Jurnal Teknik Informatika C.I.T Medicom, 16 (6) (2025) 334-341

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Optimization of academic performance prediction using linear regression with selectk-best

M. Rangga Ramadhan Saelan
Department Data Science, Nusa Mandiri University, East Jakarta, Indonesia

Article Info

Article history

Received : 28 Dec, 2024
Revised : Jan 28, 2025
Accepted : Jan 31, 2025

Kata Kunci:

Feature selection;
Linear Regression;
Prediction;
SelectK-Best;
Student performance.

Abstract

This study discusses the prediction of student performance by considering factors that can influence academic performance. In this research, the SelectK-Best feature selection technique and linear regression were used to enhance the accuracy of the prediction. The selection of this topic is based on the importance of understanding the factors that influence student performance and how feature selection can help build more efficient models. The methods applied in this study include data exploration through EDA, the use of SelectK-Best to select the most significant features, and linear regression to build the prediction model. The evaluation metrics show that the model with feature selection achieved MAE of 0.6293, MSE of 0.5945, RMSE of 0.7711, and R^2 Score of 0.9144, demonstrating the model's excellent performance. In contrast, the model without feature selection did not produce better results than the model with feature selection. This emphasizes the importance of applying feature selection techniques in building more accurate prediction models. This study contributes to predicting student performance through the use of systematic and effective methods, while also opening opportunities for further research in the context of education and more diverse data.

Corresponding Author:

M. Rangga Ramadhan Saelan,
Department Data Science
Nusa Mandiri University
Jl. Jatiwaringin Raya No. 2, Jakarta Timur, 13620, Indonesia
rangga.mgg@nusamandiri.ac.id

1. Introduction

Student performance is an important topic in the world of education, especially among students. Factors that influence student performance, such as lifestyle, study habits, and physical and mental health, have been widely studied[1],[2]. Student performance is influenced by various aspects, including demographic background, learning style, level of involvement in learning activities, and support from the family and school environment[3],[4],[5]. One factor that has not been widely explored is the effect of alcohol consumption on students' academic performance[6],[7]. Various previous studies have shown that alcohol consumption behavior can have a negative impact on students' learning ability and motivation as well as mental resilience[8],[9],[10]. However, there is still a lack of research that comprehensively integrates these factors to predict student performance, especially by utilizing more advanced analytical techniques.

According to research published in the Journal of Educational Technology by Toga Ari Harmawan and Lucia Sri Istiyowati[11], Big data analysis can identify significant factors that affect students' academic performance, such as discipline, learning methods, and personal note-taking

habits. A deep understanding of these factors can help design effective intervention strategies to improve learning outcomes.

Machine learning techniques, particularly feature selection using methods such as SelectK-Best, have proven effective in identifying key factors influencing student performance based on numerical data[12],[13],[14]. Application of machine learning methods in education, especially to predict student performance, has grown rapidly. Various algorithms such as support vector machines (SVM), linear regression and decision trees are widely used to analyze the factors that influence students' academic performance[15]. However, although many have used these techniques, few have focused on feature selection that can improve prediction accuracy[16]. SelectK-Best works by selecting a number of The most important features that have a significant impact on the prediction model, this is because there is often data that has too many features so that it needs to be selected into several features for work to be more efficient and effective, especially SelectK-Best is very suitable for use for numeric data [17],[18],[19]. This technique selects the most significant features to be used in a predictive model, based on statistical tests such as the Chi-Square test or ANOVA. This method is useful in improving the accuracy of predictive models by reducing data complexity and eliminating irrelevant or redundant features [20],[21]. In this context, the linear regression algorithm is used to build a simple yet effective prediction model, considering the numeric and continuous nature of the data, making the data more suitable for studying using a regression model[20],[21],[24]. CRISP-DM (Cross-Industry Standard Process for Data Mining) was chosen as the research methodology because of its structured and systematic approach in handling data mining projects, which includes understanding the problem, data preparation, model selection, evaluation, and application of results[25],[26].

