MACHINE LEARNING MODEL FOR PREDICTING PRIMARY SCHOOL SCORES BASED ON SPATIAL, SOCIO DEMOGRAPHIC AND SCHOOL–RELATED INFORMATION

Felipe Lillo, Leidy García and Pedro Severino-González

SUMMARY

Learning strategies at primary school level are important to ensure student progress. In this regards, the identification of those factors influencing students grades certainly help teachers in predicting outcomes as well as in improving teaching strategies. This study empirically investigates connections between socio-demographic, school–related and academic features of primary students. Special attention is given to spatial features and how they influence performance. In particular the Euclidean distance from city center. The research method is based on machine learning techniques which are developed from a dataset consisting of 12159 primary school students living in the city of Talca, Chile. Four machine learning models are tested: a Neural Network (NN), Random Forest (RF) a Support Vector Machine (SVM) and a Gradient Boosted Tree model (GBT). Results show similar error levels between models and confirms student age, school capacity and school distance as important determinants of score predictions.

Introduction

The adoption of learning among children has been an important topic of study in the literature, and has been addressed from many angles, given the multiplicity of factors and the multidimensionality of the variables which affect them. In the Chilean case, the evidence shows significant inequality in cognitive development (Contreras and Puentes, 2017) and in school results among children and young people (Canales and Webb, 2018; Treviño *et al*., 2016). This is worrisome considering that inequality in education translates in the long term into unequal development of human capacities (Heckman, 2011). Furthermore, achievement gaps tend to be persistent and require better policy formulation (Hanushek *et al*., 2019). The Chilean case is an interesting case study given the presence of high school segregation by income (Treviño *et al*., 2016). In this regard, Hofflinger *et al*. (2020) show that parents enroll their children according to their socioeconomic level, which has led to the fact that the majority of students who attend municipal or public schools are highly vulnerable. It is thus important to identify the factors that influence school achievement in municipal schools, considering the spatial effect, and particularly adopting more recent measurement methodologies, such as machine learning (ML).

Mjolsness and DeCoste (2001) anticipated the proliferation of machine learning (ML) in a variety of scientific fields. Currently, ML–techniques are making important progress across disciplines, providing new insights and ways to automate different human tasks (Bertolini *et al*., 2021). Education is one field where ML collaborates in the development of learning strategies based on student academic performance (Baashar *et al*., 2021). In Kumar *et al*. (2021) academic performance is defined as the knowledge obtained by the student, and is mainly affected by social, economic, personal, psychological and environmental factors. One of the most widely used measures is scores on standardized tests. In this regard, educators consider it very important to anticipate student scores since this helps with implementing efficient learning strategies (Yip, 2021).

Predicting students' academic performance poses an interesting question regarding the factors that determine such prediction and their influence on a corresponding ML–model. The analysis of this question is the overall subject of this research. Recent literature addressing ML–models for student performance prediction is abundant. A significant body of papers has focused on higher education students (Yakubu and Abubakar, 2021; Oreshin *et al*., 2020; Aydoğdu, 2020; Xu *et al*., 2019; Hasnine *et al*., 2018; Kotsiantis, 2012). These papers mostly predict performance measures based on GPA from both sociodemographic and academic information. On the other hand, a more limited group of articles tackles performance prediction of primary or secondary students (Rajendran, 2021; Tarik *et al*., 2021; Costa-Mendes *et al*., 2021; Alam *et al*., 2021; Chui *et al*., 2020; Fernandes *et al*., 2019; Qazdar *et al*., 2019;). These studies mostly include prediction features similar to

KEYWORDS / Machine Learning / Primary Education / Score Prediction / Spatial Features /

Received: 11/27/2023. Modified: 01/23/2024. Accepted: 01/25/2024.

Felipe Lillo-Viedma. Graduate in Engineering Sciences and Industrial Civil Engineer, Universidad del Bío-Bío (UBB), Chile. PhD in Computer Science and Mathematics, AUT - University, New Zealand. Professor, Faculty of Basic Sciences, Universidad Católica del Maule (UCM), Chile.

Leidy García. Bachelor of Economics, Universidad de Antioquia, Colombia. Master in Business Administration, Universidad de Talca, (UTalca) Chile. PhD in Economics, Universidad de Chile, Chile. Professor, Faculty of Economics and Business, UTalca, Chile.

Pedro Severino-González (Corresponding author). Graduate in Administrative Sciences, Commercial Engineer and Master in Business Administration, UBB, Chile. Professor, UCM, Chile. Address: Department of Economics and Administration, Faculty of Social and Economic Sciences, UCM. Av. San Miguel 3605, Talca, Chile. e-mail: pseverino@ucm.cl.

