflama, kümeleme ve regresyon amaçlı olarak kullanılmaktadır (Jović, Brkić ve Bogunović, 2015).

Özellik seçim yöntemlerini; denetimli, yarı denetimli ve denetimsiz yöntemler olarak sınıflamak mümkündür. Denetimli yöntemler, ilgili özellikleri belirlemek ve seçmek için bir dizi etiketli veri veya denetimli veri kümesine ihtiyaç duymaktadır. Veri kümesindeki her nesneye atanan bu etiket, bir kategori, sıralı bir değer veya gerçek bir değer olabilir. Yarı denetimli yöntemlerde, yalnızca bazı nesnelerin etiketlenmesine ihtiyaç duyulmakta, denetimsiz yöntemlerde ise denetimli veri setine ihtiyaç duyulmamaktadır (Solorio-Fernández, Carrasco-Ochoa ve Martínez-Trinidad, 2020).

Literatürde farklı özellik seçim yöntemleri geliştirilmiştir. Çalışmada sıralama yöntemlerinden Bilgi Kazancı, Kazanç oranı ve Simetrik Belirsizlik Katsayısı, Relief-F, Korelayona Dayalı Özellik Seçim Yöntemi ve One R ölçüsü ele alınmıştır.

Bilgi Kazancı Özellik Seçim Yöntemi

Bilgi Kuramında örneklerin homojenliği hakkında bilgi veren entropi ölçüsü yaygın olarak kullanılmaktadır (Abe ve Kudo, 2005). Entropi ölçüsü 0 ila 1 arasında değişen bir ölçü olup, entropinin sıfır (0) olması, örneklerin homojen, entropinin bir (1) olması örneklerin homojen olmadığı anlamına gelmektedir (Sastry ve ark., 2010). Bilgi Kazancı, Makine öğrenmesinde yaygın olarak kullanılan entropiye dayalı özellik seçim yöntemlerindendir. Yöntem bağımlı sınıfla ilgili en çok bilgi veren özellikleri belirleyebilme özelliğine sahiptir (Win ve Kham, 2019). Bilgi Kazancı, Y özelliğindeki entropi miktarının azalmasını temsil eden X özelliğinin kullanılmasıyla elde edilen Y özelliğindeki ek bilgiyi yansıtan bir ölçüdür. Simetrik olan ölçüde, X gözlemlendikten sonra Y ile ilgili kazanılmış bilgi ile Y gözlemlendikten sonra X ile ilgili kazanılmış bilgi birbirine eşittir (Novakovıć, Strbac ve Bulatovıć, 2011; Budak, 2018).

Yöntem, sınıflandırıcılardan bağımsız olmasından dolayı birçok sınıflandırıcı ile uygulanabilme özelliğine sahiptir (Win ve Kham, 2019). Bilgi Kazancı, daha az bilgi ile daha yüksek sayısal özelliklere sahip özelliklerin seçimine yönelik yanlı davranmaktadır. Bu da yöntemin zayıflığı olarak görülmektedir (Tripathi ve Trivedi, 2016; Budak, 2018).

$$H(Y) = -\sum_{y\epsilon Y} p(y) log_2 \left(p(y)\right) \quad (1)$$

$$H(Y\backslash X) = -\sum_{x\epsilon X} p(x) \sum_{y\epsilon Y} p(y\backslash x) log_2 (p(y\backslash x)) \quad (2)$$

$$Bilgi\ Kazancı = H(Y) - H(X\backslash Y) \quad (3)$$

Burada; H(X) X'in entropisi iken H(X\Y) ise Y gözlendikten sonraki X'in entropisini göstermektedir (Win ve Kham, 2019).

Kazanç Oranı Özellik Seçim Yöntemi

Kazanç Oranı, Bilgi Kazancındaki yanlılığın üstesinden gelmek için geliştirilen simetrik olmayan bir ölçüdür (Novakovıć, Strbac ve Bulatovıć, 2011; Tripathi ve Trivedi, 2016; Janabi ve Kadhim, 2018).

$$Kazanç\ Oranı = \frac{\text{Bilgi Kazancı}}{\text{H(X)}} \quad (4)$$

Kazanç Oranı, sınıflandırıcı karmaşıklığı ve doğruluk özelliğinden dolayı, basit Bilgi Kazancı ölçütlerinden daha iyi performans gösterme eğilimindedir (Quinlan, 1988; Rokach ve Maimon, 2005).

Simetrik Belirsizlik Katsayısı Özellik Seçim Yöntemi

Simetrik Belirsizlik Katsayısı, Bilgi Kazancı'nındaki doğal yanlılığın üstesinden gelmek amacıyla geliştirilen ölçü olup, Entropiye benzer olarak 0 ila 1 arasında değer almaktadır. Katsayının 1 olması X bilgisinin tamamen Y bilgisini tahmin edebildiğini, 0 olması Y ile X arasında hiçbir ilişki olmadığını göstermektedir (Hall, 1999; Budak, 2018; Janabi ve Kadhim, 2018).