In the context of scientific research, CRISP-DM offers a comprehensive framework for data analysis. For example, in a study conducted by Lukman and Nabilah[27], It describes how CRISP-DM is applied in data mining implementation, including phases such as business understanding, data understanding, and data preparation. This study shows that the CRISP-DM methodology can be used in various research fields to analyze data effectively.

However, although there are several studies that attempt to link alcohol consumption with students' academic performance, many of them have not integrated systematic feature selection to improve prediction accuracy. In addition, most studies are limited to analyzing data separately, without considering the combined effects of various factors that may play an important role. This study aims to fill this gap by using the SelectK-Best feature selection technique to select the main factors from alcohol consumption data that have the most influence on students' academic performance, then building a prediction model using linear regression.

The uniqueness of this study lies in the combination of several analytical techniques, namely feature selection, linear regression, and CRISP-DM, to produce a more accurate prediction model of factors that affect international student performance. This study also focuses on international students as the subjects of the study, who have unique challenges related to social integration, academic adaptation, and different lifestyles compared to domestic students. Therefore, this study not only contributes to the fields of data science and machine learning, but also provides new insights in the context of education.

2. Research Methodolgy

This research method is designed to ensure the reliability, validity, and reproducibility of the results. This research approach is quantitative based with exploratory and predictive methods. The purpose of this study is to analyze the factors that influence student performance using mechine leraning algorithms through the process of feature selection, model training, and performance evaluation.

The CRISP-DM (Cross-Industry Standard Process for Data Mining) method is a systematic and structured framework for conducting data analysis-based research, including in the context of education. This framework is designed to manage analytical projects iteratively, flexibly, and easily adaptable. In this study, CRISP-DM was chosen as the methodology because of its ability to handle complex, heterogeneous, and predictive-outcome-oriented data analysis processes[26].

1. Business Understanding

This study uses the SelectK-Best method in feature selection, which allows the identification of significant factors from various existing variables, including alcohol consumption, that affect academic performance. The selection of a linear regression model as a prediction tool is due to the numerical nature of the data. Based on the focus of the application of the 2 methods, this study allows for more accurate predictions of students performance based on the factors that have been selected.

2. Data Understanding

Data sourced from a public site originating from Portugal by P. Cortez and A. Silva [28], this public data has 33 attributes with a total of 649 respondent records.

Table 1. Data Description

Attribute	Description
school	student's ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	('F' - female or 'M' - male)
age	(from 15 to 22)
address	student's home address type ('U' - urban or 'R' - rural)
famsize	family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status ('T' - living together or 'A' - apart)
Medu	mother's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian ('mother', 'father' or 'other')
traveltime	home to school (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (n if 1<=n<3, else 4)
schoolsup	extra educational support (yes or no)
famsup	family educational support (yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (yes or no)
activities	extra-curricular (yes or no)
nursery	attended nursery school (yes or no)
higher	wants to take higher education (yes or no)
internet	Internet access at home (yes or no)
romantic	romantic relationship (yes or no)
famrel	family relationships (from 1 - very bad to 5 - excellent)
freetime	free time (from 1 - very low to 5 - very high)
goout	going out with friends (from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (from 1 - very low to 5 - very high)
health	current health status (from 1 - very bad to 5 - very good)
absences	number of school absences (from 0 to 93)
G1	first period grade (from 0 to 20)
G2	second period grade (from 0 to 20)
G3	final grade (from 0 to 20, output target)

3. Data Preparation

The first step taken is Exploratory Data Analysis (EDA), which aims to conduct an initial analysis of the dataset. This EDA process is important to thoroughly explore the data, find existing

patterns, and identify anomalies or outliers that may affect the quality of the prediction model[29]. During EDA, the relationship between variables, such as alcohol consumption and student academic performance, is analyzed to understand how each factor affects the target variable. Data visualizations, such as scatterplots, boxplots, and histograms, are used to help extract insights from the dataset more intuitively. In addition, this step also includes data cleaning, such as handling missing or duplicate values, so that the dataset is ready for the next analysis stage, namely feature selection and modeling. before forming the model, the categorical data is converted into numerical form to suit the needs of the model.

