Strategies to predict students' exam attendance

Gonzalo L. Villarreal[1,2][0000−0002−3602−8211] and Verónica Artola[3][0000−0003−0833−1077]

[1] Universidad Nacional de La Plata, PREBI-SEDICI
[2] Comisión de Investigaciones Científicas, CESGI
gonzalo@prebi.unlp.edu.ar
[3] Universidad Nacional de La Plata and Comisión de Investigaciones Científica, III-LIDI
vartola@lidi.info.unlp.edu.ar

Abstract. This article presents a study on predicting student attendance to exams in a university setting. The study focused on the Concept of Algorithms, Data, and Programs course, a foundational course in systems bachelor. Two models were constructed: linear regression and polynomial regression of degree 3, aimed to predict the total number of attendees and the number of students who would pass the exam. We built a dataset that included information on student enrollment, previous exam attendance, grades, and other relevant factors. Students were classified into three groups: reduced exam, complete exam with prior attendance, and complete exam without prior attendance. The results showed that the models' predictions were accurate enough, and that they could be used to ensure appropriate classroom occupancy without overcrowding or empty rooms. The models guided the allocation of students, optimizing space utilization while providing available seats for attending students. The study identified opportunities for improvement. One limitation was the assignment of attendance probabilities to achieve the overall predicted attendance. Future work could involve predicting attendance rates for each group individually. Additionally, implementing a classification model to categorise students into pass, fail, insufficient, and non-attendance groups would provide a more comprehensive understanding of student outcomes.

Keywords: regression analysis · attendance prediction · approval prediction · effective resource planning

1 Introduction

In educational institutions, predicting student attendance for exams plays a crucial role in effective planning and resource allocation. Having reliable estimates of attendance enables educators to make informed decisions regarding seating arrangements, printing exam materials, and overall logistics. By leveraging predictive models, we can forecast student attendance with reasonable accuracy and facilitate better preparation for exams.

In this article, we will walk you through an implementation of a linear regression and a polynomial regression model to predict student attendance for exams which offer many benefits, including:

- Resource Planning: by knowing the expected number of students attending an exam, administrators can plan seating arrangements, arrange adequate exam materials, and ensure a smooth experience for both students and staff.
- Timely Communication: educational institutions can inform students about essential exam details, such as exam location, timing, and any specific instructions, well in advance
- Performance Analysis: by comparing attendance rates with exam scores, educational institutions can identify potential correlations and gain insights into factors affecting student success.
- Efficient Resource Utilization: for instance, if a lower-than-expected attendance is forecasted, institutions can consider consolidating examination rooms, saving on logistics costs and reducing the environmental impact associated with exam preparations.

The subsequent sections will outline the steps involved in implementing the regression models, including data preparation, model training, evaluation, and prediction.

2 State of the Art

In recent years, several researchers have explored alternatives to predict student attendance in educational settings. Maud Vissers ([1]) investigated the probability of predicting class attendance for students' personal development, for professors' preparation and intervention, and to optimise universities' educational programs. The author used Logistic Regression, Random Forest and Naïve Bayes in this study. He found that class attendance can be predicted based on sensor data and education data, and the best performing algorithm was the Random Forest algorithm containing GPS Location data, WiFi Location data, and Class Information data.

Muzaferija et al. ([2]) focused on the reasons why students' attendance decreased, in order to try to predict when it was going to happen, and act on causing factors in order to prevent it. They built a dataset containing 2nd-year student attendance data from two years, and although the dataset didn't contain all the details about the students and their classes it was enough to extract the patterns of student attendance behavior and create a model to predict it. In their study they found that the machine learning algorithm that created the most accurate model was the C4.5 decision tree algorithm with 77.5% accuracy, followed by a Linear Regression algorithm with 75,37% accuracy. Fernandes et al. ([3]) introduced a classification model based on Gradient Boosting Machine (GBM), the demographic characteristics of the students and the achievement grades obtained from the in-term activities were taken into consideration. In

this study, the authors observed the importance of previous year's achievement scores and attendance data for estimating students' achievement.