Simetrik Belirsizlik Katsayısı, değerini [0,1] aralığına normalize ederek daha fazla değere sahip özelliklere yönelik bilgi kazancının önyargısının üstesinden gelir. Aşağıdaki denklem ile verilir:

$$Simetrik\ Belirsizlik\ Katsayısı = 2\frac{Bilgi\ Kazancı}{H(X)+H(Y)} \quad (5)$$

Relief-F Özellik Seçim Yöntemi

Uzaklığa dayalı bir ölçü olarak geliştirilmiş olan Relief yöntemi, iki sınıflı problemlerde özellik seçim yöntemi olarak kullanılmaktadır (Win ve Khan, 2019). Relief ölçüsü, aynı ve farklı sınıflarda yer alan en yakın örneklere ait verilen özelliğin değerini dikkate alarak ve bir örneği yinelemeli örneklemler ile elde ederek değerlendirme yapmaktadır. Yöntem, sınıflar arasında ayırım yapabilme yeteneğine bağlı olarak, hedef sınıf ile ilişkili her bir özelliğe bir ağırlık atamaktadır. Yöntem, özelliklerin en küçük alt kümesini bulmasa da istatistiksel olarak ilişkili olan özellikleri seçmektedir Daha sonra, özellikler sıralanarak kullanıcı tarafından tanımlanan eşik değeri veren ağırlıklara ulaşan özellikleri seçmektedir (Kira ve Rendell, 1992; Novakovıć, Strbac ve Bulatovıć, 2011; Budak, 2018). Relief algoritmaları yaygın olarak model öğrenilmeden önce bir ön hazırlık aşamasında uygulanan öznitelik alt küme seçim yöntemleri olarak görülmektedir (Kira ve Rendell, 1992).

Heuristic yöntemlerde, özelliklerin kalitesini tahmin etmek için özelliklerin bağımsızlığı koşulu şartı aranmaktadır. Yöntemler, özellik etkileşimini konu alan problemlerde çok uygun olmamaktadır. Bu yöntemlerden farklı olarak, Relief yönteminde herhangi bir varsayım bulunmamaktadır. Yöntem, özellikler arasında güçlü bağlılıkların olduğu problemlerde, özelliklerin kalitesini doğru bir şekilde tahmin edebilmektedir (Marko ve Igor, 2003).

Korelayona Dayalı Özellik Seçim Yöntemi

Korelayona Dayalı Özellik Seçim Yöntemi, özellik alt kümelerini korelasyona dayalı olarak sıralayan basit heuristik özellik alt seti seçme yöntemlerindendir (Hall, 1999). Yöntem, tüm özellikler arasındaki korelasyonu hesaplamaktadır. Sınıf değişkenleri ile yüksek korelasyona sahip ancak; birbirleriyle düşük korelasyona sahip özelliklere sahip alt kümeyi seçmektedir. Yöntem, nominal ya da kategorik özellikler arasındaki korelasyonu ölçmektedir (Janabi ve Kadhim, 2018).

$$\boldsymbol{M_s} = \frac{k\bar{r}_{c_i}}{\sqrt{k+k(k-1)\bar{r}_{ii}}} \quad (6)$$

One R özellik seçim yöntemi

Holte tarafından önerilen, hata oranlarına göre özellikleri seçen basit bir yöntemdir. Eğitim verisinde her bir özellik için bir kural oluşturmaktadır. Ardından en küçük hataya sahip özelliği seçmektedir (Novaković, Strbac ve Bulatović, 2011). Sayısal değerli tüm özellikleri sürekli olarak ele alır ve değer aralığını birkaç ayrık aralığa bölmek için basit bir yöntem kullanmaktadır. Eksik değerleri, geçerli bir değer olarak ele alarak kayıp verileri işlemektedir (Janabi ve Kadhim, 2018). Bu yöntem, birçok özelliğe ve farklı sınıflara sahip bir dizi örnekle çalışmaktadır. Yinelemeli olarak en iyi tek bir özelliği seçer ve kuralları yalnızca o özelliğe dayandırmaktadır (Tripathi ve Trivedi, 2016).

Naive Bayes

Naive Bayes yöntemi, sınıflandırma amacıyla Bayes teoremini kullanmaktadır. Yöntem, doğru makine öğrenimi modellerinin hızla oluşturulmasını sağlayan basit, hızlı ve etkili olasılık tabanlı bir sınıflandırma yaklaşımıdır (Rish, 2001; Jabardi, 2022; Wickramasinghe ve Kalutarage, 2021). NB yönteminin hesaplama, düşük varyans, artımlı öğrenme, sonsal olasılıkların doğrudan tahmini, gürültülü verilere ve kayıp verilere karşı dayanıklılık özellikleri bulunmaktadır (Sammut ve Webb, 2017). NB, karmaşık yinelemeli parametre tahminlerine gerek duymamaktadır. Bu nedenle yapılanması ve yorumlanması kolaydır (Wu ve ark., 2008). Özellikle büyük veri kümelerinde iyi performans gösterdiği belirtilmektedir (Nikam, 2015; Jabardi, 2022). NB yöntemi, her bir ilişki için koşullu bir olasılık elde etmektedir. Elde edilen olasılık ile bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi analiz etmektedir (Muda ve ark., 2011).

NB, Bayes kuralına dayalı bir sınıflandırıcıdır.

$$\mathrm{P}(\mathrm{y}\backslash\mathbf{x}) = (\mathrm{P}(\mathrm{y})\mathrm{P}(\mathbf{x}\backslash\mathrm{y}))/\mathrm{P}(\mathbf{x}) \quad (7)$$

niteliklerin sınıfa göre koşullu olarak bağımsız olduğu varsayarak Nitelik değeri verileri için bu varsayım şunları sağlar:

$$P(x\backslash y) = \prod_{i=1}^{n} P(x_i\backslash y) \quad (8)$$

x_i, x'teki i.ninci özelliğin değerini, n ise özellik sayısnı göstermektedir.

$$P(\boldsymbol{x}) = \prod_{i=1}^{k} P(c_i)P(\boldsymbol{x}\backslash c_i) \quad (9)$$

c_i, i.ninci sınıfı, k ise sınıf sayısını göstermektedir. Eşitliğin sağ tarafında yer alan bölüm normalleştirilirse (7) numaralı eşitlik elde edilir (Webb, Keog ve Miikkulainen, 2010).

