Analyzing Factors that Influence Student Performance in Academic

Nieta Hidayani[1,*], Deshinta Arrova Dewi[2], Tri Basuki Kurniawan[3]

[1]*Postgraduate Program in Syariah Banking, Universitas Azzahra, Indonesia*

[2]*Faculty of Data Science and Information Technology, Inti International University, Malaysia*

[3]*Magister Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia*

(Received: February 22, 2024; Revised: March 30, 2024; Accepted: May 3, 2024; Available online: May 31, 2024)

Abstract

Student performance analysis is a complex and popular study area in educational data mining. Multiple factors affect performance nonlinearly, making this topic more appealing to academics. The broad availability of education adds to this interest, particularly in online learning. Although previous studies have focused on analyzing and predicting students' performance based on their classroom activities, this study did not take into account student's outside conditions, such as sleep hours, extracurricular activities, and a sample of question papers that they had practiced. These three variables are included among others in our study. In this paper, we describe an analysis of 10,000 student records, each containing information on numerous predictors and a performance index. The dataset intends to shed light on the relationship between predictor variables and the performance indicator. To create the correlation variable heatmap, we use univariate and bivariate studies to produce a linear equation. The selection of these two techniques is based on their simplicity compared to other techniques. Besides, both are the most fundamental techniques in finding the data pattern. The bivariate analysis enables us to find the relation between two variables involved in the study. Finally, we showed the actual and expected student performance outcomes using the model we constructed. Following this, we perform data preprocessing and modeling to facilitate predictive analysis. The findings demonstrate that our prediction model was 98% accurate, with a mean absolute error of 1.62.

Keywords: Education, Education Quality, Education Data Mining, Predictive Analytics

1. Introduction

Student performance analysis and prediction are the two widely explored research topics in the education system literature [1]. Although their objectives are different, the outcomes of performance analysis significantly influence the prediction studies [2]. Statistical methods may not always be sufficient for establishing the association of various factors with performance [3]. The use of sophisticated algorithms may yield impressive knowledge that could help educators as well as students. The advancement of data mining technologies has influenced many researchers to investigate more profound insights into the knowledge dissemination process. Some of them have applied data mining approaches. However, the number of studies focusing on students' performance based on external factors was quite low [4]. The growing availability of digital data captured by several academic information management systems and educational software in recent years catalyzes this process to improve the study aiming for quality of education.

Since years ago, several studies in the past have reported similar works in analyzing and predicting the students' performance. For example, Osmanbegovic and Suljic conducted a study investigating students' future performance in the end-semester results at the University of Tuzla. They considered 11 factors and used a classification model with the highest accuracy for naive Bayes [5]. Apart from that, Suyal and Mohod applied the association and classification rule to identify the students' performance. They mainly focused on finding the students who need special attention to reduce the failure rate [6]. Another study by Noah, Barida, and Egerton conducted a study to evaluate students' performance by grouping the grading into various classes using CGPA. They used different methods like Neural networks, Regression, and K-means to identify the weak performers for performance improvement [7].

***Corresponding author: Nieta Hidayani (nietahidayani9@gmail.com)**
DOI: https://doi.org/10.47738/jads.v5i2.221

The usage of assignment marks was conducted by Baradwaj and Pal who described data mining techniques that help in the early identification of student dropouts and students who need special attention. Here they used a decision tree by using information like attendance, class tests, semester, and assignment marks [8]. Others like Bhise, Thorat, and Supekar presented a method using a K-means clustering algorithm by describing it step by step[5]. Their paper mainly focused on reducing the drop-out ratio of the students and improving it by considering evaluation factors like midterm and final exam assignments. They considered different clustering techniques namely hierarchical, partitions, and categorical.

In recent years, similar works have also been reported by Mengash who utilized data mining techniques and decision-making [7] while Hasan et.al., reported on the usage of video analytics along with data mining techniques [8]. Comprehensive reviews were also given by Khan [9] and Namoun [10]. The usage of the ensemble model selected was presented by Injadat [11].