M. Yağcı ([4]) proposed a new model based on machine learning (ML) algorithms to predict the final exam grades of undergraduate students. In this model, he considered the students' midterm exam grades as the source data, combined with Department data and Faculty data for each student. He compared different ML algorithms, including logistic regression, k-nearest neighbor algorithms and random forests, among others. The model achieved a classification accuracy of 70–75%, with 71.7% of accuracy for the Linear Regression based model.

Another ML-based approach was proposed by Rashid et al. ([5]). In their work, the authors used ML techniques to predict students' attendance to classes.

Considering different reasons why students skip classes, they built a dataset by collecting 2 years of attendance and they used a variety of machine learning algorithms to predict attendance, including LR, support vector machines, and decision trees. They found that all of the algorithms were able to predict attendance with a high degree of accuracy. The authors suggest that teachers can use ML to identify students who are at risk of missing class, and can then take steps to address the needs of these students. They also suggest that ML could be used to improve the efficiency of teaching. For example, teachers could use ML to predict which students are likely to need extra help on a particular topic and thus create individualised learning plans for them.

Retention prediction studies can also contribute with variables and techniques to predict attendance. Robert D. Reason ([6]) considered high school performance (high school GPA and SAT/ACT scores) as a variable to predict students' graduation rates. He mentions that students who entered college with a high school GPA were more likely to graduate with a degree in 4 years than students who entered with a low GPA. Similarly, higher SAT scores were also associated with higher graduation rates. However, the author clarifies that the effect size of these variables was relatively small. They only predicted 12% of the variation in retention. Even though we are not predicting students' retention, we also consider the results in the initiation course as an important variable in our study. Credé et al. ([7]) review the relationship of class attendance with grades and student characteristics, and they found strong relationships with class grades and GPA, which seem to be a better predictor of college grades than any other known predictor of academic performance, including scores on standardised admissions tests such as the SAT, high school GPA, study habits, and study skills.

2.1 About the course of CADP

The course "Concepts of Algorithms, Data, and Programs" (CADP, 2023) is a first-year subject in the Systems Bachelor and Computer Science Bachelor programs. Students taking this course can be either new students who enrolled in the current year or students who enrolled in previous years but did not pass the subject at the time of enrollment. Incoming students undertake a course called "Problem Expression and Algorithms" (EPA), where they are introduced to the

basic computational thinking and programming concepts required to progress in CADP. EPA is a prerequisite for CADP, but it is not an eliminatory course per se. To pass EPA, students must only meet a minimum attendance percentage requirement. However, there is an exam at the end of EPA, and students who pass this exam receive certain benefits as rewards for their CADP coursework. These benefits may include the opportunity to take a reduced exam and a priority at the time of selecting course hours.

Once students have successfully completed EPA, they can proceed to CADP, where they delve deeper into algorithms, data structures, and programming concepts. CADP is a comprehensive course that builds upon the foundations established in EPA. It covers topics such as algorithm analysis, data representation, programming paradigms, and problem-solving strategies. The CADP course is structured into lectures, practical sessions, and assignments. Assignments and projects are designed to reinforce the learned concepts and allow students to apply their knowledge to real-world problems.

Throughout CADP, attendance records are maintained to monitor students' participation and engagement in the course. Only students with a certain percentage of attendance are able to take the exam.

At the end of CADP, students must take a final exam to evaluate their understanding of the subject matter, which assesses their knowledge of concepts covered during the course. Successful completion of the exam is a requirement for obtaining a passing grade in CADP and progressing to the subsequent courses in the curriculum. Students have three opportunities to take this exam. We would like to add that the course delivery for CADP lasts for one semester, and there are course retakes offered in the second semester. However, while the course delivery in the first semester is open to all students (both incoming and returning), the course in the second semester is limited only to those students who have completed the course in the first semester, have met the required attendance to take the exam, have taken the exam, and have received a failing grade. This distinction between students in the first and second semesters results in differences in the total number of students attending each semester, as well as the attendance rate for exams and also the pass rate. It is important to consider these distinctions when analyzing the attendance and pass rates within the context of the course. Understanding the differences in student populations and their characteristics between the two semesters provides valuable insights into the dynamics of student performance and the factors that contribute to success or failure in the course. By acknowledging these variations, educators and administrators can develop targeted strategies and interventions to address the specific challenges faced by students in both the traditional course delivery and the course retake.

