Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/43/20230854

Exploring school factors of academic achievement in Macau: The application of educational data mining in Progress in International Reading Literacy Study (PIRLS) 2016 assessment

SHEN YUE

Faculty of Education, University of Macau, Macau, China

mc25766@um.edu.mo

Abstract. International large-scale assessments, provide structured and static data, and due to their extensive databases, they can be considered as a valuable resource for Big Data in Education. In this paper, we propose an educational data mining approach to detect and analyze factors linked to academic performance in Macau schools using data from the Progress in International Reading Literacy Study (PIRLS) 2016. We conducted a secondary data analysis based on a set of socioeconomic, process, and outcome variables from PIRLS and other sources, and built decision trees to obtain a predictive model of school performance. By doing so, we were able to identify the school and student-level variables that are most significant in predicting student performance in Macau. These findings will be useful for informing educational policy decisions and shedding light on the causes of poor performance in Macau schools. Overall, our study highlights the potential of educational data mining approaches in analyzing large-scale assessment data and generating insights for educational research and practice.

Keywords: Educational Data Mining, Supervised Learning, PIRLS 2016, School Factors.

1. Introduction

In Macau, an autonomous region known for its unique blend of Chinese and Portuguese cultures, educational practices and policies can have a significant impact on students' academic performance. There has been relatively limited research using the Progress in International Reading Literacy Study (PIRLS) data to conduct exploratory investigations into the various factors associated with their performance. While some studies have used Programme for International Student Assessment (PISA) data and focused on a limited number of variables [1], there remains a need to examine these factors from an integrative perspective.

According to Lee and Shute [2], academic achievement is influenced by a variety of intricate factors that need to be comprehensively considered. They proposed the Personal and Social Contextual Factors (PSCF) framework, which emphasizes the significance of personal factors (such as student engagement and learning strategies) and social-contextual factors (such as school climate and social-family influences) in analyzing student achievement, in addition to demographic variables. While previous studies, such as Jheng [3] examining grade repetition and Ieong et al. [4] investigating socioeconomic status, have identified certain factors that affect academic performance among Macau students, there is

DOI: 10.54254/2755-2721/43/20230854

limited knowledge about the relative contributions of these factors from an integrative perspective. Chen, Sakyi, and Cui [5] explored reading self-efficacy and reading achievement in relation to different contextual factors. They observed the strongest correlations between reading self-efficacy and home learning materials, as well as the school atmosphere. Similarly, Wang et al. [6] conducted research on Macau students' reading proficiency using PISA 2018 data and an integrative theoretical model. They discovered that personal factors, such as metacognitive strategies, enjoyment of reading, and perceived difficulty, were the most significant predictors of high reading performance among Macau students. These studies provide valuable insights into the variables that impact reading literacy in the Macau setting, shedding light on the complex interplay of factors that contribute to student achievement.

2. Literature Review

2.1. Large-Scale Assessments of Reading

Reading is fundamental to learning in school and is crucial for success in future work and community activities, as well as providing enjoyment through leisure reading [7]. PIRLS 2016 is an international program that assesses the reading performance of fourth-grade students in 50 countries and territories, including Macau. This study uses educational data mining techniques on PIRLS 2016 data to identify factors linked to academic performance in Macau schools, focusing on low-performing schools needing support to improve students' reading abilities.

2.2. Educational Data Mining and Large-Scale Assessments

Educational data mining (EDM) extracts insights from unstructured educational data using prediction, relationship mining, and structure finding techniques. These methods have gained popularity for examining massive databases and are increasingly considered as reliable alternatives to traditional inferential and multivariate statistics. [8] According to Papamitsiou, Z., and Economides, A. A. [9], educational data mining (EDM) can be utilized to explore various aspects of the teaching-learning process. This includes analyzing student performance, dropout and retention rates, feedback provided to students, as well as teacher and student reflection and awareness of the learning process. While most studies employing EDM techniques focus on higher education levels and online learning environments or Massive Open Online Courses (MOOCs), large-scale exams like PIRLS offer an opportunity to apply EDM approaches to less-explored student populations. These large-scale assessments provide valuable data for researchers to employ EDM approaches and investigate variables related to student performance. These studies often utilize predictor variables such as student and school background, educational practices, and non-cognitive student outcomes in an attempt to predict student performance across various competences, including reading, math, and science. [10] [11]