BULGULAR

Değerlendirme Ölçütleri

Araştırmada performans kriteri olarak doğruluk, kesinlik, duyarlılık ve F1 skor ölçütleri kullanılmıştır. Performans kriterlerine ilişkin denklemler aşağıda verilmiştir:

Tablo 1. Karışıklık Matrisi

		Tahmin edilen sınıf		
		Hayır	Evet	Toplam
Gerçek sınıf	Hayır	TN	FP	TN+FP
	Evet	FN	TP	FN+TP
	Toplam	TN+FN	FP+TP	TN+FN+FP+TP

$$\text{Doğruluk} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (10)$$

$$Kesinlik = \frac{TP}{TP+FP} \quad (11)$$

$$Duyarlılık = \frac{TP}{TP+FN} \quad (12)$$

$$F1\ skor = 2x\ \frac{(Kesinlik\ x\ Duyarlılık)}{(Kesinlik+ Duyarlılık)} \quad (13)$$

Özellik seçme yöntemleri kullanılarak elde edilen en önemli 10 değişken Tablo 2'de verilmiştir.

Tablo 2. Naïve Bayes yöntemi kullanılarak özellik seçim yöntemlerinin performansı

IG	GR	SU	Relief F	Correlation based	One R
Son yarıyıldaki genel not ortalaması	Son yarıyıldaki genel not ortalaması	Son yarıyıldaki genel not ortalaması	Son yarıyıldaki genel not ortalaması	Cinsiyet	Son yarıyıldaki genel not ortalaması
Mezuniyette beklenen genel not ortalaması	Cinsiyet	Cinsiyet	Burs türü	Son yarıyıldaki genel not ortalaması	Mezuniyette beklenen genel not ortalaması
Haftalık çalışma saatleri	Proje/faaliyetlerin başarınıza etkisi	Mezuniyette beklenen genel not ortalaması	Cinsiyet	Bölümle ilgili seminer/konferanslara katılım	Kardeş sayısı
Anne eğitim düzeyi	Mezuniyette beklenen genel not ortalaması	Haftalık çalışma saatleri	Yaş	Derslere katılım	Haftalık çalışma saatleri
Cinsiyet	Haftalık çalışma saatleri	Proje/faaliyetlerin başarınıza etkisi	Anne eğitim düzeyi	Gelir	Dersi dinleme
Burs türü	Burs türü	Anne eğitim düzeyi	Konaklama biçimi	Proje/faaliyetlerin başarınıza etkisi	Düzenli sanatsal veya sportif aktivite yapma
Baba mesleği	Okuma sıklığı (bilimsel olmayan kitaplar/dergiler)	Burs türü	Haftalık çalışma saatleri	Anne eğitim düzeyi	Mezun olunan lise türü

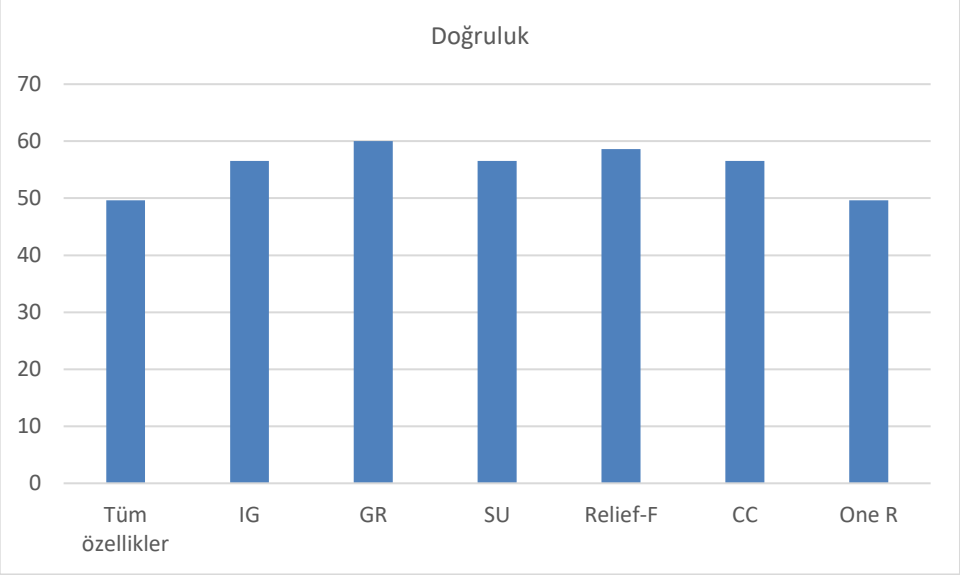
Okuma sıklığı (bilimsel olmayan kitaplar/dergiler)	Anne eğitim düzeyi	Okuma sıklığı (bilimsel olmayan kitaplar/dergiler)	Mezun olunan lise türü	Öğretmenin derste teknolojik alet kullanması	Üniversiteye Ulaşım
Proje/faaliyetlerin başarınıza etkisi	Bölümle ilgili seminer/konferanslara katılım	Anne mesleği	Okuma sıklığı (bilimsel olmayan kitaplar/dergiler)	Burs türü	Okuma sıklığı (bilimsel kitaplar/dergiler)
Baba eğitim düzeyi	Derslere katılım	Gelir	Öğretmenin derste teknolojik alet kullanması	Mezuniyette beklenen genel not ortalaması	Bölümle ilgili seminer/konferanslara katılım

Ardından seçilen özellikler kullanılarak Weka programında Naïve Bayes yöntemiyle analizler gerçekleştirilmiştir. Analiz sonuçları aşağıda gösterilmiştir.