4. Modeling

In the modeling stage, this study uses linear regression techniques to build a prediction model of student academic performance based on factors that have been selected using SelectK-Best. After conducting the Exploratory Data Analysis (EDA) process and feature selection using the SelectK-Best method, the most significant variables in influencing academic performance are selected to be included in the model. Linear regression was chosen because of the numeric and continuous nature of the data, as well as its simplicity in describing the linear relationship between independent variables, such as alcohol consumption, and dependent variables, namely the final grade (G3) which is interpreted as Student Performance. This linear regression model is then trained using training data and tested using test data to evaluate its accuracy and predictive ability.

5. Evaluation

MAE, MSE, RMSE, and R^2 Score are evaluation metrics that are widely used to measure the performance of regression models. Using some of these metrics can provide a more comprehensive view of how well the model is making predictions. In the context of this study, The use of evaluation metrics MAE, MSE, RMSE, and R^2 Score is very suitable for viewing numeric data and continuous datasets[30].

6. Deployment

The linear regression model that has been built and evaluated is applied to provide predictions of international students' academic performance based on selected alcohol consumption datasets.

3. Result and Discuss

Student academic performance is often influenced by various external and internal factors, but not all factors have a significant impact. Therefore, this study uses SelectK-Best to select the most relevant features and applies linear regression to improve prediction accuracy.

The dataset used in this study focuses on factors that affect student performance including alcohol consumption among students and how this factor can affect their academic performance. To ensure that the data is ready to be used in modeling, several main steps are carried out, namely handling missing values, data normalization, and feature selection using SelectK-Best.

1. Initial Data Exploration (EDA - Exploratory Data Analysis)
   Exploratory analysis was conducted to understand the data distribution, identify initial patterns, and detect outliers or inconsistencies in the dataset. The exploration results showed that several features had a weak correlation with the target variable, which could impact model performance if not filtered properly.
2. Data Transformation and Normalization
   Further transformations are performed such as normalizing numeric values so that the linear regression model can work optimally. This normalization helps reduce unnecessary variability and ensures that each feature has a comparable scale in modeling. At this stage, data normalization is carried out by changing the type of data objects to "integer" and "Boolean" so that the model will be easier to learn the data later.

Normalization and data type conversion from object to boolean has strong scientific reasons, especially in improving model performance and ensuring that the algorithm can work well with the data used.

3. Feature Selection with SelectK-Best
   To improve model efficiency and accuracy, the SelectK-Best feature selection technique was applied, which aims to select a number of the best features based on statistical relationships with the target variable. The feature selection results showed that only a few factors had a significant influence on academic performance, while other features contributed little or even interfered with the accuracy of the prediction.

Analysis of alcohol consumption factors and several other factors on students' academic performance by applying feature selection using SelectK-Best and linear regression models, in this study data training was carried out using models with feature selection and without feature selection to observe the differences produced in predicting student performance which is interpreted into evaluation metrics.

The feature selection process is carried out to identify the most significant factors that affect academic performance, so that the model built only uses relevant features. In this stage, SelectK-Best successfully filters out insignificant features, increases model efficiency and reduces the possibility of overfitting. This study took 10 of the 32 selected features.