In this study, we will apply EDM techniques, particularly decision tree models, to analyze the PIRLS 2016 data from Macau schools. By doing so, we hope to identify factors at the school and student level that have a higher relevance in predicting academic performance in the Macau context. Based on this, we propose the following research questions:

To what extent can school factors extracted from the school questionnaire distinguish students with high PIRLS from students with low ability levels in Macao;

Which school variables contribute more to the explanation of academic performance in Macau;

What are the characteristics of the identified academically low-performing schools in terms of student demographics, school resources, and teaching practices in Macau?

3. Methods

3.1. Data Source

In 2016, PIRLS conducted a global reading literacy survey for fourth-grade students. As part of this study, 57 schools were selected in the Macau region, and a total of 4,059 students were included in the

DOI: 10.54254/2755-2721/43/20230854

initial sample for data analysis. Following data preprocessing, the student and school questionnaire datasets were cleaned for the 57 schools in the Macau region.

3.2. Variables

In this research, utilizing the conceptual framework, we will identify and extract variables at both the student and school levels from the student dataset and the school questionnaire dataset. These variables will be used to develop a model aimed at assessing the variations in abilities across different schools.

In addition, reading behaviour and attitudes will be assessed through questionnaires. Based on the students' responses to the tests, PIRLS will provide five plausible values (PV) for each student as a fair estimate of their reading ability.

Table 1. An overview of all the variables in the models based on the school level.

Variable Label	Description	Formation
School location	Describes the high urbanization of Macau as a single city.	ACBG05A, ACBG05B (ONLY one city, omit)
School composition by student socioeconomic background	Diversity in student socioeconomic backgrounds influencing achievement.	**ACBG03** (Derived variable)
Instruction affected by resource shortages (α=.910)	Quality of instruction affected by limited resources.	**ACBG12_INST** (ACBG12AA …ACBG12BD)
Teacher working conditions and job satisfaction (α=.812)	Work environment and satisfaction affecting teaching quality.	**ACBG13_TECH** (ACBG13A, … ACBG13E.)
School average reading self-efficacy (α=.897)	Average students' confidence in their reading abilities.	**ACBG17_Sch_eff** (ACBG17A, … ACBG17N.)
School emphasis on academic success	School prioritizes high achievement and excellence.	**ACDGEAS** (Background Scales)
Safe, orderly, and disciplined school	Maintaining a secure, well-ordered learning environment.	**ACDGDAS** (Background Scales)

Note. Adapted from "PIRLS 2016 User Guide for the International Database SUPPLEMENT 1" by International Association for the Evaluation of Educational Achievement [IEA], 2018.

School characteristics include factors like location, student socioeconomic composition, resource-related instruction impacts, and teacher satisfaction. These are derived from principal responses or existing data. The socioeconomic background of a school (school SES) is measured on a scale from 1 (affluent) to 3 (disadvantaged), with adjustments for clarity. The PIRLS 2016 school questionnaire focuses on school environment and organization, using indicators of school quality listed in Table 1. The average reading self-efficacy of schools represents principal leadership.

3.3. Prediction Model

Utilizing machine learning techniques on the PIRLS 2016 dataset, we develop a prediction model to analyze factors influencing academic performance. Decision trees are chosen because it offers several advantages that make it particularly suitable for large-scale datasets like PIRLS, which may contain a significant amount of missing values. They are easy to construct, can handle qualitative predictors without the need for dummy variables, and are relatively robust to missing data. Additionally, decision trees effectively capture complex interactions between variables and produce easily interpretable results.

3.4. Data Preprocessing

A student's score for the output variable will be determined by aggregating their five plausible values (PVs). PIRLS establishes four international benchmarks as standards for achievement: Advanced

DOI: 10.54254/2755-2721/43/20230854

International Benchmark (625 points), High International Benchmark (550 points), Intermediate International Benchmark (475 points), and Low International Benchmark (400 points). In our pre-processing, we defined students with scores higher than 550 as high-level performance, numbered as 1; The rest of the students were defined as low level performance and numbered as 0. Then input variables from the questionnaire are transformed into dummy variables, and variables (shown in Table 1) are employed for data description and analysis. During data cleaning, variables with a high percentage of missing values (>80%) are eliminated. Other missing values will be replaced by means. [12].