Tablo 3. Naïve Bayes Yöntemi Kullanılarak Özellik Seçim Yöntemlerinin Performansı

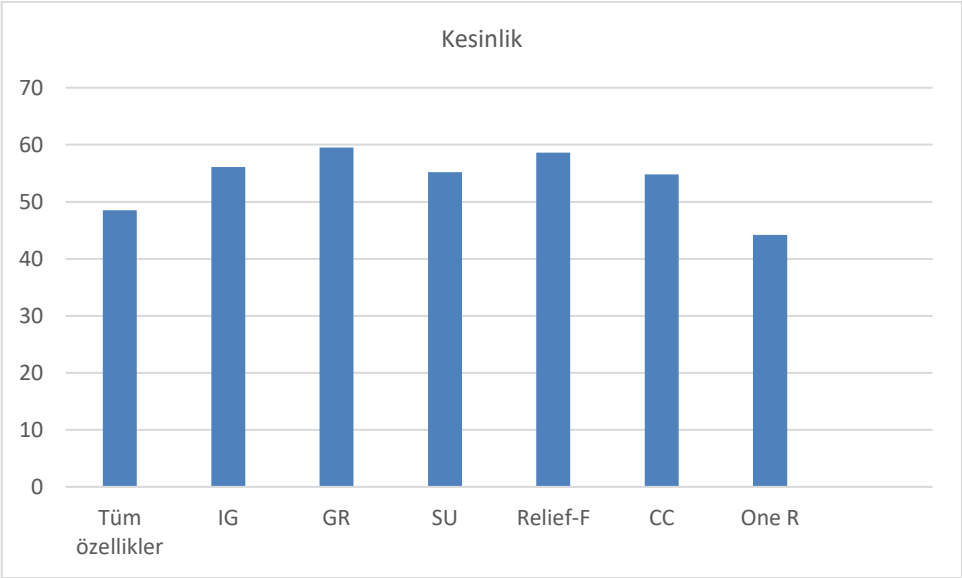
Yöntem	Özellik seçim yöntemleri	Doğruluk	Kesinlik	Duyarlılık	F1- skor
Naïve Bayes	Tüm özellikler	49.65	0.48	0.49	0.48
	IG	56.55	0.561	0.566	0.562
	GR	60.00	0.595	0.600	0.597
	SU	56.55	0.552	0.566	0.554
	Relief-F	58.62	0.586	0.586	0.580
	CB	56.55	0.548	0.566	0.553
	One R	49.65	0.442	0.497	0.465

Tablo 3'ten anlaşıldığı üzere Naïve Bayes yöntemi ile öğrenci başarısını etkileyen faktörleri belirleyen en iyi özellik seçim yöntemleri olarak GR yöntemi belirlenmiştir. One R yöntemi hariç diğer tüm yöntemlerin NB yönteminin performansını artırdığı söylenebilir.



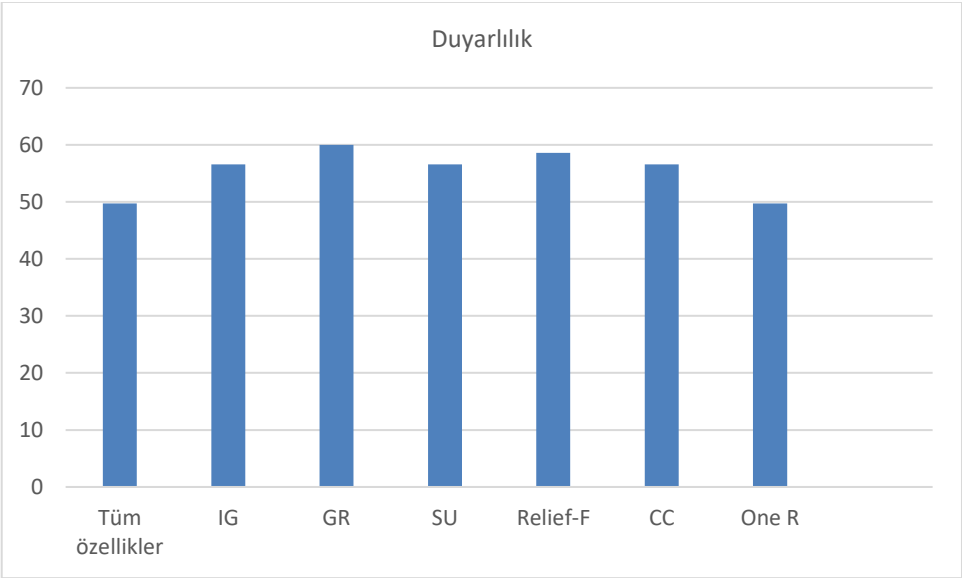
Şekil 1. Özellik seçim yöntemlerinin doğruluk oranına göre dağılımı

Şekil 1 incelendiğinde doğruluk açısından GR yönteminin daha iyi performans gösterdiği görülmektedir. One R yöntemi, tüm özelliklerin kullanıldığı analiz sonuçlarına benzer doğruluk oranına sahiptir. Diğer yöntemlerin Naive Bayes yönteminin sınıflama performansını artırdığı görülmektedir.