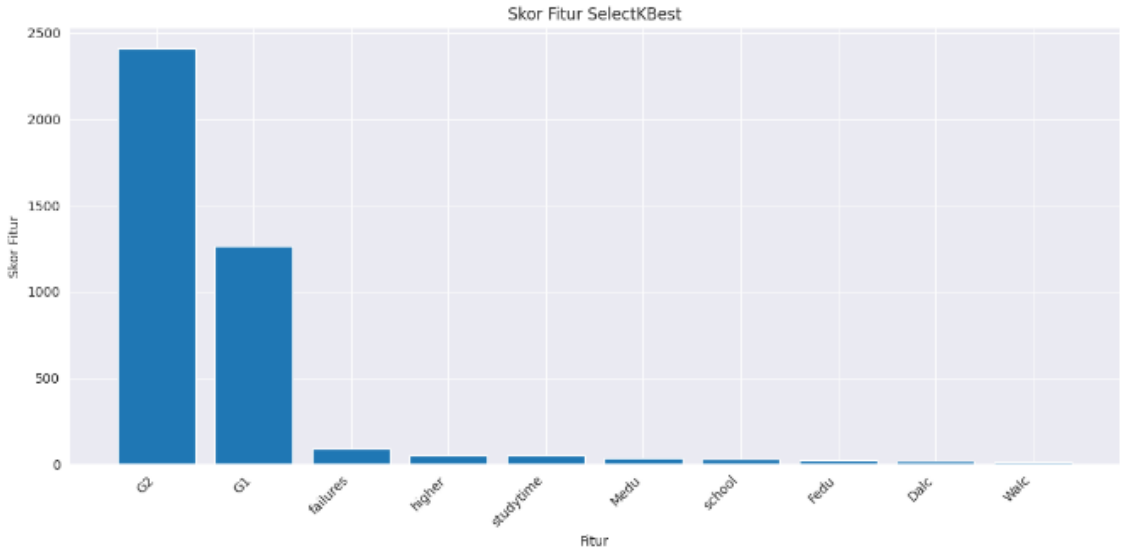

Figure 1. Top 10 Factors

After the feature selection process, a Linear Regression (LR) model was applied to predict academic performance based on the selected features. LR is chosen because it is able to describe the linear relationship between independent and dependent variables, which is very appropriate for the numerical data used in this study.

Model evaluation is carried out using several metrics, including MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R^2 Score (R-squared). This can provide a comprehensive picture of the model's performance in predicting academic achievement.

Table 2. Comparison of Evaluation Metrics of the use of selection features

Evaluation	Without Fiture Selection	With Fiture Selection
MAE	0.9553	0.6293
MSE	1.9507	0.5945

RMSE	1.3967	0.7711
R2 Score	0.8023	0.9144

Based on the evaluation results, the MAE and MSE values indicate that the application of SelectK-Best feature selection and linear regression models successfully minimized prediction errors relatively when compared to without feature selection. A low RMSE value means that the model tends to produce predictions that are close to the original values, with minimal large errors, namely MAE = 0.6293 and MSE = 0.5945.

Meanwhile, the high R^2 value with a value of 0.9144 indicates that the linear regression model can explain most of the variance in student academic performance influenced by several factors that have close relationship values. This confirms that the use of the SelectK-Best technique for feature selection and linear regression as a prediction model is quite effective in predicting student academic performance.

Overall, the results of this study show that the application of appropriate feature selection and the use of linear regression can produce a good model in predicting student academic performance. The evaluation metrics obtained support the conclusion that this model is quite accurate and reliable in describing the relationship between factors and academic performance.

4. Conclusion

This study aims to improve the accuracy of predicting students' academic performance by applying the SelectK-Best feature selection technique and linear regression models. The results of this study indicate that proper feature selection can simplify data complexity without sacrificing prediction quality, thus contributing to the development of more efficient prediction models in the field of educational analytics. By integrating feature selection methods, this study strengthens the understanding of significant factors affecting academic performance, while also offering a systematic approach to building more accurate models. This study also highlights the importance of further exploration of other feature selection techniques and comparison with more complex machine learning algorithms, especially in handling non-linear relationships in educational data. In the future, this study can be expanded by using more diverse datasets to improve model generalization, as well as adopting other feature selection techniques such as Recursive Feature Elimination (RFE) or deep learning-based methods to overcome the limitations of linear regression. In addition, the integration of this prediction model into an academic decision support system can be a further step in applying this research in a wider educational environment.