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LOS RESULTADOS DE LA ESCUELA PRIMARIA BASADOS EN LA INFORMACIÓN ESPACIAL, SOCIODEMOGRÁFICA Y RELACIONADA CON LA ESCUELA

Felipe Lillo, Leidy García y Pedro Severino-González

RESUMEN

Las estrategias de aprendizaje en la enseñanza primaria son importantes para garantizar el progreso de los alumnos. En este sentido, la identificación de los factores que influyen en las calificaciones de los alumnos ayuda sin duda a los profesores a predecir los resultados y a mejorar las estrategias de enseñanza. Este estudio investiga empíricamente las conexiones entre las características sociodemográficas, escolares y académicas de los alumnos de primaria. Se presta especial atención a las características espaciales y a cómo influyen en el rendimiento. En particular, la distancia euclidiana al centro de la ciudad. El método de investigación se basa en técnicas de aprendizaje automático que se desarrollan a partir de un conjunto de datos formado por 12159 estudiantes de primaria que viven en la ciudad de Talca, Chile. Se prueban cuatro modelos de aprendizaje automático: una Red Neuronal (RN), un Bosque Aleatorio (RF por sus siglas en inglés), una Máquina de Vectores Soporte (SVM, por sus siglas en inglés) y un modelo de Árbol Impulsado por Gradiente (GBT, por sus siglas en inglés). Los resultados muestran niveles de error similares entre los modelos y confirman la edad del estudiante, la capacidad de la escuela y la distancia de la escuela como determinantes importantes de las predicciones de puntuación.

MODELO DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE RESULTADOS DE ESCOLAS PRIMÁRIAS COM BASE EM INFORMAÇÕES ESPACIAIS, SOCIODEMOGRÁFICAS E RELACIONADAS À ESCOLA

Felipe Lillo, Leidy García e Pedro Severino-González

RESUMO

As estratégias de aprendizagem no ensino fundamental são importantes para garantir o progresso dos alunos. Nesse sentido, a identificação dos fatores que influenciam as notas dos alunos certamente ajuda os professores a prever os resultados, bem como a aprimorar as estratégias de ensino. Este estudo investiga empiricamente as conexões entre características sociodemográficas, relacionadas à escola e acadêmicas dos alunos do ensino básico. É dada atenção especial às características espaciais e como elas influenciam o desempenho. Em particular, a distância euclidiana do centro da cidade. O método de pesquisa baseia-se em técnicas de aprendizado de máquina desenvolvidas a partir de um conjunto de dados composto por 12.159 alunos do ensino fundamental que vivem na cidade de Talca, no Chile. Quatro modelos de aprendizado de máquina são testados: uma rede neural (NN, por seu acrônimo em inglês), uma floresta aleatória (RF, por seu acrônimo em inglês), uma máquina de vetor de suporte (SVM, por seu acrônimo em inglês) e um modelo de árvore com reforço de gradiente (GBT, por seu acrônimo em inglês). Os resultados mostram níveis de erro semelhantes entre os modelos e confirmam a idade do aluno, a capacidade da escola e a distância da escola como determinantes importantes das previsões de pontuação.

---

those found in higher education studies. However, psychological traits are also included in the analysis.

The prediction of drop-out levels in the school system is another aspect addressed by ML–models. Studies carried out by Rodríguez *et al*. (2023), Rodríguez and Villanueva, (2022) and Smith and Gutiérrez (2022) tackle this aspect in the Chilean case. These works evidence a research focus on ML-models for predicting desertion. However, scarce consideration is given to the identification of those features that significantly determine such desertion.

Several reviews are also found which focus mainly on the role of ML in any education, performance of ML-models or identification of predictor features (Luan and Tsai, 2021; Baashar *et al*., 2021; Albreiki *et al*., 2021; Alamri and Alharbi, 2021; Turabieh *et al*., 2021; Namoun and Alshanqiti, 2020; Rastrollo-Guerrero *et al*., 2020; Hellas *et al*., 2018). Regarding the latter, these reviews identify a significant range of predictors, which can be loosely grouped in three main categories: academic, socio–demographic, and behavioral.

One aspect barely addressed by the literature is performance predictors based on spatial information (Murphy, 2019; Gordon and Monastiriotis, 2007). This paper contributes to the research problem by analyzing how some spatial metrics (as well as some sociodemographic features) influence academic performance prediction. Specifically, a spatial metric related to distances is considered (the Euclidean distance). Information about grades, schools and socio–economics is gathered from secondary sources (students from a Chilean city). This information is employed to build four ML–models, and assess the importance of predictors.

Information on the materials and methods used, the results obtained with the instruments and the models are presented. Finally, the analysis of the results, conclusions and suggestions for future research.

Methods

Research focus and design

A quantitative research approach with a descriptive focus was selected to accomplish the objective. Primary information was cross-sectionally gathered in 2019. Such information is studied via techniques based in Machine Learning.

Data

The sample corresponds to a secondary information database provided by the Department of Municipal Education (DAEM) of Chile, for the year 2017. This includes 12,159 basic education children who attended the first to eighth grade levels (Elementary and middle school). and corresponds to 99.3% of the total children enrolled. The sample coverage by establishment considers 42 municipal educational establishments of the Municipality of Talca (87.5% of the total number of schools). That is, the sample is representative of the municipal school population.