3.5. Model Training

Applying grid search with cross-validation, we split the dataset into a training set (80%) and a test set (20%), maintaining performance level ratios. The grid search method evaluates the performance of the training set using various parameter combinations for decision trees, ultimately identifying the best parameters for the highest performance. We utilize the "GridSearchCV" feature offered by the Scikit-learn package in Python for this process. [13]

3.6. Model Evaluation and Validation

After selecting the optimal parameters, the model is trained by fitting it to the entire training dataset. Then, we assess the model's performance on the test set to evaluate its ability to generalize. To ensure robustness and reliability, we employ the k-fold cross-validation technique, dividing the dataset into k equal parts and iterating k times. By averaging the performance across all iterations, we obtain an estimate of the overall model performance. Performance metrics such as accuracy, precision, recall, and F1 score are used to evaluate the model.

4. Discussion

4.1. Results

4.1.1. RQ1: To what extent can school factors extracted from the school questionnaire distinguish students with high PIRLS from students with low ability levels in Macao?

The decision tree (DT) model, trained on school factors extracted from the school questionnaire, was able to distinguish students with high PIRLS from those with low ability levels in Macao to a certain extent. The model achieved a training accuracy of 63.61% and a testing accuracy of 62.32%. (Table 2) Although the accuracy is moderate, it demonstrates that school factors do have an influence on student performance. The model also exhibited reasonable precision (64.69%), recall (59.76%), F-score (62.13%), and area under the curve (AUC) (69.70%), indicating that it was able to balance the identification of true positives and false positives, while also distinguishing between high and low ability students.

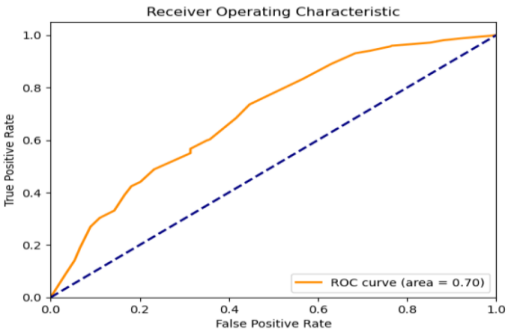

Figure 1. AUC values (the areas under the ROC curves) of the DT models.

DOI: 10.54254/2755-2721/43/20230854

Table 2. The training and testing performances of the DT model.

Training	Testing				
Accuracy(%)	**Accuracy(%)**	**Precision(%)**	**Recall (%)**	**F-Score(%)**	**AUC (%)**
63.61	62.32	64.69	59.76	62.13	69.70

4.1.2. RQ2: Which school variables contribute more to the explanation of academic performance in Macau?

Table 3 shows the most influential school variable on Macau's academic performance is school average reading self-efficacy with a coefficient of 0.4711, highlighting the role of students' belief in reading abilities. Next, resource shortages had a coefficient of 0.2998, indicating resource issues affecting education quality. Teacher conditions ranked third (0.1008), followed by student socioeconomic background (0.0605) and school environment factors. These suggest that student confidence, resources, teacher satisfaction, and school environment influence Macau's academic outcomes.

Table 3. The variables ranked according to their impact on the prediction accuracy of the model.

Ranking	Variables	Coefficient
1	ACBG17_Sch_eff	0.4711
2	ACBG12_INST	0.2998
3	ACBG13_TECH	0.1008
4	ACDG03	0.0605
5	ACDGDAS	0.0539
6	ACDGEAS	0.0139

4.1.3. RQ3: What are the characteristics of the identified academically low-performing schools in terms of student demographics, school resources, and teaching practices in Macau?

Academically low-performing schools in Macau are influenced by several interconnected factors including student demographics, school resources, and teaching approaches. These schools typically host students from lower socioeconomic backgrounds, which affects access to resources and learning opportunities. They may grapple with resource shortages that impact instructional quality. Issues related to teacher working conditions and job satisfaction can further compromise teaching quality. Moreover, such schools might not emphasize academic success as strongly and may lack a disciplined environment, contrasting high-performing schools. In essence, a blend of demographic, resource, and teaching factors defines these schools' performance.