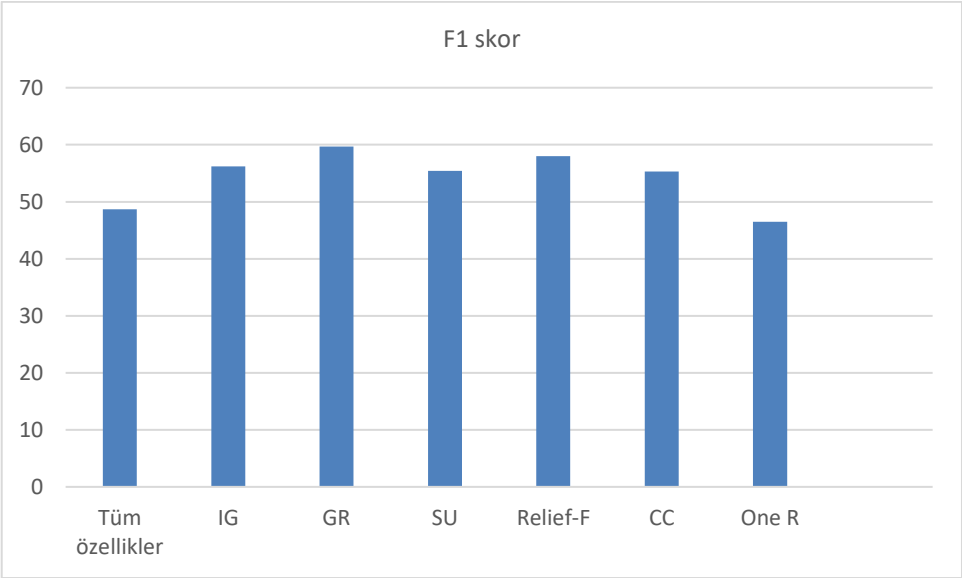
Şekil 2. Özellik seçim yöntemlerinin kesinlik oranına göre dağılımı

Şekil 2 incelendiğinde kesinlik bakımından GR yönteminin en yüksek One R yönteminin ise en düşük orana sahip olduğu görülmektedir.



Şekil 3. Özellik seçim yöntemlerinin duyarlılık oranına göre dağılımı

Şekil 3 incelendiğinde duyarlılık açısından GR yönteminin daha yüksek orana sahip olduğu görülmektedir.



Şekil 4. Özellik seçim yöntemlerinin F1 skor ölçüsüne göre dağılımı

Şekil 4 incelendiğinde F1 skor ölçüsü açısından GR yönteminin en yüksek, One R yönteminin ise en düşük orana sahip olduğu görülmektedir.

TARTIŞMA ve SONUÇ

Bu çalışmada; özellik seçim yöntemlerinden IG, GR, SU, CB, Relief-F, ve One R measure kullanılarak, üniversite öğrencilerinin başarılarını etkileyen faktörleri belirlemek amaçlanmaktadır. Özellik seçim yöntemlerinin etkisini karşılaştırmak amacıyla, NB yöntemi uygulanmıştır. NB yöntemi ile öğrenci başarısını etkileyen faktörleri belirleyen en iyi özellik seçim yöntemi %60 doğruluk oranı ile GR olarak belirlenmiştir. Analiz sonuçlarına göre; CB method hariç diğer tüm yöntemlerde öğrenci başarısını etkileyen en önemli faktörün öğrencinin son yarıyıl genel not ortalaması olduğu görülmektedir. En önemli değişkenden sonra gelen diğer önemli değişken, Relief-F özellik seçim yöntemine göre öğrencinin aldığı burs türüdür. IG yöntemine göre ise sırasıyla; mezuniyette beklenen not ortalaması, haftalık ders calışma saati, anne eğitim düzeyi, cinsiyet, öğrencinin aldığı burs türü, baba mesleği ve bilimsel olmayan kitap okuma sıklığı değişkenleridir.

Çalışmamıza benzer olarak; Hengpraprohm, Hengpraprohm ve Sudjitjoon, (2022), KNN, RF, ANN ve Linear Regression yöntemlerini kullandıkları çalışmalarında IG, GR, CB, CS özellik seçim yöntemlerini uyguladıkları çalışmalarında IG ve GR yöntemlerini en başarılı yöntemler olarak belirlemişlerdir. GR yöntemi kullanıldığında KNN yönteminin accuracy %48.53 ve RF yönteminde %76.07 olarak; IG yöntem kullandıldığında ise ANN %40.0 ve Lineer regression %54.37 olarak belirlenmiştir. Ayrıca; söz konusu çalışmada; öğrencinin son yarıyıldaki genel not ortalaması, akademik başarı beklenti puanı, bilimsel olmayan kitap okuma sıklığı ve boşanmış ya da ölmüş ebeveyn sahipliği önemli değişkenler olarak belirlenmiştir. Yine, Jabardi (2022) bu amaçla RF, AdaBoost, DT, NB ve MLP yöntemlerini kullandıkları çalışmalarında; accuracy sırasıyla; %78.62, %84.82, %74.48, %66.20 ve %73.10' olarak bulmuşlardır. Adaboost yönteminin diğer yöntemlere nazaran daha başarılı performans gösterdiğini belirtmişlerdir.

