

Students' Numeracy and Literacy Aptitude Analysis and Prediction Using Machine Learning

Tianyu Li

St. George's School, Vancouver, Canada
Email: lawrenceee0126@gmail.com

How to cite this paper: Li, T.Y. (2022) Students' Numeracy and Literacy Aptitude Analysis and Prediction Using Machine Learning. *Journal of Computer and Communications*, 10, 90-103.
<https://doi.org/10.4236/jcc.2022.108006>

Received: June 25, 2022

Accepted: August 23, 2022

Published: August 26, 2022

Copyright © 2022 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Education is one of the most pivotal services in societal development as it cultivates a wide variety of skills, especially numeracy and literacy skills. However, students may have varying masteries of these two aptitudes. Some attribute this to students' intrinsic efforts while others attribute this to students' capabilities and affiliated environments. In this work, I explore the numeracy and literacy aptitude patterns of students from various cultures based on a dataset that contains various demographic information, from which I deduced some preliminary trends. After the comparison of numerous machine learning algorithms, the optimal algorithm or combination of a few algorithms predicts students' performances by classifying students of different backgrounds into various potential outcomes. The results suggest that proper resources and supports are necessary for enhanced learning.

Keywords

Numeracy and Literacy, Machine Learning, Learning Analytics, Classification

1. Introduction

At school, subjects like math, reading, and writing constitute numeracy and literacy skills, which are fundamental in people's everyday life. Numeracy skills enable people to think critically, calculate, and thus, make decisions [1]. Literacy skills complement the former by enabling people to understand numerical operations, qualify for employment opportunities (economic benefit), and assimilate into their social surroundings [2]. Incontrovertibly, these two aptitudes are indispensable in people's lives. Thus, this project predicts people's most likely learning outcomes in these two aptitudes—including students' best subject, sub-

jects needing help, and overall learning level—and guides people for improvements according to how various levels of different factors influence learning; perhaps, students could refine their learning experiences and improve in the two substantial aptitudes by referring themselves to the results accordingly.

Good students usually share some characteristics. For instance, most of them are self-disciplined, allowing them to handle their course works. They are generally responsible as well, making them reliable people. Therefore, other factors, which may not be as simplistic as singular traits, might similarly impact students' learning outcomes. To find out about this potential phenomenon, a dataset of ten thousand public school students' information is used [3]. The data points in this dataset, which is from Dr. Royce Kimmons, are quite normally distributed with organized information. In the dataset, five factors—gender, ethnicity, parental education level, lunch quality, and test preparation participation—are accounted for each student. In each factor, there are different subgroups. The dynamics of each level of different factors are the core of exploration.

In order to systematically approach this task, my work incorporates machine learning. As a popular tool frequently utilized for all purposes, machine learning benefits analysis by holistically treating all data inputs with a standardized set of criteria. One of its most prominent advantages in this project is its ability to handle multi-dimensional data [4]. The machine can appropriately perform calculations on and synthesize all ten thousand multi-dimensional data points, which each contains five factors in a set of subgroups.

Each student, who will have a data value in each factor, is a data point, and to classify, a student will be drawn based on the training data to a group of data points that they are similar to. As a result, several most commonly used supervised classification algorithms and ensemble methods are used. Whereas some algorithms, such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), work independently, others are ensemble learning methods, including Gradientboosting and Voting Classifier. The models are compared under different tasks by their confusion matrix and accuracy scores. Each kind has its unique advantage due to the different mechanisms and thus, is useful in the project.

Along with the predictive power of the project, exploratory analysis of subgroups of different factors reveals certain patterns of how learning is impacted. People in some subgroups do have an edge over others in learning, confirming the trend people have suspected. For instance, females have generally higher literacy levels. Despite that there are far more complex conditions behind the mere factors intertwining students' comprehensions, the discovered patterns could be good starting points and should be carefully considered.