Based on Chen's work [5], school climate emerges as a stronger predictor of academic success in reading than socioeconomic status (SES), highlighting the critical role of schools in creating a positive and supportive learning environment. These findings support our previous analyses and underscore the importance of contextual factors in shaping educational outcomes.

4.2. Limitation

This study's limitations arise from its dependence on PIRLS 2016 cross-sectional data, which does not allow for establishing direct causality. Also, some variables, like school factors, are based on self-reported data, leading to potential biases. Future studies might benefit from using objective measures to validate findings.

5. Conclusion

The findings of this research illuminate the significant role of school factors, especially reading self-efficacy, in determining academic performance in Macau. The importance of resource availability, teacher conditions, and school environment can't be overstated. Moreover, the research reveals that school climate, even more than socioeconomic status, can be a key predictor of academic success. It's

DOI: 10.54254/2755-2721/43/20230854

essential to consider these factors when formulating educational policies or interventions aimed at enhancing student performance in Macau. Future studies, especially longitudinal ones, are recommended to delve deeper into these relationships and provide clearer insights.

References

[1] Karadağ, E. (Ed.). (2017). The factors effecting student achievement: Meta-analysis of empirical studies. Springer.

[2] Lee, J., & Shute, V. J. (2010). Personal and social-contextual factors in K-12 academic performance: An integrative perspective on student learning. Educational Psychologist, 45(3), 185–202. https://doi.org/10.1080/00461520.2010.493471

[3] Jheng, Y. J. (2014). Does Grade Repetition Work? Who Repeats Grades? Evidence from the Scores of Reading Literacy of PISA 2009 of Macau. Jiaoyu Yanjiu Yuekan Journal of Education Research, 242, 97-111. https://doi.org/10.3966/168063602014060242007

[4] Ieong, M. K., Cheung, K. C., Sit, P. S., Soh, K. C., & Mak, S. K. (2016). Effects of economic, social and cultural status on mathematics performance: A multilevel mediation analysis of self-regulated learning processes. Contemporary Educational Research Quarterly, 24(4), 109–143. https://doi.org/10.6151/CERQ.2016.2404.05

[5] Chen, F., Sakyi, A., & Cui, Y. (2021). Linking student, home, and school factors to reading achievement: the mediating role of reading self-efficacy. Educational Psychology, 41(10), 1260-1279. https://doi.org/10.1080/01443410.2021.1953445

[6] Wang, K., Haw, J., & Leung, S. On. (2023). What explains Macau students' achievement? An integrative perspective using a machine learning approach (¿Cuál es la explicación del rendimiento de los estudiantes macaenses? Una perspectiva integradora mediante la adopción del enfoque del aprendizaje automático). Infancia y Aprendizaje, 46(1), 71–108. https://doi.org/10.1080/02103702.2022.2149120

[7] Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). PIRLS 2011 International Results in Reading. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

[8] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61-75). Springer.

[9] Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

[10] Ayodele, M. O., Julius, A., & Ayeni, M. F. (2014). Predictive Power of Selected Variables on Students' Academic Achievement in Integrated Science. Available at: https://www.researchgate.net/publication/313329752_Predictive_Power_of_Selected_Variables_on_Students'_Academic_Achievement_in_Integrated_Science

[11] Pratama, D., & Husnayaini, I. (2022). Program for international student assessment (PISA) analysis of Asian countries using k-mean clustering algorithms. JISAE: Journal of Indonesian Student Assessment and Evaluation, 8(1), 35–44. https://doi.org/10.21009/jisae.v8i1.25445

[12] Chen, J., Zhang, Y., & Hu, J. (2021). Synergistic effects of instruction and affect factors on high- and low-ability disparities in elementary students' reading literacy. Reading and Writing, 34, 199-230. https://doi.org/10.1007/s11145-020-10120-3

[13] Mou, W. J., Liu, Z. Q., Luo, Y., Zou, M., Ren, C., Zhang, C. Y., & Tian, Y. P. (2014). Development and cross-validation of prognostic models to assess the treatment effect of cisplatin/pemetrexed chemotherapy in lung adenocarcinoma patients. Medical Oncology, 31, 59. https://doi.org/10.1007/s12032-014-0059-8