Punlumjeak ve Rachburee (2015), NB, DT, KNN ve ANN yöntemlerine ek olarak özellik seçim yöntemleri olarak genetik algoritmalar, SVM, IG, minimum redundancy ve maximum relevance gibi özellik seçim yöntemlerini kullanmışlardır. Çalışmada, en iyi performansı minimum redundancy and maximum relevance feature selection method kullanıldığı KNN yöntemi ile elde edilmiştir. NB yönteminin doğru sınıflama oranına bakıldığında en yüksek performans %83.87 ile SVM özellik seçimi yapıldığı durumda gerçekleşmiştir. Bununla birlikte; Göker, Bülbül ve Irmak (2013) ise NB, J48, Bayes Net ve RBF yöntemlerini kullandıkları çalışmalarında, IG, GR, SU, One R ve CS özellik seçim yöntemlerini kullanmışlardır. En iyi özellik seçim yönteminin %83,87 ortalama ile SU yöntemine ait olduğunu belirlemişlerdir. Ayrıca; Rahman Setiawan ve Permanasari, (2017), NB, DT ve ANN yöntemini kullandıkları çalışmalarında özellik seçim yöntemi olarak Wrapper ve IG yöntemlerini kullanmışlardır. Çalışmada, Wrapper özellik seçim yöntemi kullanılarak NB yönteminin doğru sınıflama oranı %75,41 olarak belirlenmştir. Özellik seçim yöntemlerinin kullanımının NB yönteminin performansını artırdığı belirtilmektedir. Buna ek olarak; Velmurugan ve Anuradha (2016) bu amaçla; J48, NB, Bayes Net, IBk, OneR, ve JRip yöntemlerini kullandıkları çalışmalarında; Best First Search, Wrapper, CfsSubset, CS, IG ve Relief yöntemlerini kullanmışlardır. Çalışmada en iyi sonuç CfsSubset özellik seçim yöntemiyle elde edilmiştir. Naïve bayes yönteminin sınıflama performansının bu özellik seçim yöntemine göre % 98.56 olduğu belirlenmiştir. Diğer taraftan; Anuradha ve Velmurugan (2016) çalışmalarında CB ve GR özellik seçim yöntemlerinin etkisini NB yöntemini kullanarak belirlemeye çalışmışlardır. CB yönteminin GR'ye göre daha başarılı performans gösterdiği belirtilmektedir. Çalışmada NB yöntemine ait doğru sınıflama oranı %84 olarak belirlenmiştir. Özellik seçim yöntemlerinin sınıflama performansını artırdığı belirtilmektedir. Yine, Makhtar ve ark., (2017), Priyasadie ve Isa (2021 ve Yahdin ve ark. (2021) tarafından yapılan çalışmalarda da benzer sonuçlar elde edilmiştir.

Ramaswami ve Rathinasabapathy (2012), öğrencilerin başarısını tahminlemek amacıyla mevcut çalışmada kullanılan özellik seçim yöntemleri arasından One R yöntemi hariç diğer yöntemleri kullanmışlardır. Özellik seçim yöntemlerinin tahminleme performansını artırmasının yanı sıra yöntemlerin çalışma sürelerini azalttığını belirtmektedirler.

Analiz sonuçlarına göre, kullanılan özellik seçim yöntemlerinin NB yönteminin sınıflama performansını artırdığı görülmektedir. Alan yazında bu amaçla yapılan araştırmaların sonuçları da çalışmamızı destekler niteliktedir.

ÖNERİLER

Gelecekteki çalışmalarda; Özellik seçim yöntemlerinin, veri madenciliği yöntemlerinin verimliliğini artırmak amacıyla kullanılması önerilmektedir. Diğer taraftan; farklı veri madenciliği yöntemleri kullanılarak yöntemlerin performansları değerlendirilebilir. Ayrıca; farklı veri setleri üzerinde yeni çalışmalar yapılabilir. Çalışmada özellik seçim yöntemlerinin değerlendirilmesinde Weka program kullanılmıştır. Farklı yazılımlar ile çalışmaların yapılması önerilmektedir.

Etik Metni

"Bu makalede dergi yazım kurallarına, yayın ilkelerine, araştırma ve yayın etiği kurallarına, dergi etik kurallarına uyulmuştur. Makale ile ilgili doğabilecek her türlü ihlallerde sorumluluk yazar(lar)a aittir. Mevcut çalışmada kullanılan veri seti, açık erişimli UCI veri tabanından elde edilmiştir. Bu nedenle etik kurul onayı alınmasını gerektiren bir çalışma değildir."

Yazar(lar)ın Katkı Oranı Beyanı: Bu çalışmada ilk yazarın katkı oranı %100'dür.

KAYNAKÇA

Abe, N. & Kudo, M., (2005, September 14-16,). "Entropy criterion for classifier-independent feature selection" [Oral presentation]. 9th International Conference, KES 2005, Melbourne, Australia.

Akcoltekin, A., Engin, A. O., & Sevgin, H. (2017). Attitudes of high school teachers to educational research using classification-tree method. Eurasian Journal of Educational Research, 17(68), 19-47. https://dergipark.org.tr/en/pub/ejer/issue/42457/511275

Al Janabi, K. B., & Kadhim, R. (2018). Data reduction techniques: a comparative study for attribute selection methods. International Journal of Advanced Computer Science and Technology, 8(1), 1-13.