References

[1] Dyah Vierdiana, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI KESEHATAN MENTAL DI KALANGAN MAHASISWA PERGURUAN TINGGI," *Jurnal Review Pendidikan dan Pengajaran*, vol. 7, no. 1, pp. 1553–1558, 2024.

[2] Ulfah, "Pengaruh Kesehatan Mental Terhadap Prestasi Akademik Mahasiswa Tingkat Akhir," 2023.

[3] I. Widiasanti, J. T. Adelia, L. Rosidin, M. F. Viola, and M. Daniarista, "Implementasi Penggunaan Big Data Dalam Menganalisis Faktor Yang Mempengaruhi Kinerja Siswa Dalam Hasil Ujian," *Akademika*, vol. 12, no. 01, pp. 239–250, 2023.

[4] A. M. Br Peranginangin and N. Izzati, "Analisis Faktor-Faktor Yang Mempengaruhi Keaktifan Siswa Kelas VIII-1 SMPN 11 Tanjungpinang Dalam Pembelajaran Matematika," *MATH-EDU: Jurnal Ilmu Pendidikan Matematika*, vol. 8, no. 1, pp. 24–36, Apr. 2023, doi: 10.32938/jipm.8.1.2023.24-36.

[5] S. Asyura, S. Mulyani Jamil, and F. Ilmu Kesehatan, "Faktor-Faktor yang Mempengaruhi Prestasi Belajar pada Siswa di SMP Negeri 1 Baitussalam Kabupaten Aceh Besar Factors Affecting Student's Learning Achievement in Junior High School State 1 District Baitussalam Regency Aceh Besar," 2022.

[6] J. Ablian, M. Gantang, and A. Panes, "The Effects of Alcohol Consumption on Academic Performance: A Literature Review," vol. Volume 5, pp. 77–84, Feb. 2023.

[7] M. W. Manoppo, F. F. Pitoy, and K. B. Tampi, "Hubungan Tingkat Stres dengan Konsumsi Alkohol pada Remaja," *MAHESA : Malahayati Health Student Journal*, vol. 3, no. 6, pp. 1710–1725, Jun. 2023, doi: 10.33024/mahesa.v3i6.10585.

[8] S. Calnan and M. P. Davoren, "College students' perspectives on an alcohol prevention programme and student drinking–A focus group study," *Nordic studies on alcohol and drugs*, vol. 39, no. 3, pp. 301–321, 2022.

[9] R. Ajeng, "CONSUME ALCOHOL BEHAVIOR ON STUDENTS FACULTY OF SPORT SCIENCE STATE UNIVERSITY OF SURABAYA."

[10] Julia C. Pulumbara, Sekplin A.S. Sekeon, and Wulan P.J. Kaunang, "HUBUNGAN ANTARA KONSUMSI ALKOHOL DENGAN GANGGUAN FUNGSI KOGNITIF PADA PENDUDUK DI KELURAHAN TUMUMPA DUA KECAMATAN TUMINTING KOTA MANADO," 2018.

[11] T. A. Harmawan and L. S. Istiyowati, "Big Data Dan Pemahaman Faktor Penunjang Kinerja Akademik Siswa Untuk Meningkatkan Efektivitas Pembelajaran," *JKTP: Jurnal Kajian Teknologi Pendidikan*, vol. 7, no. 1, p. 035, Apr. 2024, doi: 10.17977/um038v7i12024p035.

[12] P. P. P. Thereza, G. Lumacad, and R. Catrambone, "Predicting Student Performance Using Feature Selection Algorithms for Deep Learning Models," in *2021 XVI Latin American Conference on Learning Technologies (LACLO)*, 2021, pp. 1–7. doi: 10.1109/LACLO54177.2021.00009.

