Machine Learning in Finalizing Grades of Students' Performance in Distance Learning

Marat Nurtas[1], Temirlan Otepov[1], Aizhan Altaibek[1], Kateryna Kolesnikova[1] and Konstantin Borodkin[1]

[1] *International Information Technology University, Manas St. 34/1, Almaty, 050040, Kazakhstan*

Abstract

Nowadays, Machine learning (ML) in education is one of the less investigated areas of Data Science. However, the power of using ML in this is almost unlimited. As an example, ML in educational technology can be used for grading or testing students, improving student retention, and predicting student performance. During the pandemic, most students made the transition to distance learning, resulting in a substantial increase in the grades of select students during this period. However, as the pandemic situation resolved, these students reverted to their prior average grades. The first aim of this research is to demonstrate the predictive capacity of machine learning in forecasting students' final grades; the second aim is to examine the correlation between students' performance during the online learning phase and their final grades, while the third aim is to compare this correlation with that observed during offline periods.

Keywords

Machine learning, linear regression, educational technology, statistical learning

1. Introduction

In the realm of education, the integration of ML stands as a promising frontier within the expansive landscape of Data Science. Although this domain remains relatively underexplored, its potential to revolutionize educational practices is virtually boundless [1]. ML, when harnessed in educational technology, unveils a diverse array of applications, ranging from the automated grading of assessments to the enhancement of student retention strategies and the precise prediction of student academic performance [2,3].

Between 2007 and 2014, only a limited number of publications, approximately 4-5, focused on the theme of machine learning in education. However, from 2015 to 2017, there was a noticeable surge in interest, with as many as 20 works being published in 2017[4]. This suggests a growing enthusiasm for the application of machine learning in the educational sector during that time period. The increase in publications reflects the expanding recognition of the potential advantages that machine learning can bring to education, including personalized learning, adaptive assessment, and data-driven insights for educators and students.

The central focus of this paper lies in harnessing the formidable capabilities of ML to predict the final grades of students accurately. As the world grappled with the unforeseen challenges posed by the COVID-19 pandemic, the educational landscape underwent a seismic shift. The widespread adoption of distance learning became a necessity, catalyzing a profound transformation in the way students engaged with their academic pursuits [5,6].

One intriguing phenomenon that emerged during this period of remote learning was the dramatic increase in the grades of certain students. As classrooms transcended physical boundaries and traditional assessment methods, select students experienced a notable surge in their academic performance [7]. However, as the pandemic's grip on the world gradually

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

m.nurtas@iitu.edu.kz (M. Nurtas); oneplusone113@gmail.com (T. Otepov); a.altaibek@iitu.edu.kz (A. Altaibek); kkolesnikova@iitu.edu.kz (K. Kolesnikova); kongreat13@gmail.com (K.. Borodkin)

0000-0003-4351-0185 (M. Nurtas); 0000-0001-8431-7950 (A. Altaibek); 0000-0002-9160-5982 (K. Kolesnikova)

loosened, an equally remarkable reversion occurred, as these high-achieving students reverted to their prior average grades [8].

The main challenges in combating cheating during distance learning encompass obtaining assessment answers in advance, unfair retaking of assessments, and unauthorized assistance during assessments, necessitating measures such as proctoring systems, plagiarism detection, and promoting academic integrity to address these issues [9].

Considering these developments, the primary objective of this research endeavors to delve deep into the dynamics of student academic performance during the online learning phase. Specifically, the aim of the project is to unravel the underlying correlation between students' final grades and their performance during the unique and transformative period of online education.

The significance of this inquiry extends far beyond its immediate scope. It represents a critical exploration at the intersection of technology, education, and the human learning experience. By scrutinizing the academic trajectories of students during a time of unparalleled change, we seek to unearth valuable insights that can inform educational strategies, interventions, and policies for a post-pandemic world [10].

2. The problem statement

In the rapidly evolving landscape of education, especially in the context of the COVID-19 pandemic, it has become imperative to examine the true impact of online learning on a student's final academic achievement [11]. This study endeavors to address the following critical question: To what extent do grades obtained during online learning phases influence a student's ultimate final grade, when juxtaposed with the grades achieved during traditional offline periods?

The emergence of online education has reshaped the educational experience, with students adapting to digital classrooms during certain quarters of their academic journey [12]. Considering this, this research project seeks to delve deep into the intricate relationship between online and offline grades and their role in predicting a student's final academic performance.

