

Evaluating Algorithmic Bias in Models for Predicting Academic Performance of Filipino Students

Valdemar Švábenský
University of Pennsylvania &
Kyushu University
valdemar.research@gmail.com

Clarence James G.
Monterozo
Ateneo de Manila University
jamesmonterozo@gmail.com

Mélina Verger
Sorbonne University, LIP6
melina.verger@lip6.fr

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Maria Mercedes T.
Rodrigo
Ateneo de Manila University
mrodrigo@ateneo.edu

Miguel Zenon Nicanor
Lerias Saavedra
Ateneo de Manila University
msaavedra@ateneo.edu

Sébastien Lallé
Sorbonne University, LIP6
sebastien.lalle@lip6.fr

Atsushi Shimada
Kyushu University
atsushi@ait.kyushu-u.ac.jp

ABSTRACT

Algorithmic bias is a major issue in machine learning models in educational contexts. However, it has not yet been studied thoroughly in Asian learning contexts, and only limited work has considered algorithmic bias based on regional (sub-national) background. As a step towards addressing this gap, this paper examines the population of 5,986 students at a large university in the Philippines, investigating algorithmic bias based on students' regional background. The university used the Canvas learning management system (LMS) in its online courses across a broad range of domains. Over the period of three semesters, we collected 48.7 million log records of the students' activity in Canvas. We used these logs to train binary classification models that predict student grades from the LMS activity. The best-performing model reached AUC of 0.75 and weighted F1-score of 0.79. Subsequently, we examined the data for bias based on students' region. Evaluation using three metrics: AUC, weighted F1-score, and MADD showed consistent results across all demographic groups. Thus, no unfairness was observed against a particular student group in the grade predictions.

Keywords

fairness, online learning, distance education, performance prediction, educational data mining, learning analytics

1. INTRODUCTION

Algorithmic bias in education can be defined in a wide range of ways [7, 2]. In this paper, it denotes *group unfairness* [9], i.e., situations when the performance metrics for a machine learning model substantially differ across groups of students.

V. Švábenský, M. Verger, M. M. T. Rodrigo, C. J. G. Monterozo, R. S. Baker, M. Z. N. L. Saavedra, S. Lallé, and A. Shimada. Evaluating algorithmic bias in models for predicting academic performance of Filipino students. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 744–751, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

<https://doi.org/10.5281/zenodo.12729936>

Here, *group* is defined by a set of immutable characteristics, for example, the student's native language or hometown.

Algorithmic bias can arise in any stage of the educational data mining (EDM) pipeline [7, 23]. Subsequently, it can lead to discriminatory treatment or even harm to certain groups of students, negatively affecting their learning [19]. This often comes with severe social, ethical, or legal implications [23]. Therefore, it is essential to study algorithmic bias in order to inform decision-makers about how to possibly mitigate it.

1.1 Background

Paquette et al. [31] explored how EDM studies employed demographic data, concluding that the field needs more research to validate model performance on various student subpopulations. Baker and Hawn [7] urged the community to conduct more studies across contexts, indicating that there are two key areas where more work is needed.

First, almost all published research focuses on learners in the United States [7]. Populations in other countries were studied rarely. As part of addressing this literature gap, our paper studies these issues in relation to learners in the Philippines. Since bias may have different characteristic manifestations in different contexts, it is important to search for “evidence for algorithmic bias impacting groups that have not been previously studied” [7]. As a result, considering a perspective from another country contributes towards global understanding of the complex issue of algorithmic bias.

Second, previous work often defined groups of learners using distinctions that are too coarse [7]. For example, all Asian-Americans in the published research on algorithmic bias in education were usually treated as the same group, even though Asian countries – or even regions within a single country – have numerous cultural differences. The second unique aspect of this paper is investigating fine-grained demographic categories within the geographic regions of the Philippines. This is a response to a call for researchers to explore differences across finely-differentiated groups [8].

1.2 Motivation and Local Context

Studies conducted worldwide during the COVID-19 pandemic revealed that learning losses were larger for lower-income countries [41]. In the Philippines, learners from resource-scarce backgrounds encountered greater challenges compared to those with robust access to digital resources [10]. Students from less affluent families grappled with the scarcity of computers, mobile devices, and reliable Internet connection, which greatly limited their ability to fully engage in online learning [13]. Students from rural and low-income areas also reported lower online learner readiness [12].