Anuradha, C., & Velmurugan, T. (2016, January, 19-21). Feature selection techniques to analyse student acadamic performance using Naïve Bayes classifier [Oral presentation]. In The 3rd international conference on small & medium business. Hochiminh, Vietnam

Baker, R. S. J. D. (2010). Data mining for education. International encyclopedia of education. Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition) (pp.112-118.. Oxford, UK: Elsevier.

Bezek Güre, Ö. (2023). Investigation of ensemble methods in terms of statistics: TIMMS 2019 example. Neural Computing and Applications, 1-14. https://doi.org/10.1007/s00521-023-08969-0

Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi. 22, 21-31. https://doi.org/10.19113/sdufbed.01653

Estrera, P. J. M., Natan, P. E., Rivera, B. G. T., & Colarte, F. B. (2017). Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School Abstract. International Journal of Engineering and Technology, 3(5), 147-153. http://www.ijetjournal.org/

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of business research, 94, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

Göker, H., Bülbül, H. I., & Irmak, E. (2013, December, 04-07). The estimation of students' academic success by data mining methods. In 2013 12th International Conference on Machine Learning and Applications (Vol. 2, pp. 535-539). IEEE. Miami, FL, USA

Guan. D.. Yuan. W.. Lee. Y. K.. Najeebullah. K.& Rasel. M. K. (2014). A review of ensemble learning based feature selection. IETE Technical Review. 31(3). 190-198. https://doi.org/10.1080/02564602.2014.906859

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of Factors Effecting PISA 2015 Mathematics Literacy via Educational Data Mining. Education & Science/Egitim ve Bilim, 45(202). http://dx.doi.org/10.15390/EB.2020.8477

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM SIGKDD explorations newsletter, 11(1), 10-18. https://dl.acm.org/doi/abs/10.1145/1656274.1656278

Hall. M. (1999). Correlation-based Feature Selection for Machine Learning. [PhD Thesis]. The University of Waikato.

Hengpraprohm, K., Hengpraprohm, S., & Sudjitjoon, W. (2022). A Study of Factors Affecting Learning Efficiency on Higher Education Student Performance Evaluation Dataset Using Feature Selection Techniques. Information Technology Journal, 18(2), 34-43. https://ph01.tci-thaijo.org/index.php/IT_Journal/article/view/251051

Jabardi, M. H. (2022). Machine learning techniques for assessing students'environments'impact factors on their academic performance. International Journal of Advanced Research in Computer Science, 13(2). http://dx.doi.org/10.26483/ijarcs.v13i2.6813

Jović, A., Brkić, K., & Bogunović, N. (2015, May, 25-29). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). IEEE. Opatija, Croatia

Karasar, N. (2009). Bilimsel araştırma yöntemi (23. bs.). Ankara: Nobel Yayınlar

Kira, K..& Rendell, L.A. (1992). A practical approach to feature selection. In D. Sleeman. & P. Edwards (Eds.). Machine Learning: Proceedings of International Conference (ICML'92) (pp. 249–256). Morgan Kaufmann

Ladha. L. & Deepa. T. (2011). Feature Selection Methods And Algorithms. International Journal on Computer Science and Engineering. 3(5). 1787-1797

Makhtar, M., Nawang, H., & Wan Shamsuddin, s. N. (2017). Analysis on students performance using naïve bayes classifier. Journal of Theoretical & Applied Information Technology, 95(16). https://www.researchgate.net/profile/Hasnah-Nawang/publication/319955477_Analysis_on_students_performance_using_naIve_Bayes_classifier/links/60ebca8cb8c0d5588cee6bfa/Analysis-on-students-performance-using-naIve-Bayes-classifier.pdf

Marko. R.S., & Igor. K. (2003). "Theoretical and empirical analysis of relief and rreliefF". Machine Learning Journal. 53 23–69. https://doi.org/10.1023/A:1025667309714

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011, July, 12-13). Intrusion detection based on K-Means clustering and Naïve Bayes classification. In 2011 7th international conference on information technology in Asia (pp. 1-6). IEEE. Sarawak, Malaysia

Mythili, M. S., & Shanavas, A. M. (2014). An Analysis of students' performance using classification algorithms. IOSR Journal of Computer Engineering, 16(1), 63-69. https://www.researchgate.net/profile/Mohamed-Shanavas/publication/314445897_An_Analysis_of_students'_performance_using_classification_algorithm

s/links/58d5efff92851c44d461e5af/An-Analysis-of-students-performance-using-classification-algorithms.pdf

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science and Technology, 8(1), 13-19. http://www.computerscijournal.org/?p=1592

Novakavic. J.. Strbac. P.. Bulatovic. D. (2011). Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms. Yugoslav Journal of Operations Research. 21(1). 119-135. http://www.yujor.fon.bg.ac.rs/index.php/yujor/article/download/364/255

Phatai, G., & Luangrungruang, T. (2023, March, 18-20). A Comparative Study of Hybrid Neural Network with Metaheuristics for Student Performance Classification. In 2023 11th International Conference on Information and Education Technology (ICIET) (pp. 448-452). IEEE. Fujisawa, Japan

Phyu, T. Z., & Oo, N. N. (2016). Performance comparison of feature selection methods. In MATEC web of conferences (Vol. 42, p. 06002). EDP Sciences. https://doi.org/10.1051/matecconf/20164206002

Priyasadie, N., & Sani, M. I. (2021). Educational Data Mining in Predicting Student Final Grades on Standardized Indonesia Data Pokok Pendidikan Data Set. International Journal of Advanced Computer Science and Applications, 12(12). https://doi.org/10.14569/IJACSA.2021.0121227