Addressing this research problem involves utilizing the tool of Linear Regression, with its formula defined as follows:

$$Final\ grade = \Sigma k_i \cdot grade_i,$$

where k_i denotes the efficiency coefficient for grades during online and offline quarters, and $grade_i$ represents the grades garnered by students in these specific educational phases.

As per the formula, the sum of the coefficients should indeed equal 1, which is a fundamental property in linear functions to maintain proportionality. However, it's important to note that the constant coefficient (often referred to as the intercept) represents the baseline or starting point of the function and can be critical in understanding the behavior of the function. In linear regression or linear models, this constant term provides valuable information about the relationship between the variables [13].

In the context of comparing online and offline education, if the constant coefficient has a significant value, it can signify the difference in the starting points or base levels of the two types of education. For instance, it might indicate that even when all other coefficients are equal, there's a baseline difference in outcomes between online and offline education.

So, the constant coefficient can matter in understanding the relationship between coefficients in the context of online and offline education, especially if it shows how the two methods differ in their starting points or baseline performance.

In essence, this constraint helps maintain the linearity of the model, ensuring that the predicted values fall within an interpretable and meaningful range while accurately representing the relationship between the predictor variables and the outcome.

This inquiry serves not only to enhance our understanding of the dynamics between online and offline learning but also to offer valuable insights to educational institutions and policymakers regarding the efficacy of online education and its enduring influence on a student's

academic journey. Ultimately, the findings of this research will provide a comprehensive understanding of the complex interplay between online and offline grades and their genuine impact on a student's final academic success.

3. Methods and research

Considering the changing educational landscape, particularly during the COVID-19 pandemic, the research project explored the application of ML in the field of education, an area within Data Science that has received relatively less attention [14]. The primary objective was to harness the potential of ML, with a specific emphasis on the Linear Regression method, to predict the final grades of high school students.

Linear regression analysis is employed to forecast the value of one variable based on another variable's value. The variable being predicted is referred to as the dependent variable, while the variable used to make the prediction is known as the independent variable [15].

The central goal of the research was to create and employ ML tools capable of accurately projecting the students' final grades. The chosen approach involved the use of the Linear Regression method, a powerful statistical technique commonly applied in predictive modeling [16]. By employing this method, the intention was to construct predictive models that could estimate the students' forthcoming final grades based on various pertinent factors, including their performance during the online learning phase, historical academic data, and potentially other variables.

The rationale behind this endeavor was to offer educational institutions and stakeholders valuable insights into the potential of ML for enhancing the educational experience. Specifically, the research sought to investigate whether the students' performance during the unique online learning phase had a lasting impact on their overall academic achievement, including their final grades.

Subsequently, upon the creation of these predictive models, a comparison was made between the projected final grades and the students' actual grades following examinations. This comparison enabled an evaluation of the efficacy of the ML approach in accurately forecasting academic outcomes.

The research project represents an examination of the intersection between data science and education, showcasing how advanced analytical techniques like ML can contribute to a deeper comprehension of student performance, particularly within rapidly evolving educational contexts. By shedding light on the connection between online learning experiences and final grades, the aim is to provide valuable insights that can inform educational strategies and interventions, benefiting students in Kazakhstan and beyond.

4. Data collection and analysis

The study involved a group of 140 students in Kazakhstan who progressed from the 10th to the 12th grade. The statement that a dataset with 10 columns (features) should ideally have at least 100 rows for optimal results is a general guideline rather than a strict rule [17]. The relationship between the number of features and the number of observations in a dataset can depend on various factors, including the complexity of the data, the nature of the features, and the machine learning algorithm being used. It is noteworthy that this student cohort experienced a unique educational scenario during their 11th-grade year, and the 3rd and 4th quarters of their 10th-grade year, characterized by a shift to online learning due to the pandemic.

Table 1
Grades of first 10 students during 10th grade 1,2,3,4 quarters

№	Name	10(1)	10(2)	10(3)	10(4)
1	student 1	54	49	89	100
2	student 2	91	92	89	100
3	student 3	90	78	93	98
4	student 4	99	86	98	100
5	student 5	67	71	91	100
6	student 6	70	40	94	100
7	student 7	74	76	93	90
8	student 8	93	89	99	100
9	student 9	89	56	93	100
10	student 10	90	70	83	90

Table 1 serves as a comprehensive visual representation, offering a detailed insight into the academic performance of students throughout four distinct quarters. What makes this data particularly intriguing is the backdrop against which it unfolds—the transition of these students from the conventional offline mode of education to an entirely online learning environment during their 10th-grade year.