Altogether, the studies above highlight the importance of analyzing academic students' performance. Predictions with high accuracy in student performance are deemed important because they aid in identifying students with low academic accomplishment at an early point in their education. In universities, student retention is linked to academic performance and enrollment systems.

Previous studies have suggested steps to assist low academic performance, for example:

1) Generation of the data source of predictive variables.
2) Identification of various features or factors that affect the performance of students' learning during their academic career.
3) Construction of a prediction model with the help of classification data mining techniques based on predictive.
4) Variables that are readily identified.
5) Validation of the model developed for universities with students' performance.

Therefore, generally data mining can be applied in the educational industry to improve our understanding of the learning process by discovering and analyzing variables [12]. Data mining approaches provide more personalized education, enhance system efficiency and lower educational process costs for universities. This enables us to boost student retention rates and academic achievements in the event of student learning outcomes. The data mining prediction technique aids in identifying the most effective elements influencing students' test scores and then fine-tuning these factors to improve student performance. It presents a fresh way to look at the education system that was hidden from humanity [13], [14]. On the other hand, the inclusion of external factors that may affect students' performance has been suggested by previous studies [15] for example that includes external factors like family support, family troubles, and health conditions. The usage of bivariate analysis on student performance was conducted by [16] which focuses on analyzing the mathematics and reading pupils' achievements. Additionally, the research to find the students' performance with the critical success factor using the bivariate analysis was also suggested by the [17].

Our paper primarily focuses on the following research questions.

1) What are the internal and external factors influencing student performance in learning?
2) Which methods are used for finding these factors?
3) Is it possible to predict before course commencement?

With the above research questions, we highlight two important things. First is the importance of the performance analysis that focuses on the factors that influence student performance. Second is the importance of prediction which aims to forecast future performance based on these factors. Both analyses contribute to the action of intervention in the classroom in assisting students who are facing difficulties in coping with the subjects. This kind of action can reduce the failure rate at an early stage.

ISSN 2723-6471

The structure of our paper is organized as follows: Section 2 elaborates on the methodology adopted for our study. In Section 3, we discuss the results found in our study. Finally, in section 4 we discuss the observations as a conclusion and recommendations for future work.

2. Methodology

Data mining is the knowledge discovery process from a huge data volume[18]. The mechanism works in large datasets and in this study, the student performance in the end-semester examination is evaluated. The similar works demonstrated by [19], [20]. In our studies, the data consists of demographic profiles and other important measured variables such as the hours studied, previous scores performance (if any), extracurricular activities, sleep hours, the sample of papers they have been practicing, a performance index, etc. The performance index assesses each student's total performance. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index varies from 10 to 100, with higher numbers indicating better performance.

2.1. Data Preparation

Student-related data were collected from the sampling in the computer science department of a private university in Southeast Asia during session 2022. The dataset includes the diploma and degree students of various programs i.e. science, engineering, economics, law, arts, and medical schools. The data ranges from the first year to the final year of study. The survey does not include the postgraduate program. In this step, data stored in different tables were joined into a single set. There were 10,000 pieces of data collected with an emphasis on the internal and external variables that potentially affect the students' performance. Due to the privacy and confidentiality to protect students' data, we practiced data anonymization, and all participants have informed consent accordingly.

2.2. Data Selection and Transformation

In this step, only those fields were selected which were required for the data mining process. We ensure that all data has no missing values, no complete duplication, and modification of relevant column data types is completed. Several steps have been taken towards this as follows.

1) Handling missing values with deletion of rows or columns and imputation using the mean average of the non-missing values to fill in missing values.

2) Outliers' detection using visual inspection such as mean, median, standard deviation, and quartiles to spot the central tendency and spread of data.

3) Removing duplicates by keeping only the first occurrence of the data.

4) Handling inconsistent data using data formatting techniques for date formats, capitalization, and conversion to maintain data uniformity.