The Philippines, with its diverse landscape across the individual regions, illustrates this digital divide. Access to the Internet and cellular services within the country varies significantly, depending on socioeconomics and regional geography [38]. Urban centers generally enjoy high-speed Internet, while rural areas possess limited digital resources. In areas like Metro Manila, Central Luzon, and Cebu, both wealth and Internet speeds are higher. In contrast, regions such as Palawan, Eastern Visayas, and Northern Mindanao grapple with heightened poverty rates and slower Internet speeds.

1.3 Research Problem Statement

Our study focuses on students from a Filipino university during 2021–2022. At that time, the Philippines remained under COVID-19 lockdown measures, compelling universities to pivot to remote learning. Students were prohibited from coming on campus and had to complete their academic requirements from home, in different regions of the country. To facilitate online learning, stable Internet connection and technological resources are crucial. However, given the considerable disparities throughout the Philippines, as well as the cultural differences of different regions, we hypothesize that students’ academic achievements during this period were significantly influenced by their geographical locations.

This paper explores the following research question: *When building predictive models of academic performance of Filipino students in online courses, do the models exhibit identifiable algorithmic bias towards a certain demographic?* Here, the demographic groups are defined based on the location from which the students accessed the online courses. We assume that the location is also highly indicative of the students’ regional background, not only their place of residence. And, since different geographic locations have different properties and constraints on distance learning, we want to see whether failing to account for differences between students in different locations introduces bias in the model predictions.

2. RELATED WORK

Since our research focuses on two topics: prediction of academic performance and bias, we review papers in these areas.

2.1 Student Performance Prediction

We study predicting performance of students as it is beneficial for teaching faculty and university administrators [2]. Instructors can proactively offer assistance and support to students identified as being at risk of poor performance. Administrators can use the prediction results to guide the retention of students. Recently, related work focused on explainable prediction models [18, 37], which would better support the needs of these stakeholders.

Hellas et al. [16] published an extensive review of student performance prediction. They pointed out that the values being predicted can range from *long-term*, such as college graduation [5], through *medium-term*, such as course grades [30], to *short-term*, such as assignment performance [17]. Data sources that provide features for prediction include course data, student demographics, and psychometric information such as motivation. Our work employs course data for prediction and student demographics for bias evaluation. Commonly used classification models in EDM include Decision Tree, Random Forest, K-Nearest Neighbors, eXtreme Gradient Boosting, and Logistic Regression [24]. For the sake of comparison, we use these models as well.

2.2 Algorithmic Bias in Prediction Models

Although most EDM research did not address the issue of algorithmic bias [31], several studies have been conducted over the past decade. To prevent bias, researchers argue against using demographic variables as predictors [6, 14]. In a similar vein, the output of a predictive model should not be influenced by the student’s sensitive attributes [19]. Thus, we seek to evaluate to what extent the models are fair regarding these attributes in an Asian context. We train binary classifiers to predict whether a student will receive a grade better than the average and evaluate disparities across fine-grained groups using fairness metrics (see Section 3.5).

Lee and Kizilcec [25] argued that student success prediction models may introduce bias by failing to identify a successful student “because of their membership in a certain demographic group” [25]. The study focused on the US context and unfairness with respect to race or gender. It evaluated a Random Forest model that predicted whether students from a US university will receive at least a median grade in a course. Based on two fairness measures (equality of opportunity and demographic parity) out of three measures considered, the model exhibited bias against racial minority students and male students.

Jiang and Pardos [22] implemented a recurrent neural network for predicting course grades of students at a US university. Then, they compared the overall model’s performance (measured by accuracy, sensitivity, and specificity) against the performance of models when the students were divided into seven groups by race. Since the results exhibited some bias, the authors compared four fairness strategies, out of which adversarial learning achieved the best fairness scores.

Bayer et al. [11] studied a large population of university students in the UK. The performance of models for predicting student success was measured by AUC and false positive/negative rate. For all divisions of student groups (based on ethnicity, gender, or disability status), the models achieved slightly better predictions for the majority group.

A study of three countries – the US, Costa Rica, and the Philippines – confirmed the intuition that predictive models trained on data of students from one country generalize better to other students from the same country [29]. Thus, such usage leads to less bias compared to when the models are deployed in another country. Although this research also examined the Filipino context, it focused on model generalization and not group fairness analysis as in our paper.

3. RESEARCH METHODS

Our goal is to build and evaluate classification models that predict students' grades from submissions (such as assignments and quizzes) in university courses.

3.1 Educational Context of the Study

Our study analyzes data of a population of 5,986 students, all of whom were enrolled in a large university in the Philippines. The data set does not include any external students from other institutions. All students were at least 18 years old, and they were studying undergraduate or graduate programs across different academic disciplines taught at the university in 2021–2022.