The database considers the score variables of the standardized tests which evaluate the Level of Achievement of the Expected Learning and Skills based on the Current Curricular Framework of the Ministry of Education of Chile. These are for the areas of mathematics and language (reading comprehension) and are evaluated on a scale of 0 to 30 points, with 30 being the highest score. These tests are carried out by the DAEM with the support of experts and under the theories of response to the item and equating, guaranteeing their comparability. We also considered control variables corresponding to the socioeconomic characterization of the children such as gender, age, maternal educational level, household income level, and test scores. This also includes both sociodemographic and spatial school information (school distances from the city center) for each student.

Three categories classify student features: Socio–demographic, School–related, and Academic. These are composed by numeric as well as categorical variables. Table I describes variables and features utilized in the analysis. Special attention is given to the School distance, which is computed for each school compared to the latitude and longitude of the Talca city council (decimal degrees). The Euclidean metric is utilized in this case. Figure 1 shows schools- spatial distribution, where the yellow dot pins the city council location.

Analysis techniques

A Machine learning approach assesses the relation between socio–demographics, school related variables and academic scores. In particular, four machine learning models are analyzed: a Neural Network model (NN), Support Vector Machine (SVM), Random Forest model (RF) and Gradient Boosted Tree (GBT). The purpose of the models is to predict student scores for both mathematics and language. The average score between these subjects is also considered for prediction.

The corresponding dataset is pre-processed so that all categorical input variables are encoded by using One–Hot encoding. Hence, dummy variables are created for those variables containing several

TABLE I
DIFFERENTIALLY EXPRESSED PROTEINS IN MILLET ROOT FERTILIZED WITH SI VS. WITHOUT SI. CELAYA, GTO. S-S, 2019

Category	Feature	Variable Type	Encoding
Socio-demographic	Gender	Categorical	0: Male 1: Female
	Mother educational	Categorical level	1: No education 2: Pvt-Sec Incomplete 3: Primary complete 4: Secondary complete 5: Higher education
	Income level	Categorical	1: First quintile 2: Second quintile 3: Third quintile
	Age	Numerical	Discrete (years)
School related	Number of teachers	Numerical	Discrete
	School grade	Numerical	Discrete
	Capacity usage	Numerical	Continuous
	School latitude	Numerical	Continuous
	School longitude	Numerical	Continuous
	School distance	Numerical	Continuous
Academic	Score mathematics	Numerical	Discrete
	Score language	Numerical	Discrete

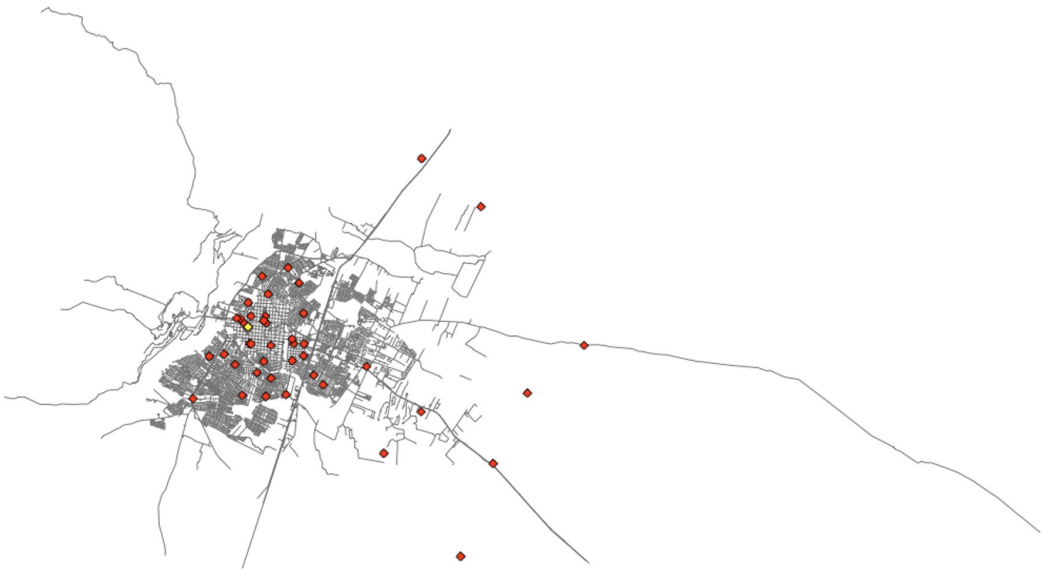

Figure 1. Schools' spatial distribution.

categories. Since missing values also concentrate mainly in age and academic scores features, these are replaced by the corresponding averages per grade and school, respectively. Finally, latitude and longitude are dropped since they are only used to compute school distances.

To identify multicollinearity issues, correlation coefficients between numeric predictive features (Socio-demographic and School related) are computed. This addresses the need for feature reduction.