Punlumjeak, W., & Rachburee, N. (2015, October, 29-30). A comparative study of feature selection techniques for classify student performance. In 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 425-429). IEEE. Chiang Mai, Thailand

Quinlan. J.R.. (1988). Decision trees and multivalued attributes. J. Richards. ed.. Machine Intelligence. 11: 305-318. Oxford University Press

Rahman, L., Setiawan, N. A., & Permanasari, A. E. (2017, November, 01-02). Feature selection methods in improving accuracy of classifying students' academic performance. In 2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE) (pp. 267-271). IEEE. Yogyakarta, Indonesia

Ramaswami, M., & Rathinasabapathy, R. (2012). Student performance prediction. International Journal of Computational Intelligence and Informatics, 1(4), 231-235. https://www.academia.edu/download/34746728/IJCII_1-4-38_-_Ramasamy.pdf

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

Rokach. L.. Maimon. O.. (2005). Data Mining and Knowledge Discovery Handbook. Springer New York Dordrecht Heidelberg London.1285.

Sachin, R. B., & Vijay, M. S. (2012, January, 07-08). A survey and future vision of data mining in educational field. In 2012 second international conference on advanced computing & communication Technologies, IEEE, 96-100. Rohtak, India

Sastry PM. Krishnan R. Ram .B.V.S.. (2010). Classification and identification of teluguh and written characters extracted from palm leavesusing decision tree approach. ARPN Journal of Engineering and Applied Sciences. 5 (3): 22-32. https://www.researchgate.net/profile/Narahari-Sastry/publication/242591979_Classification_and_identification_of_Telugu_handwritten_characters_extracted_from_palm_leaves_using_decision_tree_approach/links/56152f4508aed47facefb7bd/Classification-and-identification-of-Telugu-handwritten-characters-extracted-from-palm-leaves-using-decision-tree-approach.pdf

Şevgin, H. & Önen, E. (2022). Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE- 2016 Case. Education & Science/Egitim ve Bilim, , 47(211). http://dx.doi.org/10.15390/EB.2022.10575

Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. International Journal of Advanced Computer Science and Applications, 11(8), 492-502. https://www.academia.edu/download/64408036/Paper_62-Study_on_Dominant_Factor_for_Academic_Performance.pdf

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. Artificial Intelligence Review, 53(2), 907-948. https://doi.org/10.1007/s10462-019-09682-y

Tripathi. A..& Trivedi. S. K. (2016, 24-24 October). Sentiment analyis of Indian movie review with various feature selection techniques. In 2016 IEEE international conference on advances in computer applications (ICACA) (pp. 181-185). IEEE. Coimbatore

UCI Machine Learning Repository: Higher Education Students Performance Evaluation Dataset Data Set. https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset#

Velmurugan, T., & Anuradha, C. (2016). Performance evaluation of feature selection algorithms in educational data mining. Performance Evaluation, 5(02). https://www.researchgate.net/profile/Velmurugan-Thambusamy/publication/311773948_Performance_Evaluation_of_Feature_Selection_Algorithms_in_Educational_Data_Mining/links/585a381108ae3852d256dfb0/Performance-Evaluation-of-Feature-Selection-Algorithms-in-Educational-Data-Mining.pdf

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714.

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293. https://doi.org/10.1007/s00500-020-05297-6

Win. T. Z..& Kham. N. S. M. (2019). Information gain measured feature selection to reduce high dimensional data. In Seventeenth International Conference on Computer Applications (ICCA 2019) (Vol. 68. No. 73. pp. 1-5). https://meral.edu.mm/record/3413/files/ICCA%202019%20Proceedings%20Book-pages-79-84.pdf

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14, 1-37. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/content/pdf/10.1007/s10

115-007-0114-2.pdf&casa_token=Nx6o3BNiK_kAAAAA:6VTMnyDyNk9J_-wfg09CuJHptY6ERpasiIJfKIJ3weOKbWQNi5mWeJrC1_Yxcs0C3x6XMDx8OIi0b-i6sw

Yahdin, S., Desiani, A., Gofar, N., & Agustin, K. (2021). Application of the relief-f algorithm for feature selection in the prediction of the relevance education background with the graduate employment of the universitas sriwijaya. Computer Engineering and Applications Journal, 10(2), 71-80. https://doi.org/10.18495/comengapp.v10i2.369

Yılmaz, N., & Sekeroglu, B. (2019, August). Student performance classification using artificial intelligence techniques. In International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (pp. 596-603). Cham: Springer International Publishing.

Zaffar, M., Hashmani, M. A., Savita, K. S., & Rizvi, S. S. H. (2018). A study of feature selection algorithms for predicting students academic performance. International Journal of Advanced Computer Science and Applications, 9(5). https://www.researchgate.net/profile/Maryam-Zaffar/publication/325574028_A_Study_of_Feature_Selection_Algorithms_for_Predicting_Students_Academic_Performance/links/5b28ac4ba6fdcca0f09c62fa/A-Study-of-Feature-Selection-Algorithms-for-Predicting-Students-Academic-Performance.pdf

Zaffar, M., Hashmani, M. A., Savita, K. S., Rizvi, S. S. H., & Rehman, M. (2020). Role of FCBF feature selection in educational data mining. Mehran University Research Journal Of Engineering & Technology, 39(4), 772-778. https://search.informit.org/doi/pdf/10.3316/informit.459135063399758