The rest of the paper is organized as follows: Section 2 details the designs, results, and inquiries as well as the machine learning algorithms. Section 3 presents several related works to my project to draw parallels. Finally, Section 4 concludes the remarks of this research.

2. Related Works

Boran Sekeroglu *et al.* [5] recognize the significance of technology in education. In fact, AI is already a popular tool for both educators and students. Their work compares three different algorithms, from which Backpropagation (BP) classifies and Support Vector Machine (SVM) predicts students' performances in two datasets with an accuracy of around 80%. With the possibility of increasing accuracy from different data selections, machine learning algorithms can predict and classify educational data.

Abdelmajid Chaffai *et al.* [6] implement the C5.0 algorithm, which has a 90% accuracy in this case, to predict new high school students' first-year academic results and a k-means clustering algorithm to recommend students to a majoring department that they are the most suitable for. Based on pre-higher education data and first-year scores over nine years of students in a high school, the program can predict scores and recommend majoring departments for future students through an interactive platform.

Nikola Tomasevic *et al.* [7] conducted a study to predict students' academic performance and identify those at high risk of dropping out. The various machine learning algorithms used for prediction involved in the work include Decision Trees, the SVM model, and Naive Bayes (NB). To compare machine learning algorithms by their optimal performance, the optimal combination of input data, or factors, is determined from the dataset, which encompasses students' demographics, engagement at school, and scores.

3. Data Exploration

As aforementioned, numeracy and literacy skills are indispensable in almost every aspect of people's daily lives. However, not all students can competently understand the relevant knowledge. Thus, for students, especially those needing assistance, to improve the two skills, mechanisms of efficient studying should be explored. Prominently, good students' social-cultural backgrounds provide them exclusive resources in addition to in-school support that's also available to those in need.

The dataset includes valuable information for this investigation. The subgroups of each cultural factor are as follows: Gender (Male/Female), Ethnicity (Group A-E), Parents' Education Level (Some High School/High School/Associate's Degree/Bachelor's Degree/Master's Degree), Lunch (Standard/Reduced Quality), and Test Preparation Course (Completed/None).

Due to the inherent limitation of the dataset, students' performances on exams—math, reading, and writing scores—are the references to infer numeracy (Math) and literacy (Reading + Writing) skills.

To comprehensively examine the influences of different subgroups, the exploratory analysis is divided into subsections of inquiries.

3.1. The General Trend of Students' Numeracy and Aptitude Skills

As the first inquiry, finding the general trend gives a standard to judge students'

scores for the following steps.

As shown in **Figure 1**, all subject scores are quite symmetrically distributed with the means of all subject scores around 70. More specifically, students are better at reading than at writing and math, the subject that most students are the least comfortable with. The total score distribution is skewed left, and its mean is slightly above 200.

Thus, students generally have better literacy skills than numeracy skills. There are also students that significantly study worse than others.

3.2. Total Score Patterns among Subgroups of Each Factor

Part of our hypothesis acknowledges the different influences subgroups have on learning. Total scores are the metric to begin with.

As shown in **Figure 2**, females are generally better in two aptitudes aggregated. The scores of the entire group as a whole are also more concentrated.

As shown in **Figure 3**, students' overall aptitude level positively correlates with parental education level. Additionally, the mean total score of students whose parents have an associate's degree or higher is above all students' mean scores. The score distributions of students whose parents have higher education levels also have less variation or outliers.

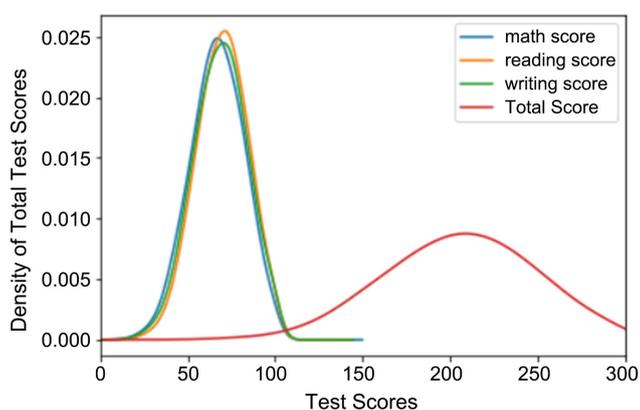


Figure 1. Density map of subject and total scores.