This transition marked a profound shift in the educational landscape for these students, one that brought about notable transformations in their academic achievements. One of the most striking observations within this dataset is the remarkable surge in students' scores, notably from the 2nd to the 3rd quarter. Such an increase is not only noteworthy but also warrants a deeper examination.

To uncover the underlying factors contributing to this surge, it's crucial to turn our attention to the change in academic conduct. During the transition to online education, there was a significant uptick in the prevalence of academic misconduct, with instances of cheating increasing from a relatively modest 30% to a staggering 60% [18]. This shift in academic integrity levels raises questions about the complex interplay between learning environments and academic outcomes.

Furthermore, this data not only paints a picture of the evolving educational landscape but also highlights the challenges and opportunities that emerged during the shift to online learning. It underscores the need for educators, institutions, and policymakers to navigate the intricacies of these changes effectively, promoting a learning environment that fosters academic excellence while addressing the challenges associated with academic misconduct.

In the research study, the dataset was partitioned into two distinct sets: a training set consisting of 70% of the data and a test set comprising the remaining 30%. Notably, a subset of 10 students, as originally presented in Table 1, was selected to showcase the outcomes. Specifically, the study aimed to demonstrate and compare the predicted scores with the actual scores these students achieved in their final assessments.

5. Results and future works

In Figure 1, we can observe how linear regression works with one variable. If we halt to examine the efficiency coefficients for grades during online and offline quarters, represented as k_i, within the previously mentioned formula:

$$Final\ grade = \Sigma k_i \cdot grade_i$$

in our case, there are 12 variables, representing four quarters of 10th grade, four quarters of 11th grade, and four quarters of 12th grade. This results in a 13-dimensional space to fully represent the data, which is challenging to visualize graphically. However, it is possible to calculate data

metrics and perform analysis to grasp the connections and patterns within this complex dataset, all without the necessity of generating a visual depiction.

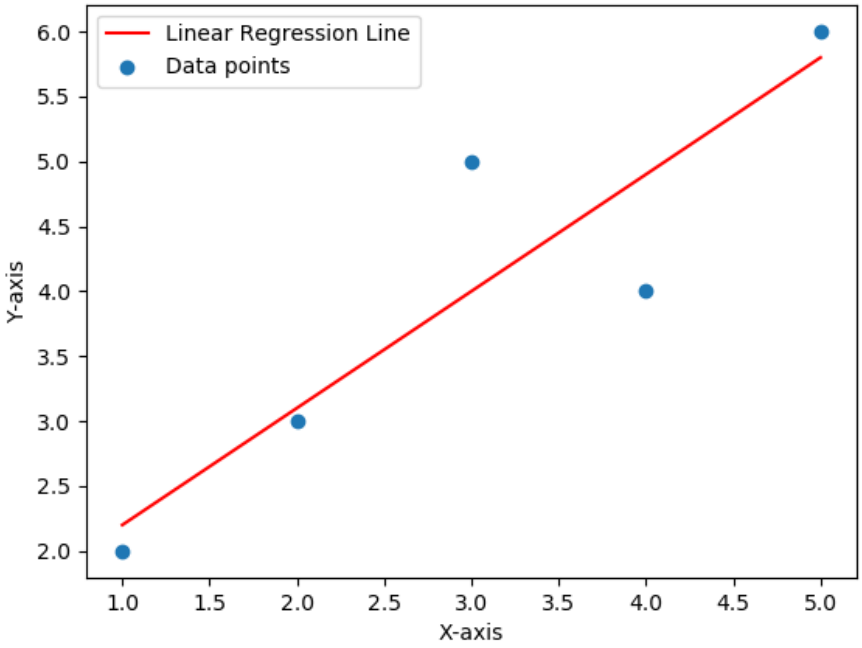

Figure 1: Linear regression in with one variable (generated by Python)

The model was trained using the Python programming language, specifically with the TensorFlow library developed by Google. TensorFlow is a widely used open-source machine learning framework that is particularly known for its capabilities in building and training deep neural networks.