[13] N. Kartik, R. Mahalakshmi, and K. A. Venkatesh, "Predicting Students' Performance Using Feature Selection-Based Machine Learning Technique," in *Proceedings of Data Analytics and Management*, A. Swaroop, Z. Polkowski, S. D. Correia, and B. Virdee, Eds., Singapore: Springer Nature Singapore, 2024, pp. 389–397.

[14] Monica Tiara Gunawan, Jeane Yosefa Tine, and Chatarina Enny Murwaningtyas, "Model decision tree untuk prediksi prestasi akademik matematika siswa kelas VIII SMP Frater Don Bosco Manado," *Jurnal Pendidikan Informatika dan Sains*, vol. 13, no. 2, 2024.

[15] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 4067721, Jan. 2024, doi: https://doi.org/10.1155/2024/4067721.

[16] R. Habibie Sukarna *et al.*, "ANALISIS PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN METODE ALGORITMA MACHINE LEARNING DAN FEATURE SELECTION," vol. 8, no. 2, 2024.

[17] N. R. Abid-Althaqafi and H. A. Alsalamah, "The Effect of Feature Selection on the Accuracy of X-Platform User Credibility Detection with Supervised Machine Learning," *Electronics (Switzerland)*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010205.

[18] M. Saelan and A. Subekti, "K-BEST SELECTION UNTUK MENINGKATKAN KINERJA ARTIFICIAL NEURAL NETWORK DALAM MEMPREDIKSI RANGE HARGA PONSEL," *INTI Nusa Mandiri*, vol. 19, pp. 10–16, Jul. 2024, doi: 10.33480/inti.v19i1.5554.

[19] Andy Hermawan, Nila Rusiardi Jayanti, Zia Tabaruk, Faizal Lutfi Yoga Triadi, Aji Saputra, and M.Rahmat Hidayat Syachrudin, "Membangun Model Prediksi Churn Pelanggan yang Akurat," *Merkurius : Jurnal Riset Sistem Informasi dan Teknik Informatika*, vol. 2, no. 6, pp. 67–81, Oct. 2024, doi: 10.61132/merkurius.v2i6.398.

[20] G. L. Pritalia, "Analisis Komparatif Algoritme Machine Learning pada Klasifikasi Kualitas Air Layak Minum," 2022.

[21] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, *A Chi-Square Statistics Based Feature Selection Method in Text Classification*. 2018. doi: 10.1109/ICSESS.2018.8663882.

[22] A. Qoiriah and Y. Yamasari, "Prediksi Nilai Akhir Mahasiswa dengan Metode Regresi (Studi Kasus Mata Kuliah Pemrograman Dasar)."

[23] M. A. Kurniawan, B. Very Christioko, M. R. Abidin, S. V. Rivaldo, and T. R. U. Utami, "Analisis Prediksi Kinerja Akademik Menggunakan Regresi Linear Berganda."

[24] D. Jatikusumo and R. R. Hidayat, "OPTIMASI PENENTUAN LOKASI BENCANA ALAM DENGAN REGRESI LINIER SEDERHANA DAN BERGANDA," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5257.

[25] D. Ruswanti, D. Susilo, and R. Riani, "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45," *Go Infotech: Jurnal Ilmiah STMIK AUB*, vol. 30, no. 1, pp. 111–121, Jun. 2024, doi: 10.36309/goi.v30i1.266.

[26] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[27] S. N. Luqman *et al.*, "Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM," 2021.

[28] P. Cortez and A. Silva, "Student Alcohol Consumption," Kaggle. Accessed: Jan. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

[29] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory Data Analysis," in *Secondary Analysis of Electronic Health Records*, M. I. T. C. Data, Ed., Cham: Springer International Publishing, 2016, pp. 185–203. doi: 10.1007/978-3-319-43742-2_15.

[30] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.