3 Methodology

To address the problem of predicting student attendance for an exam, two different models were developed: one based on linear regression (LR) and another

using a polynomial regression of degree 3 (PR). These models allowed us to predict both the total number of attendees and the total number of students who would pass the exam. Having two predictive models allows us to compare results between them. Additionally, while the LR model is simpler, the PR model may better capture the statistical variations between the first and second semesters, considering the specific course delivery characteristics described earlier. In the context of the CADP course, where different instructional modes and student cohorts are involved, the polynomial model's ability to account for statistical variations between semesters can be advantageous. It can capture nuances and dynamics specific to each semester, such as the difference in enrollment and attendance rates for the traditional course delivery and the course retake.

We followed a systematic methodology that involved the construction and preparation of a comprehensive dataset, with data for the period 2019-2023 (Villarreal, 2023). The dataset encompasses information about students eligible to take the exam, students who actually appeared for the exam, and the number of students who achieved different grades. The grades are classified as "Approved" (an exam that was well done, although it may have had some errors), "Disapproved" (an exam that was incorrect but showed an attempt to solve the problem), and "Insufficient" (a very incomplete exam).

Additionally, the dataset includes information about the academic year, semester, exam number and course modality, categorised as "In-person" (during regular on-campus sessions), "Virtual" (during the COVID pandemic when classes were conducted online), and "Hybrid" (during the transition period after the pandemic, combining in-person and virtual elements). During the COVID pandemic, since both classes and exams were conducted online, students' attendance were higher than usual, although pass rates were a bit lower. Not taking into consideration this parameter may introduce severe distortions in our models.

The steps involved in constructing the dataset and preparing it for the models are as follows:

1. Data Collection: We collected information from various sources, including student records, attendance registers, and exam grading records.
2. Data Cleaning: We carefully cleaned the dataset by removing a few duplicates, but specially completing missing values or inconsistent entries. Additionally, we performed data validation checks to ensure the accuracy and integrity of the data.
3. Feature Engineering: To enhance the predictive power of the model, we performed feature engineering. This involved transforming the categorical variables, such as semester, exam number, and course modality, into numerical representations using appropriate encoding techniques.
4. Linear and Polynomial Regression Models: We implemented a LR model and a PR of degree 3 model. Both models take into account the features derived from the dataset.
5. Model Training and Evaluation: We divided the dataset into training and testing sets to train the polynomial regression model. The model was trained using the training set, and its performance was evaluated on the testing set.

6. Prediction and Analysis: Once the model was trained and evaluated, we utilised it to make predictions on new data. These predictions were then analyzed to gain insights into student attendance patterns and to identify factors that significantly impact attendance.

Based on the predictions obtained from the models, the students were categorised into three groups:

1. Students taking a reduced exam (group Reduced): This group consisted of students who had previously passed the EPA exam. They were eligible to take a shorter version of the exam.
2. Students taking a complete exam and attended previous exam (Group Complete): This group comprised students who were attending the complete exam and had also attended the previous exam.
3. Students taking a complete exam and did not attend previous exam (Group Complete Absent): This group consisted of students who were attending the complete exam but had not attended the previous exam.

Each student was assigned to one of these groups based on their eligibility and attendance history. Using this information, the attendance percentages for each group were estimated in such a way that the total attendance would match the predicted attendance from each model (table 1).

Table 1. Number of students in each of the three groups, and attendance estimation for both models for the second exam.

Group	Number of students	Linear Regression Attendance Estimation	Polynomial Regression Attendance Estimation
Reduced	262	84,10%	86%
Complete	541	79%	80,10%
Complete Absent	966	11%	11,37%

By incorporating this categorization and adjusting the attendance percentages accordingly, we aimed to align the predicted total attendance with the attendance estimated by the selected model. This approach allowed for a more accurate representation of the different student groups and their corresponding attendance patterns.