2.3. Univariate Analysis and Visualization

In this step, we explore each variable in a data set, separately. We look at the range of values, as well as the central tendency (mean, mode, and median) of the values to describe the pattern of response to the variable. We also take into consideration dispersion of range, variance, maximum, minimum quartiles, and standard deviation. By doing this, we can screen data and evaluate whether the data meets our assumptions and criteria.

2.4. Bivariate Analysis and Visualization

In this step, we examine how two different things are related. We aim to determine if there is a statistical link between the two variables and, if so, how strong and in which direction that link is. By doing this, we can observe and investigate how two variables are connected and find trends and patterns in the data.

2.5. Data Processing and Modeling

In this step, we split the data into training and testing data sets with a ratio of 80:20. The purpose of this is to train and develop models. By doing this, we can estimate different variables to compare the model performance. We decided to employ the linear regression model to calculate the score of the model on the training data. The real and predicted values are generated at this step to distinguish the model performance. We aim to achieve a low mean absolute error to come up with a formula for the regression model.

3. Result and Discussion

In this section, we will show the results of the analysis that we have performed on the collected data.

3.1. Data Preparation

Figure 1 below shows the five most important variables that we analyze to find factors that influence academic students' performance and Figure 2 shows the total amount of data that we collected.

In [3]:
```
# see top 5 rows

data.head()
```

Out[3]:

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced
0	7	99	Yes	9	1
1	4	82	No	4	2
2	8	51	Yes	7	2
3	5	52	Yes	5	2
4	7	75	No	8	5

In [6]:
```
# see dimensions

data.shape
```

Out[6]:
```
(10000, 6)
```

Figure 1. The five most important variables in our study

Figure 2. The total amount of data collected.

The hours of studies in Figure 1 capture numerical values representing the hours spent by the respondents to study in the semester of 2022. The previous scores were the earlier performance done by the respondents. The extracurricular activities capture whether the student's names are registered in the enrolment. Sleep hours represent the sleep time average done by the respondents in the semester of 2022. The sample question papers practiced show the number of question papers practiced in the semester of 2022.

3.2. Data Selection and Transformation

The following figure 3 shows the results after cleaning all data so they do not contain missing values and duplication.

In [9]:
```
# see quick info of category values

data.describe(include = object)
```

Out[9]:

	Extracurricular Activities
count	10000
unique	2
top	No
freq	5052

Figure 3. The clean data as a result of data transformation

Using Python programming, we check the properties of each data variable for example in Figure 3 out of several variables that we observed, extracurricular activities are one of the variables that have 2 distinct values i.e. yes and no. In this column (and other columns), we identified the missing values and repetitive values that need to be removed. Apart from that, standardized data entry and formatting have taken place as well.

3.3. Univariate Analysis and Visualization

The following figures show the visualization of the univariate analysis results. Figure 4 illustrates the hours studied by the collected students' data. We can see that most students averagely studied in one hour during the 2022 semester. Figure 5 shows that half of the student population averagely scored above 60 whereas most of the population does not participate in extracurricular activities as shown in Figure 6. Figure 7 extracts the information that most students sleep for 8 hours per day. Figure 8 shows most students have a similar number of sample questions that mostly perhaps they study in a group.

ISSN 2723-6471

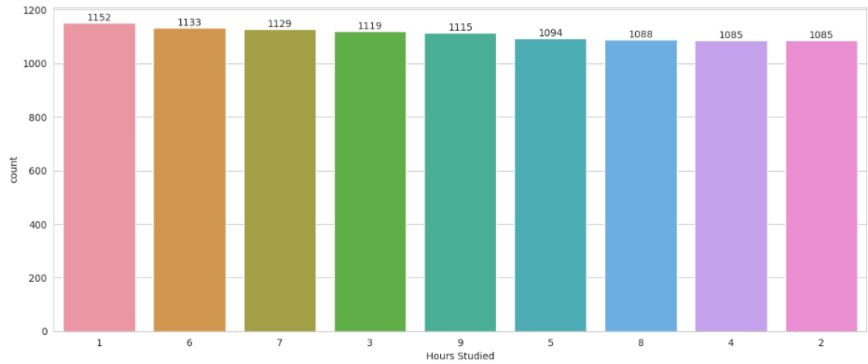

Figure 4. Average hours of student study in the 2022 semester

From the respondents, we can see the highest duration for the study is one hour which has been illustrated by Figure 4. Respectively, the next higher duration of the study is 6 hours 7 hours, 3 hours, 9 hours, 5 hours, and 8 hours. This indicates the duration of the study varies among respondents and potentially affects the performance.