Student data were collected from courses that used the university's Canvas LMS [20]. The courses spanned a wide range of offerings: from humanities to natural sciences and engineering. Many of the courses were part of the university's common core curriculum, which includes languages, philosophy, psychology, etc., while other courses were specific to the college majors, e.g., programming or marketing. The course data were combined to ensure broader generalizability (and, if courses were considered separately, the data pool for some of them would be too small). All courses ran fully remote due to COVID-19 restrictions at the time of data collection.

3.2 Collection of Student Data

The research data contain two types of information: (1) students' usage of Canvas LMS (i.e., their actions in the system) to perform the prediction modeling, and (2) students' approximate location when working with the LMS (i.e., their region) to evaluate the algorithmic bias.

3.2.1 Canvas LMS Data (Learning Traces)

Appendix A details the technical aspects of the data collection. Overall, the data of the 5,986 students represent the following possible actions within the LMS:

- accessing a learning resource (e.g., viewing a PDF file with lecture slides),
- uploading/editing a file (e.g., a homework attachment),
- posting in the discussion forum,
- submitting an assignment/quiz,
- receiving a grade (e.g., for submitting an assignment).

Each student in the data set is uniquely identified by a 5-digit Canvas student ID ("student ID" further in the text). For privacy, this ID is different from the student ID issued by the university. Each course is also identified by a 5-digit ID.

3.2.2 Demographic Data (Geographic Location)

For each student ID, our records contain the city and region where the student resided when connecting to Canvas. There are 18 possible categorical values: either one of the 17 official regions defined by the Philippine Statistic Authority [33], or, in very rare cases, abroad (outside the country's borders).

Infrequently, it happened that a student connected to Canvas from multiple regions throughout the three semesters (this could happen when a student visited a family member, for example). In this case, we assigned the student to their

majority location, i.e., the one from which they connected the most often throughout the considered time period.

Since some of the 18 regions had very low representations, we subsequently grouped regions that were related, resulting in five larger meaningful clusters. The clusters were determined according to the standard division [33] based on the main island groups of the Philippines. The clusters are:

- National Capital Region (NCR), i.e., the Metro Manila urban area including the capital city – 4,816 students,
- Luzon (except the NCR) – 580 students,
- Mindanao – 138 students,
- Visayas – 131 students,
- Abroad (see the explanation above) – 49 students.

For all clusters (except Abroad), the locations inside that cluster have mostly similar context in terms of economy, culture, and language. Internet access might sometimes vary even within the regions on a smaller level of granularity. NCR is separated since it is the densest region, and its context is different from the other regions as it is very highly urbanized.

The student counts across the five regions sum to 5,714, since we lacked the location data for 272 students. So, we used the full data set of 5,986 students for the modeling (since these were valid log data from the population), and the data of a subset of 5,714 students for the bias evaluation.

3.2.3 Data Anonymization and Ethical Issues

The university had permission to use students' Canvas and location data (which the students were supposed to report) for research purposes. These data are paired only with the pseudonymous 5-digit Canvas student ID, so the researchers handling the data could not access other information about a particular student. All data were anonymized and used in accordance with the country's and university's regulations.

3.3 Preprocessing of Canvas LMS Data

Due to the huge size of the raw Canvas LMS data set, we automatically preprocessed it using a dedicated Python script (see Appendix B). The preprocessing involved computing the values of predictor and target variables, which are explained below. Then, the preprocessed data were used for building models that predict the student grades (see Section 3.4).

3.3.1 Target Variable (Grade)

Since different courses/assignments/quizzes in the data set use different grading scales, we normalized the grade as a decimal number between 0.0 and 1.0, which is the ratio between (a) the number of points that the student received and (b) the maximal number of points that was possible to be obtained in that particular context.

The average grade in the data set is 0.721. For the binary classification task, we aim to determine whether the student will receive a grade *worse than an average grade* (class 1, "unsuccessful") or not (class 0, "successful"). Out of the total of 211,945 grade entries (rows in the preprocessed data), 76,556 (36.1%) are below the average grade.

3.3.2 Predictor Variables (Features)

To predict the grade, we extracted features from events that a particular student performed in Canvas prior to receiving that grade within the context of a particular course. In other words, we explored whether and to what extent are student actions in Canvas predictive of their grade. We used 27 features grouped in the following categories:

- Count of events (examples: the number of resources accessed, the number of uploaded homework files, the number of submitted quizzes).
- Conditioned count of events (examples: the number of forum posts with a minimal length of 100 characters, the number of submissions that were not past the deadline, the average number of learning resources accessed per day of being active in the system).
- Timing of events (examples: the average/standard deviation of the time gap between two accesses of a learning resource or submitting a file).