The NN model is constructed by assigning feature variables as the input layer, while the output layer corresponds to scores on mathematics as well as language subjects. The number of hidden layers is 3, and the number of nodes in each layer is 200. The activation function used in these layers is the "Relu" function, a normal kernel initializer is also employed, and "Linear" is used as the activation function of the last layer. The loss function of the NN model is mean squared error (MSE) and the objective function is optimized using the Adam optimizer. The mean absolute error is also computed. Moreover, 200 epochs are considered and the model is fitted by splitting train and validation dataset in 0.8 and 0.2, respectively (hereafter, all models are split in the same way).

The SVM model is set as a regressor model, since numeric variables are estimated. The main parameter in this model is the kernel function which is set to Radian Basis Function (RBF).

RF is a machine learning method for either classification or regression, and works by constructing decision trees from a training data set (Ho, 1995). Some advantages of RF are in handling regression and classification, the need for a small number of tuning parameters, suitability for complex problems, avoiding overfitting, estimation of variable importance, and ability for accurately dealing with small size problems (Zhang and Ma, 2012). The performed random forest test has 60 trees (number of estimators). Both Mean Squared Error (MSE) and Mean Absolute Error (MAE) are also computed to analyze performance. Variable importance is estimated via Permutation Feature Importance (PFI), which is an approach that normalizes a biased measure based on a permutation tree test and returns the statistical significance for each feature (Altmann *et al*., 2010). In this case, the significance is measured by the R^2 (coefficient of determination) regression score function.

GBT provides a prediction modelling approach as an ensemble of weak prediction models, which are commonly decision trees (Hastie *et al*., 2009). A gradient-boosted tree model is built in a stage-wise form, and generalizes other boosting methods by allowing optimization of an arbitrary differentiable loss function. For implementation, the presented GBT model considers a maximum tree depth of 10 for base learners while other parameters remain at default.

The NN and SVM models are implemented by Keras and Sklearn Python libraries. "RandomForestRegressor" and "XBGRegressor" are also implemented for RF and GBT models, respectively. Finally, the "PermutationImportance" package is imported to estimate feature ranking in both tree models (scoring = r2 and number of iteration = 5).

Results

Dataset summary

The categorical features for the dataset are statistically summarized by Figure 2. Female students overpass male students. Student mothers have mostly completed their primary school. Families are mainly part of the second income quintile. Table II shows a

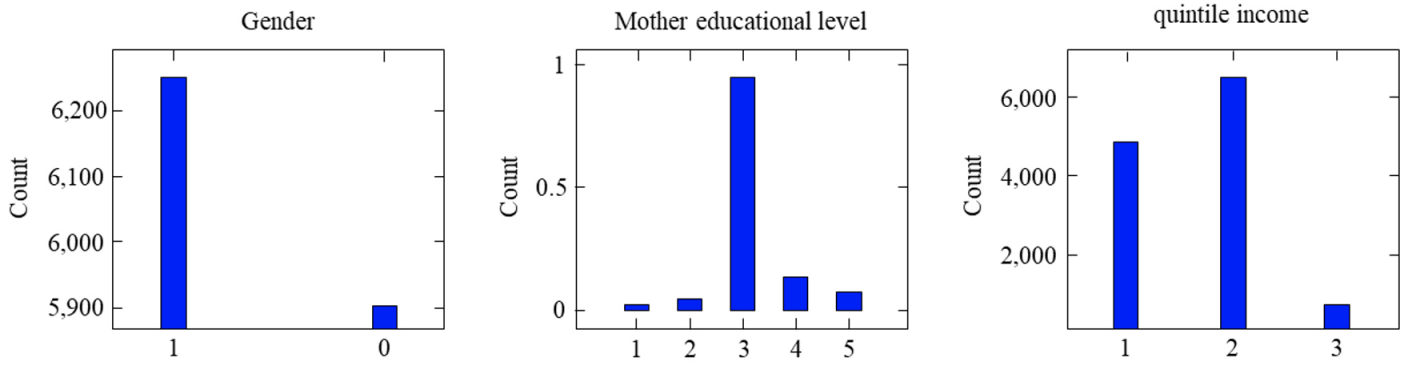

Figure 2. Schools' spatial distribution.

TABLE II
NUMERIC VARIABLE SUMMARY

	School-distance	School grade	Language-score	Math-score	Num-teachers	Capacity usage	Age
mean	2.254	4.156	14.786	14.893	36.085	0.681	9.617
std	2.334	2.023	5.015	5.990	14.060	0.141	2.099
min	0.432	1.000	0.000	0.000	4.000	0.246	6.351
25%	0.980	2.000	11.000	10.000	26.000	0.585	7.356
50%	1.794	4.000	15.000	14.000	32.000	0.746	9.479
75%	2.897	6.000	18.000	19.000	47.000	0.770	11.520
max	13.442	8.000	28.000	30.000	80.000	0.991	13.800

descriptive summary of numeric variables from the data set. Furthermore, input feature correlation is computed to identify highly interdependent numeric variables. Table III shows the corresponding results. As a result, "School Grade" is removed due to being highly correlated with "Age".