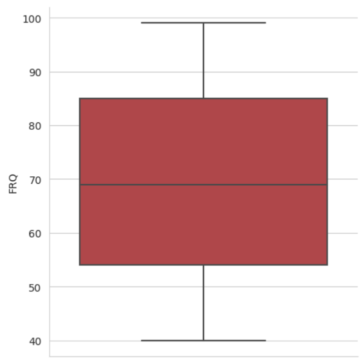

Figure 5. The average scores of students in the 2022 semester

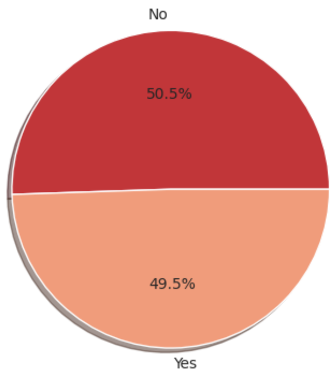

Figure 6. The extracurricular activities of students in the 2022 semester

The data collected in the 2022 semester shows the distribution of the scores are in the range of 40 and 100 whereby most of the respondents score from 55 to 85 which generally is a satisfactory performance. This is depicted in Figure 5. There were no outliers detected during the analysis that conclude the respondent's marks are alike which is most likely they used to learn together or frequently learn in a group of study.

However, the extracurriculars show a fair distribution of data showing the interest of the respondents in doing extracurricular activities is not high. This is shown in Figure 6. The involvement of extracurricular activities tends not to affect the students' performance. This is understandable since the extracurricular activities are voluntary.

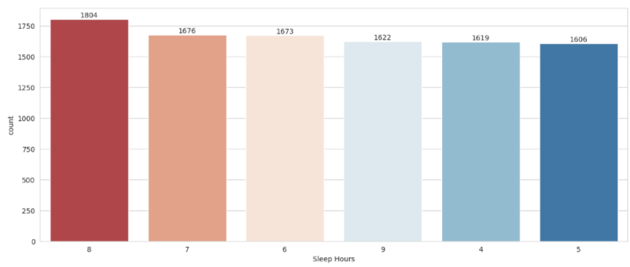

Figure 7. The average sleep hours of students in the 2022 semester

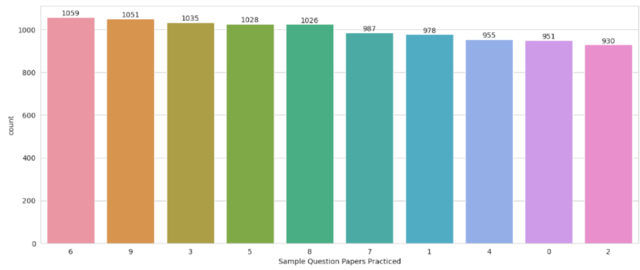

Figure8. The number of sample question papers studied in the 2022 semester

The average sleep hours of respondents in the survey are normal mostly they spend 8 hours a day for sleep. Respectively, the next sleep hours captured in the survey are 7 hours, 6 hours, 9 hours, and 4 hours. This can be seen in Figure 7.

Furthermore, the study in this paper includes the number of question papers studied in the 2022 semester as depicted in Figure 8. The highest count captured in the survey was 6 papers, 9 papers, 3 papers, and 5 papers. This variation indicates this variable may affect the performance of students.

3.4. Bivariate Analysis and Visualization

The following figures show the visualization of the bivariate analysis results to understand how two different variables are related. Figures 9 to 13 are generated from the collected data with 6 main features in the research i.e., Hours studied, previous score, sleep hours, sample question papers practiced, and performance index as the main target.