3.4 Models’ Training and Cross-Validation

To provide a comparison and find a best-performing model, we selected five binary classifiers: Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGB), and Logistic Regression (LR). Their performance was evaluated by AUC ROC and weighted F1-score. As a baseline, we also trained a “dummy” classifier, which assigned each data point to the majority class (0), yielding AUC and weighted F1-score of 0.5. We used Python 3.10 and the Scikit-learn library [32] for the implementation.

The training set consisted of 211,945 samples and 27 features for all 5,986 students. For each model, feature scaling was performed for each predictor variable to unitize the values used for modeling. In addition, we experimented with feature selection by dropping features one-by-one (backward feature selection) or features that were correlated, but removing any feature worsened the model performance.

Model training involved 10-fold student-level cross-validation and a grid search of hyperparameters only on the training data. The training-validation split was stratified, and all samples for a given student appeared “exactly once in the test set across all folds” [36]. The evaluation metrics were computed as an average across the 10 cross-validation runs.

3.5 Models’ Bias Measurement

As mentioned in Section 1, we used a group fairness approach to measure algorithmic bias. For an initial check, we examined the distribution of grades in each group to see whether bias or a major difference is already present. The boxplot (Figure 1) and kernel density estimate plots [40] of grades indicated a similar distribution of grades across groups.

Subsequently, we compared the models’ performance across the five region groups to identify potential unfair predictive results. To do so, we chose complementary metrics: two traditional and one novel. Firstly, we used the two predictive performance metrics used for model evaluation, AUC ROC and weighted F1-score (both ranging from 0 to 1). Secondly, we used a dedicated fairness metric that evaluates the model behavior rather than its predictive performance, MADD

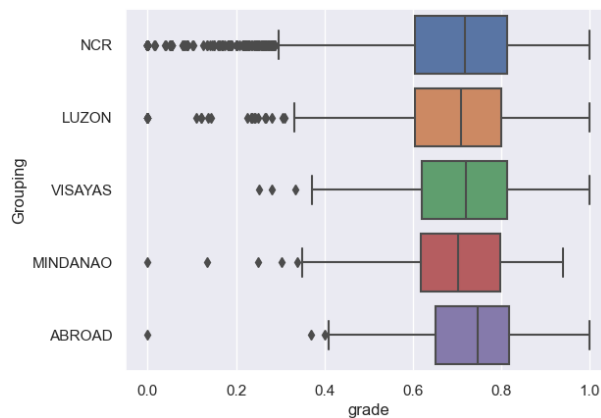


Figure 1: Grade statistics across the five student groups.

(ranging from 0 to 2, with smaller values being better) [39]. It evaluates how different the predicted probability distributions are between groups, even though the groups are not provided as features to the model and thus do not influence the probability distributions.

Thus, the first two metrics evaluate to what extent a model was as accurate for a group as for the others. The third one quantifies to what extent a model produced varying prediction distributions for a group over the others. We computed AUC ROC and weighted F1-score for each group, and MADD for each group against all other students (one-versus-rest approach). Therefore, for a model to be considered fair in this context, it should obtain for each group: AUC ROC and weighted F1-score with low variance (close to other groups), and low MADD score (close to 0).

4. RESULTS AND DISCUSSION

We first look at the classifiers alone, then evaluate their bias.

4.1 Binary Classification Models

In Table 1, the column “All” reports the prediction results, with standard deviation reported for each measurement in parentheses. Although the performance of Logistic Regression and KNN is rather unsatisfactory, they are still better than the naive baseline, and the other three models perform well enough to warrant the bias evaluation.

Within the literature on student performance prediction modeling, our best obtained AUC of 0.75 is slightly on the lower end compared to what was reported in the past. Examples of related work reported AUCs from 0.83 to 0.91 [34] or up to 0.95 depending on the week of the semester [21]. Nevertheless, 0.71 [4], 0.75 [35], and 0.77 [1] (as the best-performing among the compared models) were recently published.

4.2 Algorithmic Bias

Table 1 reports the bias evaluation. Each model produced consistent, even constant, AUC and weighted F1-score for each of the region groups. For instance, for the Random Forest model, the AUC ranged from 0.74 to 0.77 and the weighted F1-score from 0.78 to 0.82 for all the groups. Thus, the models did not produce discrepancies in predictive performance between groups.