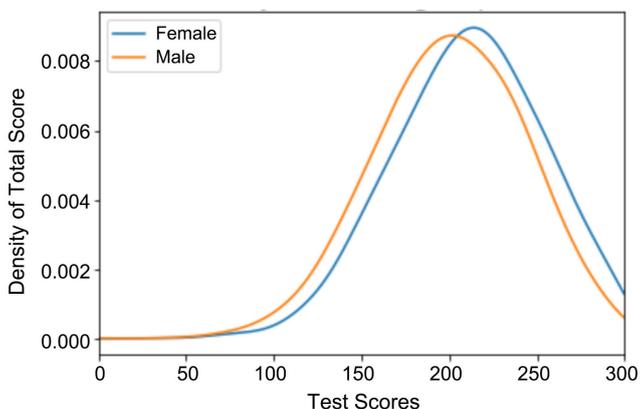


Figure 2. Total score distribution by genders.

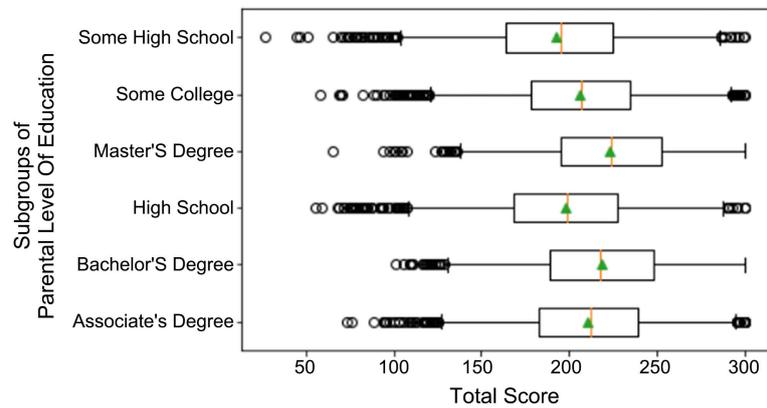


Figure 3. Total score distribution by subgroups of parental education level.

Thereby, a higher parental education level is the most conducive to overall aptitude development.

Similarly, students' overall aptitude level positively correlates with ethnicity group letter (with the mean scores of students in Group C and higher above all students' mean score), lunch quality, and test preparation course participation.

3.3. Subject Score Patterns among Subgroups of Each Factor

While the overall aptitude level is explored, how the subgroups affect each subject is equally important, if not more. Finding the better subgroups at each subject helps determine the most efficient improvement method.

As shown in **Figure 4**, females are generally better than males at writing, having a mean score 10 points higher than that of males. There is also a significant number of female students earning full scores in writing. More importantly, the standard deviation of female students' writing scores is also lower than that of male students' writing scores. All of these patterns are reflected in reading scores as well.

However, not all subject score patterns among the two genders follow the total score patterns.

As shown in **Figure 5**, males generally score higher in math with a 5-points margin in mean scores. Specifically, males' math scores have a higher standard deviation than females' do though.

Thereby, males should be models in math while following how females learn reading and writing. Males as an entire group may also emulate how females score consistently.

Subject score trends among subgroups of other demographic factors are consistent with the total score trend among them.