3.1 Predicting students that will pass the exam

Although it was not the purpose of these models to predict the results of the exams, the predictions from the initial regression models were further utilised as an input for a subsequent predictive model to estimate the percentage of students who would pass the subject. To achieve this, the predicted attendance values obtained from the regression models were combined with the previous

dataset. This data was used to train new predictive models, again using LR and PR, which aimed to forecast the percentage of students who would successfully pass the subject. Once the predictive models were trained and evaluated for accuracy, it was applied to the current cohort of students to make predictions on the probability of passing the subject.

4 Prediction results

At the moment of writing these lines, the first exam of the year had already been taken, so the dataset also included attendance and results data for this cohort. Thus, this model has been used to predict attendance and pass rates for the second and third exam (in the later we have also included data from the former into the dataset).

For the second exam, there were a total of 14 classrooms available, each with a different maximum capacity. These classrooms were organised into three groups, corresponding to the same groups into which the students were classified (reduced, complete with previous attendance, and complete without previous attendance). Next, the students were assigned to the classrooms based on their respective groups. The goal was to achieve an occupancy level between 40 and 90 percent in each classroom. This approach aimed to maximise the utilization of available space while ensuring that all attending students had a seat (table 2).

Table 2. Classroom organization and student distribution.

Class-room	Max. Capacity	Group	Assigned Students	Estimated students by LR	Percent-age of estimated atten-dance by LR	Estimated students by PR	Percent-age of estimated atten-dance by PR
1	50	Reduced	52	43,73	87,46%	44,72	89,44%
2	50	Reduced	51	42,89	85,78%	43,86	87,72%
3	50	Reduced	53	44,57	89,15%	45,58	91,16%
4	120	Reduced	106	89,15	74,29%	91,16	75,97%
5	200	Complete	151	119,29	59,65%	120,95	60,48%
6	120	Complete	89	70,31	58,59%	71,29	59,41%
7	100	Complete	69	54,51	54,51%	55,27	55,27%
8	100	Complete	70	55,30	55,30%	56,07	56,07%
9	200	Complete	162	127,98	63,99%	129,76	64,88%
10	80	Absent	248	27,28	34,10%	28,20	35,25%
11	80	Absent	247	27,17	33,96%	28,08	35,10%
12	50	Absent	155	17,05	34,10%	17,62	35,25%
13	50	Absent	157	17,27	34,54%	17,85	35,70%
14	50	Absent	159	17,49	34,98%	18,08	36,16%

By distributing the students in this manner, we were able to effectively allocate the available resources and accommodate the predicted attendance levels for each group. The varying classroom capacities allowed for flexibility in accommodating different numbers of students within each group. The allocation of students to classrooms based on their respective groups ensured that the seating arrangements were optimised, and the available space was utilised efficiently. This approach aimed to strike a balance between maximizing capacity utilization and ensuring that all attending students had a seat. Overall, the results of this allocation strategy enabled the creation of an organised and conducive environment for the exam, where students had appropriate seating arrangements according to their attendance status. By achieving an optimal occupancy level (4) in each classroom, we were able to utilise the available space effectively while accommodating the expected number of attending students.

Table 3. Actual attendance of students and predictions from both estimators for the second exam.

Global Attendance	Percentage of global attendance	Linear Regression Global Attendance estimation	Linear Regression Attendance accuracy	Polynomial Degree 3 Global Attendance Estimation	Polynomial of degree 3 global attendance accuracy
721	40,57%	757,43	95,19%	774,90	93,04%

Table 4 shows the percentage of students that attended the exam, and the accuracy of each model. Note that the occupation level of all classrooms were between 38% and 84%, with a mean of 56,35% and a standard deviation of 13,28% . Even though both models predicted with high precision the attendance percentage with 2 3 percent difference (table 3), classroom attendance distribution was not as accurate as expected. The lack of precision in the predictions can be attributed to the estimated assignment of probabilities to each student group. The approach used in this study involved assigning probabilities to achieve the overall predicted attendance, rather than predicting attendance rates for each group individually. However, data obtained from the second exam was used to calculate a much better probability of attendance for each group, as shown in Table 4, by adjusting the percentage of students who actually were present in each classroom combined with the group assigned to each classroom.