Table 1: Results of modeling and bias evaluation. Average values across the 10-fold cross-validation are listed, with standard deviations in parentheses. The column “All” reports the performance of binary classification models on the data of 5,986 students, sorted descending by the AUC value. The remaining columns report per-group model performance on the data of 5,714 students.

| | | All | NCR | Luzon | Mindanao | Visayas | Abroad | Mean | $\Delta(\text{All})$ |
|-----|------|-------------|-------------|-------------|-------------|-------------|-------------|------|----------------------|
| RF | AUC | 0.75 (0.00) | 0.75 (0.01) | 0.74 (0.02) | 0.76 (0.04) | 0.75 (0.03) | 0.77 (0.04) | 0.75 | 0 |
| | F1 | 0.79 (0.00) | 0.79 (0.01) | 0.78 (0.02) | 0.80 (0.05) | 0.78 (0.03) | 0.82 (0.03) | 0.79 | 0 |
| | MADD | N/A | 0.14 (0.02) | 0.17 (0.03) | 0.31 (0.08) | 0.31 (0.08) | 0.55 (0.22) | 0.30 | N/A |
| XGB | AUC | 0.73 (0.02) | 0.73 (0.02) | 0.73 (0.01) | 0.74 (0.03) | 0.73 (0.03) | 0.74 (0.05) | 0.73 | 0 |
| | F1 | 0.78 (0.01) | 0.78 (0.02) | 0.77 (0.02) | 0.79 (0.05) | 0.76 (0.04) | 0.80 (0.04) | 0.78 | 0 |
| | MADD | N/A | 0.12 (0.02) | 0.16 (0.03) | 0.28 (0.06) | 0.28 (0.09) | 0.52 (0.20) | 0.27 | N/A |
| DT | AUC | 0.70 (0.01) | 0.70 (0.01) | 0.70 (0.02) | 0.71 (0.04) | 0.71 (0.02) | 0.73 (0.03) | 0.71 | + 0.01 |
| | F1 | 0.74 (0.01) | 0.74 (0.01) | 0.73 (0.02) | 0.76 (0.05) | 0.74 (0.02) | 0.79 (0.04) | 0.75 | + 0.01 |
| | MADD | N/A | 0.16 (0.04) | 0.20 (0.06) | 0.39 (0.08) | 0.36 (0.05) | 0.62 (0.25) | 0.34 | N/A |
| LR | AUC | 0.62 (0.00) | 0.62 (0.00) | 0.61 (0.01) | 0.63 (0.04) | 0.61 (0.02) | 0.63 (0.04) | 0.62 | 0 |
| | F1 | 0.60 (0.00) | 0.60 (0.01) | 0.60 (0.01) | 0.60 (0.03) | 0.60 (0.02) | 0.62 (0.03) | 0.60 | 0 |
| | MADD | N/A | 0.11 (0.04) | 0.14 (0.07) | 0.24 (0.06) | 0.21 (0.05) | 0.45 (0.19) | 0.23 | N/A |
| KNN | AUC | 0.55 (0.01) | 0.55 (0.01) | 0.55 (0.01) | 0.55 (0.02) | 0.54 (0.03) | 0.58 (0.09) | 0.55 | 0 |
| | F1 | 0.59 (0.01) | 0.59 (0.01) | 0.59 (0.01) | 0.62 (0.06) | 0.57 (0.04) | 0.62 (0.10) | 0.60 | + 0.01 |
| | MADD | N/A | 0.15 (0.05) | 0.19 (0.07) | 0.28 (0.07) | 0.30 (0.05) | 0.47 (0.20) | 0.28 | N/A |

The MADD results also indicate that the models are quite fair, as they are much closer to 0 than to 2. For instance, for the Random Forest model, per-group MADD ranges from 0.14 to 0.55. For all groups other than Abroad, MADD is on the low side with low variance, meaning that the models generate similar probability distributions across the groups. The highest MADD values are always for the Abroad students, likely because this group is the least represented. Thus, there might not be enough predictions for Abroad students to compare their probability distributions against other groups.

Overall, we did not observe evidence of unfair predictions – even despite the fact that three groups (Mindanao, Visayas, and especially Abroad) had smaller numbers of students. Smaller samples have more variance and therefore might exhibit more bias, and although there was some indication of this for the Abroad group, it was not substantial.

Since the average grade (0.72) and the median grade (0.83) in the data set were relatively high, one could argue that teachers’ grading was more lenient during the mandated online learning, as reported in [27]. This skew of the grades, along with the similar grade distribution between the groups (see Figure 1), may explain the low variance across the groups. However, we obtained enough differentiation to build accurate models, so this limitation does not seem to apply.