3.4. Comparison of Each Factor's Subgroups in Scoring High and Passing a Subject and the Overall Aptitude

In education, two goals are almost the most significant ones pursued: developing elite students and the least, equipping everyone with basic knowledge. Elite students would become the most competent to innovate for society while others,

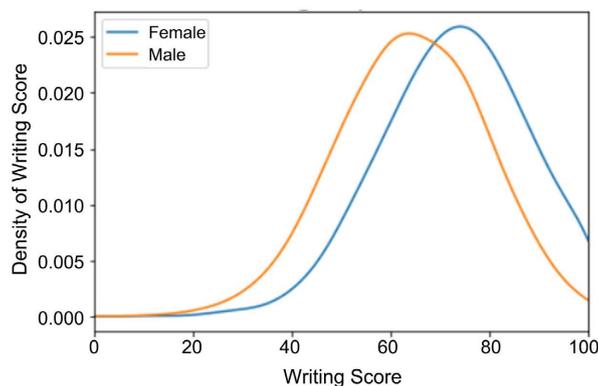


Figure 4. Writing score distribution by genders.

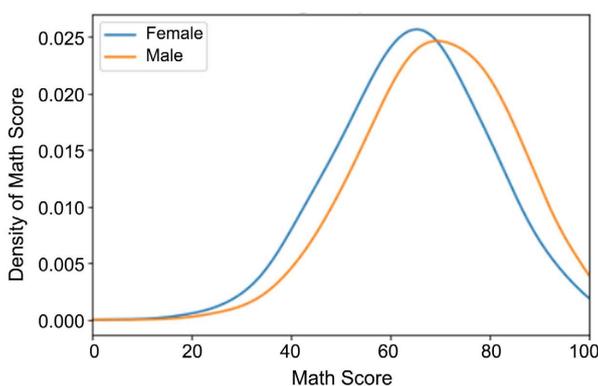


Figure 5. Math score distribution by genders.

who have enough knowledge to perform certain jobs, could contribute to society by realizing the innovations. As a result, both groups are resources integral to development. The demographic status of both groups is explored for students in need to refer to. In my work, elite students are defined as those scoring 90% or higher while the passing cutoff is 60%.

Further confirming the trends previously discussed, the trends at the two thresholds provide more details. A commonality between all trends in this inquiry is that in the same score area, students in subgroups that are more advantaged to become elites are also more advantaged to have the basic knowledge. Moreover, the extent that subgroups affect students is greater in excelling than in passing each subject.

For instance, comparing the reading scores of two genders, females have an almost three times higher likelihood to earn a high score (as shown in **Figure 6**) while both genders have comparable rates of passing. Thus, females' advantage in reading is well-established.

Besides, as shown in **Figure 7** and **Figure 8**, there is a nuance in ethnicity trends: Group A has a slightly higher high score rate than Group B, which is a subgroup of a higher letter. However, other trends are still generally consistent with previous findings as to the high score rates and passing rates still positively correlate with the group letter. Interestingly, Group E's advantage in both passing and high-scoring is exceptionally significant. Therefore, ethnicities of higher

group letters should be models in both education goals, especially Group E.

Lastly, lunch turns out to have even more significant subgroup differences. As shown in **Figure 9**, the total score high-scoring rate of students with standard lunch is 6.5 times higher than that of students with lunch of lower quality. The passing rate margin is also more obvious than other factors'. Hence, lunch quality heavily affects students learning outcomes.

Therefore, certain demographic status does increase students' likelihood to both pass and, particularly, excel in an aptitude.

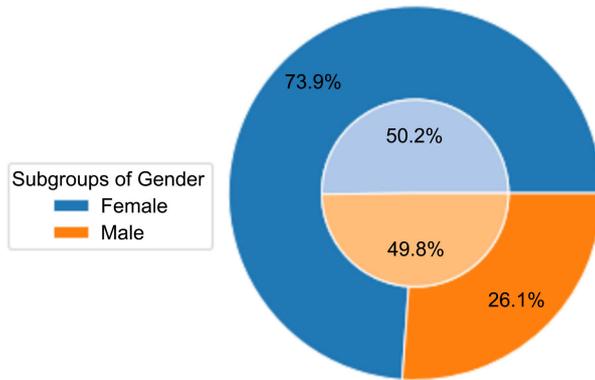


Figure 6. High reading score and population ratio among genders.