For the third exam, after data from the second exam was added to the dataset, a similar approach was taken. However, we decided to split the Reduced group into two subgroups, absent reduced (students that never attended any exams) and reduce (students that attended at least once). A different criteria was taken for the complete groups: we considered as complete and absent all students who never attended any exam or who attended only once and got insufficient. Another important consideration for the third exam is that not all classrooms were available: classrooms 7, 10 and 11 were occupied by other courses, so we

Table 4. Accuracy of estimators in each classroom for the second exam. An accuracy of 100% means a perfect prediction, while an accuracy of 100% ± 25% means a fair enough prediction. Note that classrooms 10 to 14, in which students were absent in previous exams, tend to obtain less accurate predictions. However, even though the number of assigned students were very high, the occupation level was similar to most classrooms.

Classroom	Real attendance percentage	LR accuracy	PR accuracy
1	76,00%	86,89%	84,97%
2	68,00%	79,27%	77,52%
3	84,00%	94,23%	92,15%
4	66,67%	89,74%	87,76%
5	49,50%	82,99%	81,85%
6	52,50%	89,60%	88,37%
7	46,00%	84,39%	83,23%
8	42,00%	75,95%	74,91%
9	57,00%	89,08%	87,85%
10	61,25%	179,62%	173,77%
11	50,00%	147,22%	142,43%
12	38,00%	111,44%	107,81%
13	46,00%	133,18%	128,85%
14	52,00%	148,66%	143,82%

Table 5. Attendance rate per classification group after the second exam. The average classroom attendance rate column indicates a good distribution of students that ensures good attendance rate in all classrooms.

Group	Assigned of students	Actual students	Global attendance rate	Average classroom attendance rate
Reduced	262	194	74,04%	74,5%
Complete	541	364	67,28%	69,4%
Absent	966	158	16,35%	15%

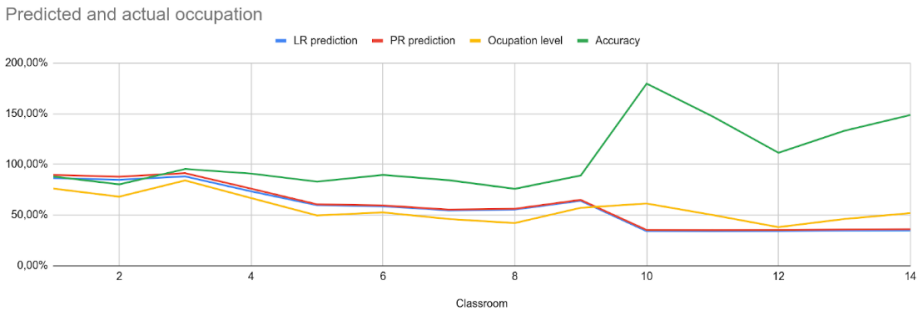

Fig. 1. Models' predictions and accuracy per classroom.

had about 260 less seats available. However, since the LR model predicted an attendance rate of 33,35% of students and PR model predicted 35,75% attendance, this was not a real issue to attend. Results showed an actual attendance rate of 36,5%, so in this case PR model performed better than LR model (table 4), with an error of only 0.75% (12.7 students over 600). For the allocation of students in classrooms we took a similar approach for the reduced exam students, but a different one from complete and absent: considering a near 1:2 relation for these groups (467 students for the complete group, and 976 for the absent group), we distributed one complete for every two absent in each classroom. To measure the estimation classroom occupation level, we calculated the expected value E for the whole complete group (complete and absent) as:

$$E = \frac{C \times P(C) + CA \times P(CA)}{C + A} \tag{1}$$

where C is the total number of students in the Complete group, CA is the total number of students in the Complete Absent group, and P(C) and P(AC) are the probabilities for students in each group attending the exam. These probabilities are based on the estimation of both LR and PR models (Table 6).

Table 6. Attendance estimations using the expected value E for both complete and complete absent groups.