These findings represent a contrast to the previous literature on bias in education; for example, the regional distribution of students mattered in [28], which discovered failures to generalize across urban, suburban, and rural populations across regions of two states in the US. The performance of predictive models trained on separate populations significantly differed from each other and also from the overall model. However, this effect pertained to a younger population (K-12 students rather than university students) and for a different prediction problem (affect detection rather than grades).

5. CONCLUSION AND FUTURE WORK

This paper assessed fairness of prediction models for grades in the context of online learning during COVID-19 lockdowns. The models were evaluated in a novel setting: for multiple groups located across the Philippines, with groups having very different numbers of students. Contrary to our expectations, the evaluation did not demonstrate the consequences of the digital divide in the Philippines. Machine learning models for predicting student grades worked comparably well for students all across the country (and abroad). Supplementary materials for this work are linked in Appendix B.

While this is a promising result, further research is needed to demonstrate the generalizability of these findings in different contexts in the Philippines and beyond. While the location did not influence the results here, it mattered in other past research for a different prediction task [28], so the EDM field needs to further investigate where it does and does not matter. Moreover, the smaller samples for some groups might not be entirely representative of the respective regions, hence, there is a need to replicate the results on larger samples.

Future work should also research whether bias manifests not only in groups defined along a single dimension (e.g., the location), but also across intersections of demographic variables [26, 42] (e.g., the student’s location and age). This deeper investigation can identify specific biases that might get lost or evened out in an overall evaluation.

6. ACKNOWLEDGMENTS

We thank Ezekiel Adriel D. Lagmay for data cleaning and anonymization. Valdemar Švábenský thanks Ateneo de Manila University for providing the LS Internationalization Grant that enabled him to visit and collaborate with the research group of Maria Mercedes T. Rodrigo. Méline Verger’s work was supported by Sorbonne Center for Artificial Intelligence (SCAI) and the *Direction du Numérique Educatif*.

7. REFERENCES

- [1] Z. Alamgir, H. Akram, S. Karim, and A. Wali. Enhancing Student Performance Prediction via Educational Data Mining on Academic data. *Informatics in Education*, 2023. <https://doi.org/10.15388/infedu.2024.04>.
- [2] H. Anderson, A. Boodhwani, and R. S. Baker. Assessing the Fairness of Graduation Predictions. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM)*, 2019.
- [3] Apache. Parquet, 2024. <https://parquet.apache.org/>.
- [4] A. Asselman, M. Khaldi, and S. Aammou. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31(6):3360–3379, 2023. <https://doi.org/10.1080/10494820.2021.1928235>.
- [5] L. Aulck, D. Nambi, N. Velagapudi, J. Blumenstock, and J. West. Mining University Registrar Records to Predict First-Year Undergraduate Attrition. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM)*, 2019. <https://eric.ed.gov/?id=ED599235>.
- [6] R. S. Baker, L. Esbenshade, J. Vitale, and S. Karumbaiah. Using Demographic Data as Predictor Variables: a Questionable Choice. *Journal of Educational Data Mining*, 15(2):22–52, 2023. <https://doi.org/10.5281/zenodo.7702628>.
- [7] R. S. Baker and A. Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1052–1092, 2021. <https://doi.org/10.1007/s40593-021-00285-9>.
- [8] R. S. Baker, E. Walker, A. Ogan, and M. Madaio. Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context*, 1(1):1–13, 2019. <https://doi.org/10.5281/zenodo.4057223>.
- [9] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. <http://www.fairmlbook.org>.
- [10] J. S. Barrot, I. I. Llenares, and L. S. Del Rosario. Students’ online learning challenges during the pandemic and how they cope with them: The case of the Philippines. *Education and information technologies*, 26(6):7321–7338, 2021. <https://doi.org/10.1007/s10639-021-10589-x>.
- [11] V. Bayer, M. Hlosta, and M. Fernandez. Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally? In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, editors, *Artificial Intelligence in Education*, pages 71–75, Cham, 2021. Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_12.
- [12] I. G. Clemen, H. Ali, A.-N. Abdulmadid, and J. H. Jabbar. Education During COVID-19 Era: Readiness of Students in a Less-Economically Developed Country for E-Learning. *IMCC Journal of Science*, 1(2):94–101, 2021. <http://hdl.handle.net/10419/234960>.
- [13] J. V. Cleofas and I. C. N. Rocha. Demographic, gadget and internet profiles as determinants of disease and consequence related COVID-19 anxiety among Filipino college students. *Education and Information Technologies*, 26(6):6771–6786, 2021. <https://doi.org/10.1007/s10639-021-10529-9>.
- [14] L. Cohausz, A. Tschalzev, C. Bartelt, and H. Stuckenschmidt. Investigating the Importance of Demographic Features for EDM-Predictions. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 125–136, Bengaluru, India, July 2023. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.8115647>.
- [15] A. Haim, R. Gyurcsan, C. Baxter, S. T. Shaw, and N. T. Heffernan. How to Open Science: Debugging Reproducibility within the Educational Data Mining Conference. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 114–124, Bengaluru, India, July 2023. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.8115651>.
- [16] A. Hellas, P. Ihanola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao. Predicting academic performance: A systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, page 175–199, New York, NY, USA, 2018. Association for Computing Machinery. <https://doi.org/10.1145/3293881.3295783>.
- [17] E. Hicks, A. Cook, K. Malasri, A. Zaman, and V. Phan. Keep It Relevant! Using In-Class Exercises to Predict Weekly Performance in CS1. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2022*, page 154–160, New York, NY, USA, 2022. Association for Computing Machinery. <https://doi.org/10.1145/3478431.3499357>.
- [18] M. Hoq, P. Brusilovsky, and B. Akram. Analysis of an explainable student performance prediction model in an introductory programming course. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 79–90, Bengaluru, India, July 2023. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.8115693>.
- [19] Q. Hu and H. Rangwala. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM)*, 2020. <https://eric.ed.gov/?id=ED608050>.
- [20] Instructure. Canvas LMS, 2024. <https://www.instructure.com/canvas>.
- [21] Y. Jang, S. Choi, H. Jung, and H. Kim. Practical early prediction of students’ performance using machine learning and eXplainable AI. *Education and information technologies*, 27:12855–12889, 2022. <https://doi.org/10.1007/s10639-022-11120-6>.
- [22] W. Jiang and Z. A. Pardos. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’21, page 608–617, New York, NY, USA, 2021. Association for Computing Machinery.