Figure 9 shows the statistical evidence of variables that we have observed. We can see that all main features have 10.000 data without any missing values. Accordingly, the mean for hours of study is 4.99 with the average previous score is 69.44. The mean of sleep hours is 6.5 hours with 4.5 samples of question papers practiced. The mean of the performance index is 55.22. These values are quite reasonable indicating that the 6 main features may affect the students' performance in academic.

Figure 10 illustrates the correlation of hours of study with the performance index or success rate. The graphs suggest the more hours students studied, the greater their success rate, hence the students are encouraged to spend more hours studying to achieve better performance.

Figure 11 shows the relationship between extracurricular activities with the performance index or higher grades. The study suggests that students who participate in extracurricular activities help to a very small extent in obtaining high grades. This implies that extracurricular activities are voluntary and based on students' preferences because this feature is not dominant to the other features.

The association between the performance index and the number of hours of sleep is shown in Figure 12. According to the graph, students' performance index may rise the more hours they sleep. Therefore, it is advised that students engage in constructive activities that support their performance improvement and remember to get enough sleep.

Figure 13 shows the heatmap of correlation that reflects there are strong positive relationships between the performance index and features on the map. In sequential order, the heatmap suggests variables that possess high correlations with the performance index. They start with previous scores (0.92), followed by hours of studies (0.37), sleep hours (0.05), and hours of sleep (0.04). These facts are strengthened by the findings as shown in figures 10 and 12.

```
# display quick info of numeric values of variables observed

data.describe()
```

	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Performance Index
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	4.992900	69.445700	6.530600	4.583300	55.224800
std	2.589309	17.343152	1.695863	2.867348	19.212558
min	1.000000	40.000000	4.000000	0.000000	10.000000
25%	3.000000	54.000000	5.000000	2.000000	40.000000
50%	5.000000	69.000000	7.000000	5.000000	55.000000
75%	7.000000	85.000000	8.000000	7.000000	71.000000
max	9.000000	99.000000	9.000000	9.000000	100.000000

Figure 9. The statistical evidence of the observed variables

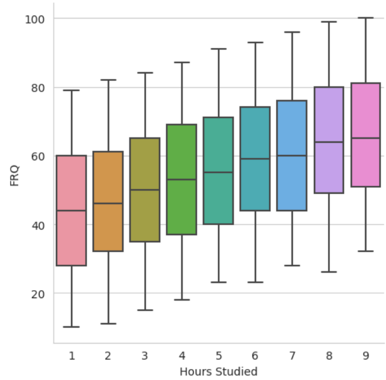

Figure 10. Bivariate analysis of hours studies and success rate

ISSN 2723-6471

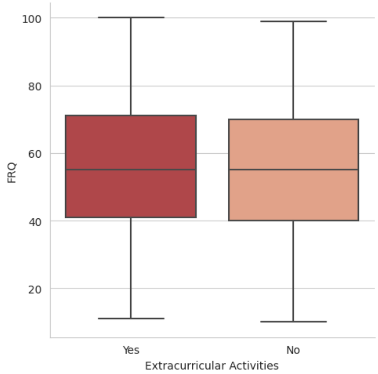

Figure 11. Bivariate analysis of extracurricular activities and high grades

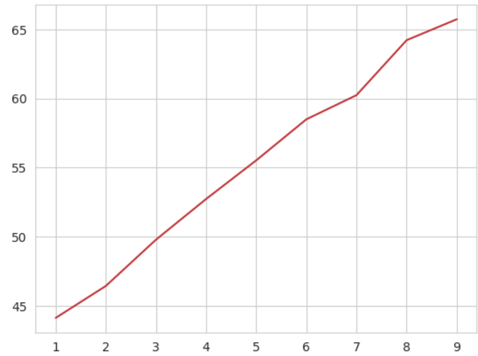

Figure 12. Bivariate analysis of hours of sleep and performance index.