Group	Number of students	LR Estimation	PR Estimation
Global estimation	1645	*33,35%*	*35,75%*
Reduced	282	83,00%	86,20%
Complete	467	78,10%	80,20%
Reduced Absent	60	4,90%	7,00%
Complete Absent	976	7,50%	9,00%
C + CA	1443	E = 28,59%	E = 30,69%

Table 7. Actual attendance of students and predictions from both estimators for the third exam

Global Attendance	Percentage of global attendance	Linear Regression Global Attendance estimation	Linear Regression Attendance accuracy	Polynomial degree 3 Global Attendance Estimation	Polynomial of degree 3 global attendance accuracy
600	36,50%	551,3	91,88%	587,31	97,88%

As mentioned above, even though the PR model performed better than the LR model for the third exam, both models performed accurately enough to make a good distribution of students, auxiliars and exams. Table 8 shows the accuracy

Table 8. Actual attendance, models' estimation and models' estimation per. Classrooms 1, 2, 3 and 14 were assigned to reduced exams only, classroom 13 to the reduced absent group, and the rest for complete and complete absent groups with a 1:2 ratio.

Classroom	Attendance percentage	LR estimation	LR accuracy	PR estimation	PR accuracy
1	77,78%	73,78%	110,97%	76,62%	106,85%
2	84,44%	75,62%	111,67%	78,54%	107,52%
3	73,33%	71,93%	101,95%	74,71%	98,16%
4	75,00%	50,27%	151,34%	53,96%	140,99%
5	46,50%	53,03%	88,15%	56,93%	82,12%
6	78,33%	50,51%	155,82%	54,22%	145,16%
8	63,33%	57,75%	132,25%	61,99%	123,20%
9	44,50%	52,32%	83,24%	56,16%	77,54%
12	38,00%	42,88%	89,81%	46,03%	83,66%
13	20,00%	6,15%	325,20%	7,38%	271,00%
14	36,00%	66,40%	54,22%	68,96%	52,20%

levels achieved in each classroom, with a mean of 105,10% $\pm$ 62,67% std. dev. and a median of 98,16%. The error rate value was higher than expected, nevertheless most classrooms were occupied between 40% and 80%, which fulfills the purpose of this work. We observed that the error rate was primarily due to classroom 13, which presents a special case in which both models predicted less than half of the students that actually attended. However, this situation was expected due to the uncertainty of the group classified as *reduced absent*, and it did not present a real issue since its occupancy level was 20% (10 students over 41, with predictions between 3,08 and 3,69).

A last prediction of the models was the rate of students that would pass the exam. In the second exam, the LR model predicted a 19,5% of approval rate (139 students), while the PR model predicted a 20,42% approval rate (146 students). The real approval rate was 18,88%, corresponding to 135 students , so even though the predictions weren't perfect, the models performed well enough to estimate the number of students who would pass the second exam. For the third exam, the LR model predicted a 15,93% approval rate, while the PR model predicted a 16,02% approval rate . In this case, the real approval rate was 19,01%, so the models performed less accurately in the third exam, but again close enough (with a difference of about 20 students). The precision of these early estimations are very useful to gain several weeks in advance to organise CADP for the second semester (retake format) and Programming Workshop, the course that follows CADP in the same semester.

5 Analysis and conclusions

The primary objective of the attendance prediction models was to estimate the number of students expected to attend the exam, allowing for effective allocation of resources and seating arrangements. While the models may not have provided

exact attendance figures, their predictions provided a reliable guideline for organizing the exam logistics. It is worth noting that although the predictions of the models were not exact, they were accurate enough to ensure an appropriate level of occupancy in all the classrooms used, avoiding situations where classrooms were overcrowded or nearly empty. LR model performed slightly better than PR model, both for attendance and approval estimations. However, as the dataset expands and more years of course data are incorporated, it is anticipated that the models will become increasingly accurate and robust in their predictions, and we expect the PR model to improve its accuracy by capturing the statistical variations between the both semesters.