Machine learning models

Results associated to the machine learning regression models appear in Table IV. The models perform very similarly in each predictive feature. Nonetheless, errors diminish when Average Score is computed. Scores are evenly concentrated around the mean, which improves models- performance.

Finally, feature importance is computed from the tree models. Figure 3 shows feature importance score for each sociodemographic characteristic based on permutation importance for RF and GBT models. Variables "mel" and "qin" are a result of the One-Hot encoding performed on features "mother educational level" and "quintile income", respectively. As described, "Age" and "Capacity usage" rank as the most determinant predictors for all output variables.

Discussion

ML regression models could reasonably predict student scores from different input features. In general, the models tested in this work showed homogeneous error estimates. This could be partially explained by the convenient sizes of both the training and testing set (Shalev-Shwartz and Srebro, 2008). The good performance of the SVM model agrees with arguments presented by Somvanshi *et al.* (2016) and Mahesh (2020) regarding the easy implementation of these models for either predicting or classifying

TABLE III
FEATURE CORRELATION

	Num-teachers	Capacity usage	School distance	Age	School grade
Num-teachers	1.000	0.231	-0.359	0.044	0.044
Capacity usage	0.231	1.000	-0.194	-0.038	-0.038
School-distance	-0.359	-0.194	1.000	-0.029	-0.029
Age	0.044	-0.038	-0.029	1.000	1.000
School grade	0.044	-0.038	-0.029	1.000	1.000

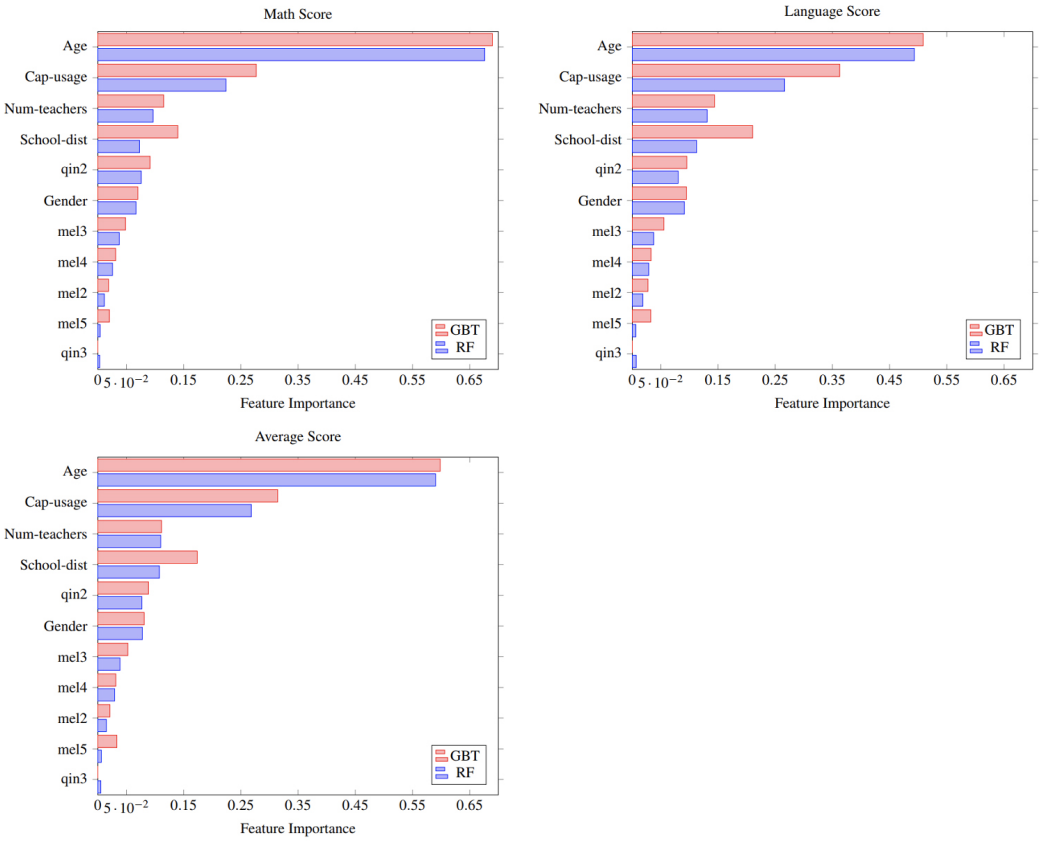

Figure 3. Feature Importance according to RF and GBT.