In Figure 2, the depicted content is the presented code. This code likely corresponds to a specific section of the research or document, and its inclusion in the figure serves to visually illustrate or provide reference to a particular code snippet or algorithm discussed in the text.

```
model = tf.keras.Sequential([
    tf.keras.layers.Dense(1, input_dim=number_of_grades, activation='linear')
])
# Compile the model
model.compile(optimizer='sgd', loss='mean_squared_error')
# Train the model
model.fit(X_train, y_train, epochs=50, verbose=1)
# Get the coefficients (weights) and intercept (bias)
weights = model.layers[0].get_weights()[0]
bias = model.layers[0].get_weights()[1]
print("Coefficients (Weights):", weights)
print("Intercept (Bias):", bias)
```

Figure 2: Algorithm Implementation in Python

Here, two variables are printed: efficiency coefficients and bias. Efficiency coefficients, as denoted previously as k_i are the values that need to be determined and are now being printed or

displayed. It should be clarified that there were 5 quarters of online education and 7 quarters of offline education, resulting in 5 efficiency coefficients for online education and 7 efficiency coefficients for offline education.

Bias in the context of machine learning is often characterized as a systematic error stemming from incorrect assumptions made during the model's training process. More technically, bias can be defined as the discrepancy between the average predictions made by the model and the actual ground truth values. This discrepancy indicates the presence of systematic inaccuracies in the model's output, which can lead to deviations from the true values it is intended to predict. Addressing bias is a critical aspect of improving the overall performance and accuracy of machine learning models [19].

The predicted grades of students are shown in Table 2, and subsequently predicted grades are translated into letter grades. Grade ranges that correspond to each letter grade:

- A: 90-100;
- B: 80-89;
- C: 70-79;
- D: 60-69;
- E: Below 60.

Table 2
Predicted grades of first 10 students with Linear regression

№	Name	Predicted grade	Letter grade
1	student 1	75,608	C
2	student 2	90,24	A
3	student 3	87,248	B
4	student 4	92,792	A
5	student 5	82,776	B
6	student 6	82,44	B
7	student 7	80,704	B
8	student 8	91,824	A
9	student 9	81,456	B
10	student 10	80,688	B

To assess the performance of the regression model, the Relative Root Squared Error (RRSE) was computed using the following formula:

$$RRSE = \sqrt{\frac{\Sigma(y_i - y_i^p)^2}{\Sigma(y_i - \overline{y})^2}},$$

where:

- y_i - real grades of students;
- y_i^p - predicted grades of students;
- $\overline{y}$ - average real grades of students [20].

Following the calculation, it was determined that the RRSE value approached approximately 0.11, and the average bias was equal to 0.022. This result suggests that the regression model exhibits a favorable fit, indicating its effectiveness in predicting the desired outcomes [21].

For online learning quarters, the average efficiency coefficient is 0.072, whereas for offline learning, it stands at 0.16. This signifies that, when comparing these two coefficients, grades obtained during online education exhibit an approximately 2.2 times lower impact than those acquired during offline education.

In future research endeavors, the plan is to delve deeper into the realm of ML, specifically focusing on the powerful classification method, to uncover novel insights and perspectives [2].

This approach will enable a generation of distinct sets of results, further enriching the understanding of the educational landscape.

Moreover, forthcoming investigations will introduce an additional layer of complexity by considering the teacher's level of expertise as a crucial influencing factor. Within the school context from which the data was collected, teachers span a spectrum from novice to seasoned professionals, with six distinct levels of expertise. This intricate parameter is known to exert a profound influence on students' academic performance, and as such, it warrants comprehensive exploration.

6. Conclusion

In conclusion, the analysis of the efficiency coefficients and bias in the context of machine learning has provided valuable insights into the predictive capabilities of the regression model employed in forecasting students' grades. The distinct sets of efficiency coefficients for online and offline education underscore the differential impact of these learning modes on academic outcomes, with online education exhibiting approximately 2.2 times lower influence compared to offline education.

The assessment of predicted grades, translated into letter grades, through the Linear regression model, as presented in Table 2, further reinforces the model's effectiveness in capturing and predicting students' academic performance. The Relative Root Squared Error (RRSE) calculation, yielding a value of approximately 0.11, along with an average bias of 0.022, signifies a favorable fit and accuracy of the model in predicting the desired outcomes.

Looking ahead, future research endeavors are poised to delve deeper into the realm of machine learning, particularly focusing on powerful classification methods, to uncover novel insights and perspectives within the educational landscape. The inclusion of the teacher's level of expertise as a crucial influencing factor in forthcoming investigations adds an additional layer of complexity, acknowledging its profound impact on students' academic performance. By exploring these nuanced aspects, future research aims to contribute further to our understanding of the multifaceted dynamics influencing educational outcomes and inform potential improvements in educational practices.