- <https://doi.org/10.1145/3461702.3462623>.
- [23] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*, pages 174–202. Routledge, 2022. <https://www.taylorfrancis.com/chapters/edit/10.4324/9780429329067-10/algorithmic-fairness-education-ren%C3%A9-kizilcec-hansol-lee>.
- [24] C. Korkmaz and A.-P. Correia. A review of research on machine learning in educational technology. *Educational Media International*, 56(3):250–267, 2019. <https://doi.org/10.1080/09523987.2019.1669875>.
- [25] H. Lee and R. F. Kizilcec. Evaluation of fairness trade-offs in predicting student success. In *FATED (Fairness, Accountability, and Transparency in Educational Data) Workshop at EDM 2020*, 2020. <https://arxiv.org/abs/2007.00088>.
- [26] L. Li, L. Sha, Y. Li, M. Raković, J. Rong, S. Joksimovic, N. Selwyn, D. Gašević, and G. Chen. Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. In *13th International Learning Analytics and Knowledge Conference, LAK'23*, page 499–508, New York, NY, USA, 2023. Association for Computing Machinery. <https://doi.org/10.1145/3576050.3576119>.
- [27] M. M. L. Moreno, M. M. T. Rodrigo, J. M. R. Torres, T. J. Gaspar, and J. A. Casano. Cura Personalis During COVID-19: Student and Faculty Perceptions of the Pedagogy of Compassion. *Computer-Based Learning in Context*, 2023. <https://doi.org/10.5281/zenodo.7256911>.
- [28] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3):487–501, 2014. <https://doi.org/10.1111/bjet.12156>.
- [29] A. Ogan, E. Walker, R. Baker, M. M. T. Rodrigo, J. C. Soriano, and M. J. Castro. Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, 25:229–248, 2015. <https://doi.org/10.1007/s40593-014-0034-8>.
- [30] N. Ong, J. Zhu, and D. Mosse. Towards Including Instructor Features in Student Grade Prediction. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 239–250, Durham, United Kingdom, July 2022. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.6853063>.
- [31] L. Paquette, J. Ocumpaugh, Z. Li, A. Andres, and R. Baker. Who’s Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, 12(3):1–30, 2020. <https://eric.ed.gov/?id=EJ1274018>.
- [32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [33] Philippine Statistics Authority. Directory, 2024. <https://psa.gov.ph/directory>.
- [34] N. Rohani, K. Gal, M. Gallagher, and A. Manataki. Early Prediction of Student Performance in a Health Data Science MOOC. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 325–333, Bengaluru, India, July 2023. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.8115721>.
- [35] H. Sahlaoui, E. A. A. Alaoui, S. Agoujil, and A. Nayyar. An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models. *Education and Information Technologies*, pages 1–37, 2023. <https://doi.org/10.1007/s10639-023-12007-w>.
- [36] Scikit-learn developers. StratifiedGroupKFold, 2024. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html.
- [37] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser. Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 98–109, Durham, United Kingdom, July 2022. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.6852964>.
- [38] S. Sy, A. Araneta, H. Rahemtulla, B. Carrasco, and S. Balgos. Mapping digital poverty in the Philippines using AI/Big Data and Machine Learning, 2021. <https://stories.thinkingmachin.es/mapping-digital-poverty-in-the-philippines/>.
- [39] M. Verger, S. Lallé, F. Bouchet, and V. Luengo. Is Your Model “MADD”? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 91–102. International Educational Data Mining Society, 2023. <https://doi.org/10.5281/zenodo.8115786>.
- [40] M. Waskom. seaborn.kdeplot, 2024. <https://seaborn.pydata.org/generated/seaborn.kdeplot.html>.
- [41] M. Wu, Q. Yu, S. L. Li, and L. Zhang. Geographic and gender disparities in global education achievement during the COVID-19 pandemic. *International Journal of Applied Earth Observation and Geoinformation*, 111:102850, 2022. <https://doi.org/10.1016/j.jag.2022.102850>.
- [42] Z. Xu, J. Olson, N. Pochinki, Z. Zheng, and R. Yu. Contexts Matter but How? Course-Level Correlates of Performance and Fairness Shift in Predictive Model Transfer. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, page 713–724, New York, NY, USA, 2024. Association for Computing Machinery. <https://doi.org/10.1145/3636555.3636936>.