TABLE IV
ERRORS FOR ML-MODELS

Prediction feature	NN		RF		GBT		SVM	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Math Score	0.0296	0.1373	0.0291	0.1362	0.0291	0.1363	0.0290	0.1365
Score language	0.0266	0.1293	0.0263	0.1285	0.0262	0.1283	0.0261	0.1296
Average Score	0.0222	0.1185	0.0219	0.1171	0.0221	0.1172	0.0223	0.1194

NN: Neural Network, RF; Random Forest, SVM: Support Vector Machine, GBT: Gradient Boosted Tree model, MSE: Mean Squared Error, MAE: Mean Absolute Error.

attributes. Although models could be improved by applying some hyperparemeter tunning technique, this is not applied since is beyond the scope of this work.

From the RF–model, feature importance shows "Age" and "Capacity usage" as important score predictors. The results on the importance of age as a predictor of school achievement are consistent with the literature (Peng *et al*., 2019; Selmeczy *et al*., 2021).

The literature is inconsistent regarding the capacity of use as a variable equivalent to the size of the school. On the one hand, authors such as Schwartz *et al*. (2013) point out that smaller schools have better results. On the other hand, Ross (2019) showed both positive and negative school size effects depending on students' characteristics. They found that class size does not significantly affect school performance in Chilean children, but the size of the class is not exactly the used capacity of the establishments, and the latter is associated more with the available resources. Therefore, it is to be expected that the more capacity and resources schools have, the higher their student achievement should be.

We also found an important difference between school achievement depending on the location of the educational establishment, in our case measured as the distance to the city center. According to Li and Qiu (2018) and Broer *et al*. (2019), this could be explained because more rural areas tend to have fewer resources and families with lower socioeconomic characteristics, which affect test performance. However, this is not the most determinant feature in our case.

We therefore show that the learning of children in public schools in Chile, measured in test scores, is influenced by the characteristics of the schools, especially by use capacity and the distance to the city center. This provides new background on the need to formulate policies that reduce schools' educational and spatial segregation. This is relevant considering that educational institutions reflect segregated neighborhoods and, in turn, the education that children receive reinforces segregation (Owens, 2020). Nonetheless, the study timeframe is a limitation that need to be addressed by future studies. In particular, the Covid-19 pandemic certainly influenced student academic performances so that future work could compare how the analyzed determinants had been affected by the Covid-19 phenomenon.

Conclusions

This paper addressed the estimation of primary student scores from sociodemographic, academic, and school–related features from an empirical perspective. To attain this, four machine learning models were developed and evaluated.

The study showed a similar behavior of the ML-models, which could be further improved by applying hyperparameter tunning methods. In particular, RF models provided a feature importance ranking that helped determine the "influence" of each feature on the predicted variable. According to this ranking, school—related features are predominant determinants. In particular, school distance from the city center arises as an important determinant. Student age is at the top, though which is consistent with some findings in the literature. This ranking tool provided by RF-models is helpful for educators, since both learning and evaluation strategies can be developed based on knowledge from this ranking. Future work can experimentally address the performance of evaluation instruments designed from student sociodemographic features. More understanding is also needed about how spatial variables and learning outcomes are related.

Finally, the use of machine leaning in education is a promising research area, since prediction and classification are important for educational policy development. In particular, the authors emphasize Tree Base Models such as RF and GBT since they can provide more insight.

Glossary

– Socio-demographic features: social and demographic characteristics associated with students.

– School–related features: characteristics associated with either school infrastructure or resources.

Academic features: characteristics associated with student academic performance.

– Spatial features: characteristics associated with a specific geographical location.

– Euclidean distance: distance metric that computes the length of a straight line joining two points.

– Machine learning model: a model that is built specifically from empirical data for either predicting or classifying.

– One-hot encoding: technique for handling categorical variables. It converts categorical data into numerical data.

– School capacity: school capacity usage measured as percentage.

Ethics Statement

The study considers secondary information from schoolchildren; it was reviewed and approved by the Ethics Committee of the University of Talca, ID. 39-2021.

Data Statement

The database described in the methodology section is not publicly available since the study subjects are minors. On the other hand, all data from the result section can be fully employed by other studies.

ACKNOWLEDGMENTS

This research was supported by the Chilean National Agency for Research and Development (ANID) under Fondecyt Regular number 1210450 of the year 2021.

REFERENCES

Alam TM, Mushtaq M, Shaukat K, Hameed IA, Umer Sarwar M, Luo S (2021) A Novel Method for Performance Measurement of Public Educational Institutions Using Machine Learning Models. *Applied Sciences 11*: 9296. https://doi.org/10.3390/app11199296

Alamri R, Alharbi B (2021) Explainable Student Performance Prediction Models: A Systematic Review. *IEEE Access 9*: 33132–33143. https://doi.org/10.1109/access.2021.3061368

Albreiki B, Zaki N, Alashwal H (2021) A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences 11*: 552. https://doi.org/10.3390/educsci11090552

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: A corrected feature importance measure. *Bioinformatics 26*: 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Aydoğdu Ş (2020) Predicting student final performance using artificial neural net-works in online learning environments. *Education and Information Technologies 25*: 1913–1927. https://doi.org/10.1007/s10639-019-10053-x