By considering the predicted attendance levels, the allocation of students to classrooms could be carefully planned to achieve an optimal utilization of space: classrooms were neither overcrowded nor sparsely populated. This outcome is crucial in maintaining a smooth and efficient exam administration process while ensuring that all attending students had a seat available to them. While it is always desirable to have precise attendance predictions, the fact that the models provided sufficiently accurate estimates to avoid overcrowded or empty classrooms should be considered a strength of the approach.

One of the limitations of this study is that the assignment of attendance probabilities for different student groups was done in a way to achieve an overall attendance percentage according to the selected model. A future improvement could involve predicting not only the overall attendance but also the attendance percentages for each specific student group. This enhancement to the predictive models can be used to gain more granular insights into attendance patterns and better understand the dynamics within different student cohorts.

As noted before, the adjustments per group (tables 5 and 8) were made by using *really* simple average approaches, with a small difference for the third exam in which we mixed students for the complete and absent groups. However, a much better adjustment could be achieved by running an optimization algorithm per group, such as minimum square error (MSE). For instance, let's consider attendance observations for the second exam, together with attendance predictions based on the LR model. We have already calculated the accuracy of the model in each classroom, and we also know the group each classroom belongs to (either reduced, complete or absent). Having this information into consideration, we can set a limitation per group using the real number of students which actually attended the exam, and calculate the best value of X for the same group (estimated attendance rate of the group X) by minimizing the square error between 1 (perfect estimation) and the model estimation per classroom.

Another improvement to this model would be implementing a classification model to categorise students into four possible groups: 1) Will pass, 2) Will fail, 3) Will receive an insufficient grade, and 4) Will not attend the exam. The classification model could be built using techniques such as logistic regression, decision trees, or support vector machines. By utilizing a classification model, we can go beyond predicting the overall pass rates and gain insights into the likelihood of individual students falling into different performance categories. This

information can provide a more detailed understanding of student outcomes, and would enable educational institutions to identify students who may require additional support or intervention early on, facilitating timely interventions to improve their chances of success.

To implement this improvement, we would need to gather additional data that captures factors influencing student performance, such as prior academic records and engagement in coursework. One advantage of a classification model is the amount of available data: while regression models are based on less than 25 records (3 exams per semester, two semesters per year, 4 years of data plus data from this year), a classification model could base its predictions on several thousands records (between 500 and 3000 per semester). However, gathering records and constructing and preparing this dataset would require more work than the one used in this study.

References

1. Vissers, M.: Predicting Students' Class Attendance. Master Thesis Data Science Business and Governance. Tilburg University, School of Humanities. Tilburg, The Netherlands. October 2018
2. Muzaferija, I., Mašetić, Z., Jukic, S., Kečo, D.:Student Attendance Pattern Detection and Prediction. Journal of Engineering and Natural Sciences. 3. https://doi.org/10.14706/JONSAE2021313
3. Fernandes, Eduardo; Holanda, Maristela; Victorino, Marcio; Borges, Vinicius; Carvalho, Rommel; Cordeiro Galvão van Erven, Gustavo: Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research. 94. 10.1016/j.jbusres.2018.02.012.
4. Yağcı, M. (2022) 'Educational data mining: prediction of students' academic performance using machine learning algorithms'. Smart Learn. Environ. 9, 11 (2022). https://doi.org/10.1186/s40561-022-00192-z
5. Rashid, E., Ansari, M.D., Gunjan, V.K., ; Khan, M.: Enhancement in Teaching Quality Methodology by Predicting Attendance Using Machine Learning Technique. In: Gunjan, V., Zurada, J., Raman, B., Gangadharan, G. (eds) Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough. Studies in Computational Intelligence, vol 885 . Springer, Cham. https://doi.org/10.1007/978-3-030-38445-6-17
6. Reason, R.: Student Variables that Predict Retention: Recent Research and New Developments. Journal of Student Affairs Research and Practice, 40(4), pp. 704-723. https://doi.org/10.2202/1949-6605.1286
7. Credé, M., Roch, S. G., ; Kieszczynka, U. M.: Class Attendance in College: A Meta-Analytic Review of the Relationship of Class Attendance With Grades and Student Characteristics. Review of Educational Research, 80(2), pp. 272–295. https://doi.org/10.3102/0034654310362998