Figure 13. Heatmap of correlation among variables

3.5. Data Processing and Modeling

This section presents the actual and predicted values as results of the previous steps. We can see that our proposed model has produced closer values to the actual performance. Figure 14 shows the samples of the actual and predicted performance. Figure 15 shows the distribution of values generated from data in Figure 14. We have calculated the mean absolute error of this model, and it shows 1.612 which is quite low indicating a good quality of our model. The accuracy of our model is 98.89%. We produced the coefficient values of 2.75, 0.17, 1.02, 0.51, and 0.38. The model intercept is 30.92. Hence, our model has suggested the equation of the multiple linear regression model as below and can be used as a predictive model before class commencement.

(2.75 × Hours Studied) + (0.17 ×Sample Question Papers Practiced) + (1.02 × Previous Scores) + (0.51 × Extracurricular Activities) + (0.38 × Sleep Hours) - 30.92 (model intercept)

	Actual Performance	Predicted Performance
6252	51.0	54.7
4684	20.0	22.6
1731	46.0	47.9
4742	28.0	31.3
4521	41.0	43.0
...	...	...
6412	45.0	46.9
8285	66.0	62.7
7853	16.0	16.8
1095	65.0	63.3
6929	47.0	45.9

Figure 14. The actual and predicted students' performance

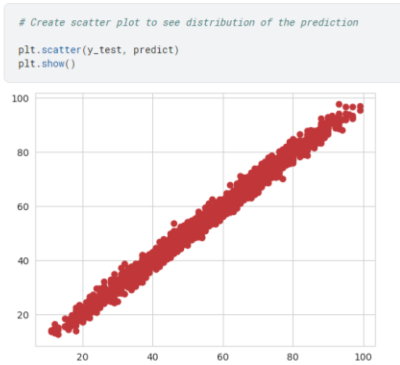

Figure 15. The scatter plot of distributed values

3.6. Model Evaluation

Handling overfitting is essential in our research especially since our research is more of a linear regression. Several methods we applied during the experiments by improving the quality of the training data to ensure meaningful patterns were captured during the data analysis. Besides, we ensured the model did not contain high variance data. Furthermore, we also use Lasso regression to maintain model reliability on both the training and testing datasets by tuning the alpha values. The following Figure 16 shows the values of the mean absolute error (1.612) and R^2 score (0.98.89).

```
# see mean absolute error

mean_absolute_error(y_test,predict)
```

```
1.612
```

```
# see score

r2_score(y_test,predict)
```

```
0.9889704960519785
```

Figure 16. The MSE and R^2 score of the data

4. Conclusion

In this paper, the univariate and bivariate analyses are performed on 10,000 students' performance data in the 2022 semester with the visualization of the important variables. The main contribution of our research is the employment of the external variable that promotes the predictive modeling that we suggested. This study can facilitate researchers and education practitioners in exploring the factors that may influence students' academic performance. The factors may vary in different semesters across schools however the techniques that we use in the study can be an example of doing such research. Future works may include applying data processing techniques using different variables in another program study with additional information such as parental involvement, socio-economic status, etc. Apart from that, this research can be expanded using different data mining techniques such as random forest, neural network, clustering, ensemble methods, etc. With this approach, we may get the overall overview of factors influencing students' performance in academic that support the corrective or intervention actions that help the students to perform better in earlier stages. The findings of the study may support the learning strategy and actions that will be taken by educators and policymakers.

5. Declarations

5.1. Author Contributions

Conceptualization: D.A.D., and T.B.K.; Methodology: D.A.D.; Software: T.B.K; Validation: N.H., D.A.D., and T.B.K.; Formal Analysis: N.H., D.A.D., and T.B.K.; Investigation: N.H.; Resources: D.A.D.; Data Curation: N.H.; Writing Original Draft Preparation: N.H. and D.A.D.; Writing Review and Editing: N.H. and D.A.D.; Visualization: D.A.D.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] O. Edin, "Data Mining Approach for Predicting Student Performance Economic," *Econ. Rev. J. Econ. …*, vol. X, no. 1, pp. 3–12, 2012.