7. References

[1] Kucak, D., Juricic, V., & Dambic, G. (2018). Machine Learning in Education - a Survey of Current Research Trends. In B. Katalinic (Ed.), Proceedings of the 29th DAAAM International Symposium (pp. 0406-0410). Published by DAAAM International. ISBN 978-3-902734-20-4, ISSN 1726-9679, Vienna, Austria, doi: 10.2507/29th.daaam.proceedings.059.

[2] Stimpson, A. J., & Cummings, M. L. (2014). Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. IEEE Access, 2, 78–87. doi:10.1109/access.2014.2303071.

[3] Nafea, I. T. (2018). Machine learning in educational technology. Machine learning-advanced techniques and emerging applications, 175-183.

[4] Korkmaz, C., & Correia, A. P. (2019). A review of research on machine learning in educational technology. Educational Media International, 56(3), 250-267.

[5] Figaredo, D. D., Jaurena, I. G., & Encina, J. M. (2022). The impact of rapid adoption of online assessment on students' performance and perceptions: Evidence from a distance learning university. Electronic Journal of e-Learning, 20(3), pp224-241.

[6] Bashkireva, T., Bashkireva, A., Morozov, A., Severin, A., Fateeva, N., Baykova, L., & Severina, E. (2022). Adaptation of students to distance learning in COVID-19 conditions in terms of ultradian rhythms of the cardiovascular system. In E3S Web of Conferences. Vol. 211, p. 04010.

[7] Elzainy, A., El Sadik, A., & Al Abdulmonem, W. (2020). Experience of e-learning and online assessment during the COVID-19 pandemic at the College of Medicine, Qassim University. Journal of Taibah University Medical Sciences, 15(6), 456-462.
[8] Skar, G. B. U., Graham, S., & Huebner, A. (2022). Learning loss during the COVID-19 pandemic and the impact of emergency remote instruction on first grade students' writing: A natural experiment. Journal of Educational Psychology, 114(7), 1553.
[9] Rowe, N. C. (2004). Cheating in online student assessment: Beyond plagiarism. Online Journal of Distance Learning Administration, 7(2), 1-10.
[10] Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2021). Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education. Postdigital Science and Education, 3(3), 715-742.
[11] Le Thi Minh Que. (2021). Online Teaching and Learning in Higher Education During Covid-19 Pandemic: Vietnamese Students' Perspective. IUP Journal of Information Technology, 17(3), 23-48.
[12] Foo, Cc., Cheung, B., & Chu, Km. (2021). A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic. BMC Medical Education, 21, 141, 2021. doi:10.1186/s12909-021-02575-1.
[13] Iwase, K. O. S. (1989). Linear regression through the origin with constant coefficient of variation for the inverse Gaussian distribution. Communications in Statistics-Theory and Methods, 18(10), 3587-3593.
[14] Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. Review of Education, 9, e3310. doi:10.1002/rev3.3310.
[15] What is linear regression, 2023, URL: https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.
[16] Maulud, D., & Abdulazeez, A. M., A Review on Linear Regression Comprehensive in Machine Learning. JASTT, 1(4), 140-147, 2020.
[17] Smolic H., How Much Data Is Needed For Machine Learning? 2022, URL: https://graphite-note.com/how-much-data-is-needed-for-machine-learning#:~:text=Generally%20speaking%2C%20the%20rule%20of,100%20rows%20for%20optimal%20results.
[18] Newton, P. M., & Essex, K., How Common is Cheating in Online Exams and did it Increase During the COVID-19 Pandemic? A Systematic Review. Journal of Academic Ethics, 2023. doi:10.1007/s10805-023-09485-5.
[19] Wickramasinghe, S., Bias & Variance in Machine Learning: Concepts & Tutorials, 2021. URL: https://www.bmc.com/blogs/bias-variance-machine-learning/#:~:text=Bias%20is%20considered%20a%20systematic,prediction%20and%20the%20ground%20truth.
[20] Coding Prof, 3 Ways to Calculate the Root Relative Squared Error (RRSE) in R, 2022, URL: https://www.codingprof.com/3-ways-to-calculate-the-root-relative-squared-error-rrse-in-r/.
[21] Padhma M., Models, Data Science Blogathon, 2023, URL: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/.