APPENDIX

A. TECHNICAL DETAILS OF THE DATA COLLECTION FROM CANVAS LMS

The data collection took place over the period of three consecutive semesters in two academic years (AY):

- Second semester of AY 2020–2021, divided into two quarters, from February 10, 2021 (with the logging starting on March 9, 2021) to April 7, 2021, and then from April 12, 2021 to June 5, 2021.
- First semester of AY 2021–2022, which ran from August 26, 2021 to December 18, 2021.
- Second semester of AY 2021–2022, which ran from January 31, 2022 to May 28, 2022.

Figure 2 illustrates the data collection architecture for this project. Events from Canvas in JSON format were sent to a message queue for preprocessing and anonymization using open-source software created at the university (<https://github.com/zzenonn/canvas-data-collector>).

The data then underwent an Extract, Transform, and Load (ETL) process using Amazon Web Services tools. Finally, the data were converted to the format Parquet [3] for faster analysis and efficient storage.

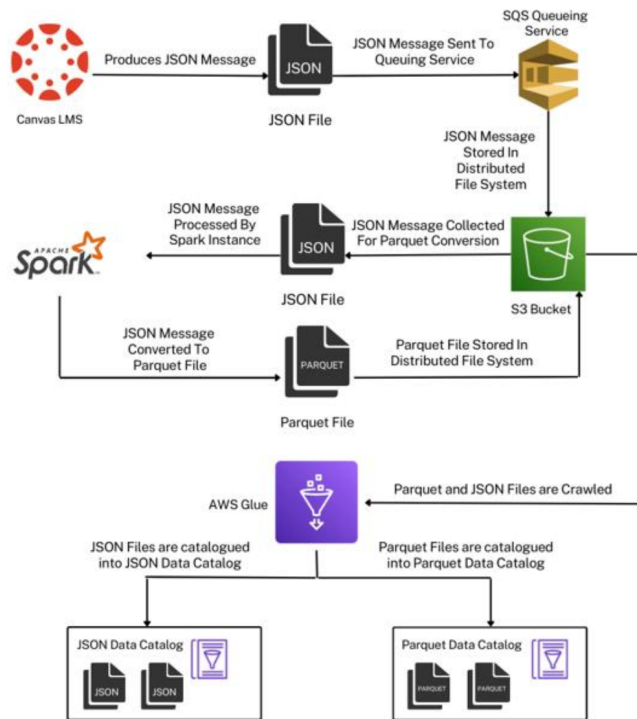


Figure 2: Architecture of the data collection from Canvas.

For this research, the raw data were subsequently exported as a CSV file. The size of this file is more than 7.6 GB, and it contains 15 columns and 48,740,270 rows (i.e., 8,142 entries on average per student), describing numerous activities and learning behaviors throughout the three semesters. Each action (one row in the CSV file) contains a timestamp with millisecond precision.

B. SUPPLEMENTARY MATERIALS

To support reproducibility within the EDM conference [15], our code is publicly available at:

<https://github.com/pcla-code/2024-edm-bias>.

We are unfortunately unable to share the data due to the regulations of the university in the Philippines.