[2] M. M. M. Sayali Rajesh Suyal, "Quality Improvisation Of Student Performance Using Data Mining Techniques," *Int. J. Sci. Res. Publ.*, vol. 4, no. September 2014, pp. 115–123, 2015.

[3] O. F. Noah, B. Barida, and T. O. Egerton, "Evaluation of Student Performance Using Data Mining Over a Given Data Space," *Nigeria*, vol. 1, no. 4, pp. 101–104, 2013.

[4] S. P. Baradwa, Brijesh Kumar, "Mining Educational Data to Analyze Students‟ Performance," *Int. J. Immunopathol. Pharmacol.*, vol. 15, no. 1, pp. 59–63, 2002, doi: 10.1177/039463200201500108.

[5] B. R.B, "'Importance of Data Mining in Higher Education System,'" *IOSR J. Humanit. Soc. Sci.*, vol. 6, no. 6, pp. 18–21, 2013, doi: 10.9790/0837-0661821.

[6] M. Kumar, A. J. Singh, and D. Handa, "Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques," *Int. J. Educ. Manag. Eng.*, vol. 7, no. 6, pp. 40–49, 2017, doi: 10.5815/ijeme.2017.06.05.

[7] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, no. 1, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[8] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Appl. Sci.*, vol. 10, no. 11, pp. 1-9, 2020, doi: 10.3390/app10113894.

[9] A. Khan and S. K. Ghosh, *Student performance analysis and prediction in classroom learning: A review of educational data mining studies*, vol. 26, no. 1. Education and Information Technologies, pp. 1-7, 2021. doi: 10.1007/s10639-020-10230-3.

[10] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, pp. 1–28, 2021, doi: 10.3390/app11010237.

[11] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Syst.*, vol. 200, no. 1, pp. 1-12, 2020, doi: 10.1016/j.knosys.2020.105992.

[12] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’ performance prediction using machine learning techniques,” *Educ. Sci.*, vol. 11, no. 9, pp. 1-9, 2021, doi: 10.3390/educsci11090552.

[13] I. Papadogiannis, V. Poulopoulos, and M. Wallace, “A Critical Review of Data Mining for Education: AR TI CL E IN FO AB STR A CT,” *Int. J. Educ. Res. Rev.*, vol. 1, no. 1, pp. 353–372, 2020, [Online]. Available: www.ijere.com

[14] T. T. Ting *et al.*, “Educational big data mining: Mediation of academic performance in crime among digital age young adults,” *Online J. Commun. Media Technol.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.30935/ojcmt/14026.

[15] M. Carnoy and S. Loeb, “Does external accountability affect student outcomes? A cross-state analysis,” *Educ. Eval. Policy Anal.*, vol. 24, no. 4, pp. 305–331, 2002, doi: 10.3102/01623737024004305.

[16] R. Krakehl, A. M. Kelly, K. Sheppard, and M. Palermo, “Physics teacher isolation, contextual characteristics, and student performance,” *Phys. Rev. Phys. Educ. Res.*, vol. 16, no. 2, p. 20117, 2020, doi: 10.1103/PhysRevPhysEducRes.16.020117.

[17] I. Damjanov, B. A. Fenderson, M. Hojat, and E. Rubin, “Curricular reform may improve students’ performance on externally administered comprehensive examinations,” *Croat. Med. J.*, vol. 46, no. 3, pp. 443–448, 2005.

[18] M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” *Smart Learn. Environ.*, vol. 9, no. 1, pp. 1-11, 2022, doi: 10.1186/s40561-022-00192-z.

[19] Q. Meng and W. Jia, “Influence of psychological hardiness on academic achievement of university students: The mediating effect of academic engagement,” *Work*, vol. 74, no. 4, pp. 1515–1525, 2023, doi: 10.3233/WOR-211358.

[20] M. Theobald, “Self-regulated learning training programs enhance university students’ academic performance, self-regulated learning strategies, and motivation: A meta-analysis,” *Contemp. Educ. Psychol.*, vol. 66, no. May 2021, pp. 1-9, 2021, doi: 10.1016/j.cedpsych.2021.101976.