Baashar Y, Alkawsi G, Ali N, Alhussian H, Bahbouh HT (2021) *Predicting student's performance using machine learning methods: A systematic literature review*. At the 2021 International Conference on Computer & Information Sciences (ICCOINS). Kuching, Malasia. July 13-15, pp. 357–362 https://doi.org/10.1109/ICCOINS49721.2021.9497185

Bertolini M, Mezzogori D, Neroni M, Zammori F (2021) Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications 175*: 114820. https://doi.org/10.1016/j.eswa.2021.114820

Broer M, Bai Y, Fonseca F (2019) A Review of the Literature on Socioeconomic Status and Educational Achievement. *Socioeconomic inequality and educational outcomes*. Springer. Cham. New York, USA. pp. 7–17.

Canales A, Webb A (2018) Educational Achievement of Indigenous Students in Chile: School Composition and Peer Effects. *Comparative Education*

Review 62: 231–273. https://doi.org/10.1086/696957

Chui KT, Liu RW, Zhao M, De Pablos PO (2020) Predicting Students' Performance with School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine. *IEEE Access 8*: 86745–86752. https://doi.org/10.1109/access.2020. 2992869

Contreras D, Puentes E (2017) Inequality of Opportunities at Early Ages: Evidence from Chile. *The Journal of Development Studies 53*: 1748–1764. https://doi.org/10.1080/00220388.2016.1262025

Costa-Mendes R, Oliveira T, Castelli M, Cruz-Jesus F (2021) A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies 26*: 1527–1547. https://doi.org/10.1007/s10639-020-10316-y

Fernandes E, Holanda M, Victorino M, Borges V, Carvalho R, Erven GV (2019) Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research 94*: 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Gordon I, Monastiriotis V (2007) Education, Location, Education: A Spatial Analysis of English Secondary School Public Examination Results. *Urban Studies 44*: 1203–1228. https://doi.org/10.1080/00420980701302387

Hanushek EA, Peterson PE, Talpey LM, Woessmann L (2019) The Unwavering SES Achievement Gap: Trends in U.S. Student Performance. *NBER Working Paper w25648*. https://doi.org/10.2139/ssrn.3357905

Hasnine MN, Akcapinar G, Flanagan B, Majumdar R, Mouri K, Ogata H (2018) *Towards final scores prediction over clickstream using machine learning methods*. At the 26th International Conference on Computers in Education Workshop Proceedings, APSCE. Metro Manila, November 26-36. pp. 399–404. http://hdl.handle.net/2433/237325

Hastie T, Tibshirani R, Friedman J (2009) Boosting and additive trees. *The Elements of Statistical Learning*. Springer, New York, USA. pp. 337–387.

Heckman JJ (2011) The Economics of Inequality: The Value of Early Childhood Education. *American Educator 35*: 31–35.

Hellas A, Ihantola P, Petersen A, Ajanovski VV, Gutica M, Hynninen T, Knutas A, Leinonen J, Messom C, Liao SN (2018) *Predicting academic performance: a systematic literature review*. At the ITiCSE '18: 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education. Larnaca, July 2-4. pp. 175–199. https://doi.org/10.1145/3293881.3295783.

Ho TK (1995) *Random decision forests*. At the Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal, August 14-16. pp. 278-282. https://doi.org/10.1109/ICDAR.1995.598994

Hofflinger A, Gelber D, Cañas ST (2020) School choice and parents' preferences for school attributes in Chile. *Economics of Education Review 74*: 101946. https://doi.org/10.1016/j.econedurev.2019.101946

Kotsiantis SB (2012) Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review 37*: 331–344. https://doi.org/10.1007/s10462-011-9234-x

Kumar S, Agarwal M, Agarwal N (2021) Defining and Measuring Academic Performance of Hei Students-A Critical Review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12*: 3091–3105.

Li Z, Qiu Z (2018) How does family background affect children's educational achievement? Evidence from Contemporary China. *The Journal of Chinese Sociology 5*: 1–21. https://doi.org/10.1186/s40711-018-0083-8

Luan H, Tsai CC (2021) A Review of Using Machine Learning Approaches for Precision Education. *Educational Technology & Society 24*: 250–266.

Mahesh B (2020) Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR) 9*: 381–386.

Mjolsness E, DeCoste D (2001) Machine Learning for Science: State of the Art and Future Prospects. *Science 293*: 2051–2055.

Murphy S (2019) School location and socioeconomic status and patterns of participation and achievement in senior secondary mathematics. *Mathematics Education Research Journal 31*: 219–235. https://doi.org/10.1007/s13394-018-0251-9

Namoun A, Alshanqiti A (2020) Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences 11*: 237. https://doi.org/10.3390/app11010237

Oreshin S, Filchenkov A, Petrusha P, Krasheninnikov E, Panfilov A, Glukhov I, Kaliberda Y, Masalskiy D, Serdyukov A, Kazakovtsev V, Khlopotov M, Podolenchuk T, Smetannikov I, Kazlova D (2020) *Implementing a Machine Learning Approach to Predicting Students' Academic Outcomes*. At the 2020 International Conference on Control, Robotics and Intelligent System. Xiamen, China. October 27-29. pp: 78–83. https://doi.org/10.1145/3437802.3437816

Owens A (2020) Unequal Opportunity: School and Neighborhood Segregation in the USA. *Race and Social Problems 12*: 29–41. https://doi.org/10.1007/s12552-019-09274-z

Peng P, Wang T, Wang C, Lin X (2019) A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin 145*: 189–236. https://doi.org/10.1037/bul0000182

Qazdar A, Er-Raha B, Cherkaoui C, Mammass D (2019) A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies 24*: 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Rajendran S (2021) Predicting Factors Impacting Student Academic Performance using Machine Learning Algorithms. *SSRN*: 1–38. http://doi.org/10.2139/ssrn.3898302

Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A (2020) Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences 10*: 1042. https://doi.org/10.3390/app10031042

Rodríguez P, Villanueva A (2022) Design, Development, and Evaluation of a Predictive Model for Regular School Dropout in the Chilean Educational System. Hosseini S, Peluffo DH, Nganji J, Arrona-Palacios, A (eds) *Technology-Enabled Innovations in Education. Transactions on Computer Systems and Networks*. Springer, Singapore. pp 493-505. https://doi.org/10.1007/978-981-19-3383-7_40

Rodríguez P, Villanueva A, Dombrovskaia L, Valenzuela JP (2023) A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies 28*: 10103-10149. https://doi.org/10.1007/s10639-022-11515-5

Ross MA (2019) *Resources, Pupil-Type, or Personal Attention: Investigating the Relationship between School Size and Student Achievement on Pennsylvania Standardized Tests*. Doctoral dissertation. Youngstown State University. Youngstown, USA. 24 pp. https://search.proquest.com/openview/54b0eb7c6ec9f2db8279eb562bdd99e0/1?pq-origsite=gscholar&cbl=18750&diss=y

Schwartz AE, Stiefel L, Wiswall M (2013) Do small schools improve performance in large, urban districts? Causal evidence from New York City. *Journal of Urban Economics 77*: 27–40. https://doi.org/10.1016/j.jue.2013.03.008

Selmeczy D, Ghetti S, Zheng LR, Porter T, Trzesniewski K (2021) Help me understand: Adaptive information-seeking predicts academic achievement in school-aged children. *Cognitive Development 59*: 101062. https://doi.org/10.1016/j.cogdev.2021.101062

Shalev-Shwartz S, Srebro N (2008) *Svm optimization: inverse dependence on training set size*. At the Proceedings of the 25th International Conference on Machine Learning. New York, USA. July 5-9. pp. 928–935. https://doi.org/10.1145/1390156.1390273

Smith J, Gutiérrez C (2022) Una aplicación de aprendizaje automático (machine learning) en políticas públicas. Predicción de alerta temprana de deserción escolar en el sistema de educación pública de Chile. *Mutidisciplinary Business Review 15*: 20–35. https://doi.org/10.35692/07183992.15.1.4

Somvanshi M, Chavan P, Tambade S, Shinde S (2016) *A review of machine learning techniques using decision tree and support vector machine*. At the 2016 International Conference on Computing Communication Control and automation (ICCUBEA). Pune, India. August 12-13. pp. 1–7. https://

doi.org/10.1109/ICCUBEA.2016.7860040

Tarik A, Aissa H, Yousef F (2021) Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science 184*: 835–840. https://doi.org/10.1016/j.procs.2021.03.104

Treviño E, Valenzuela JP, Villalobos C (2016) Within-school segregation in the Chilean school system: What factors explain it? How efficient is this practice for fostering student achievement and equity? *Learning and Individual Differences 51*: 367–375. https://doi.org/10.1016/j.lindif.2016.08.021

Turabieh H, Azwari SA, Rokaya M, Alosaimi W, Alharbi A, Alhakami W, Alnfiai M (2021) Enhanced Harris Hawks optimization as a feature selection for the prediction of student performance. *Computing 103*: 1417–1438. https://doi.org/10.1007/s00607-020-00894-7

Xu X, Wang J, Peng H, Wu R (2019) Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior 98*: 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Yakubu MN, Abubakar AM (2021) Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes 51*: 916–934. https://doi.org/10.1108/ K-12-2020-0865

Yip MC (2021) The linkage among academic performance, learning strategies and self-efficacy of japanese university students: A mixed-method approach. *Studies in Higher Education 46*: 1565–1577. https://doi.org/10.1080/03075079.2019.1695111

Zhang C, Ma Y (2012) *Ensemble Machine Learning: Methods and Applications*. Springer. USA. 332 pp.