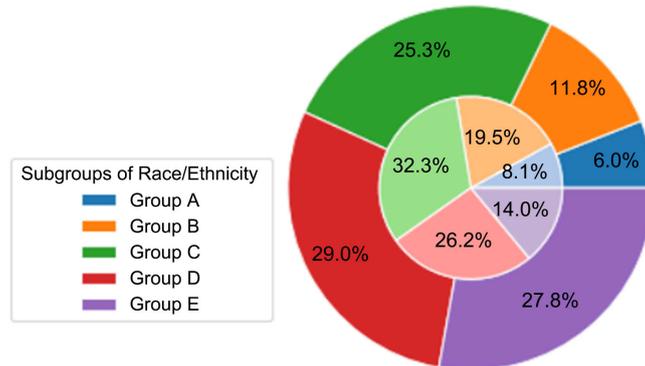


Figure 7. High total score and population ratio among ethnicity subgroups.

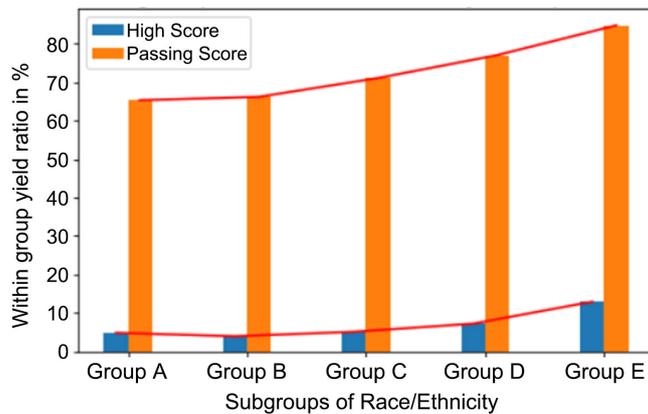


Figure 8. High and passing yield ratio among ethnicity subgroups.

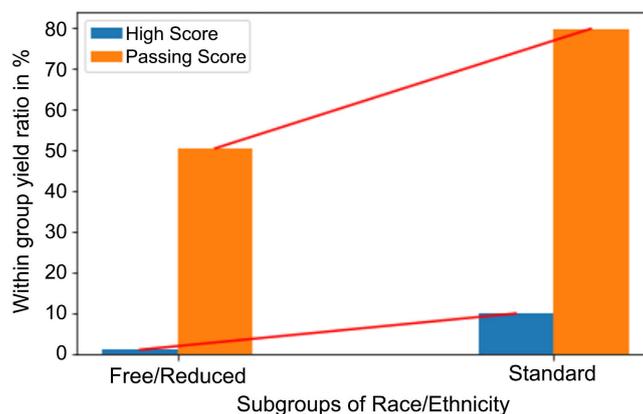


Figure 9. High total score and population ratio among lunch subgroups.

4. Data Analysis and Prediction

In this paper, several machine learning algorithms are implemented to either determine the significance of each factor or, more importantly, predict every student's learning outcome.

To start with, Principal Component Analysis (PCA) is a dimensionality reduction method that maintains as much information as possible [8]. It creates uncorrelated principal components, which are combinations of original variables into new ones with maximum variance. Moreover, although the dynamics of subgroups in a demographic factor are discovered, PCA enables the ranking of the importance of different demographic factors.

Additionally, numerous classification algorithms are incorporated—including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, Random Forest Classifier (RFC), Gaussian Naive Bayes (GNB), Gradient Boost Classifier (GBC), and Voting Classifier.

KNN is a supervised machine learning algorithm that can perform classification and regression [9]. It classifies based on the most frequent class that data points near the input data are in.

SVM is a classic supervised machine learning algorithm that helps classification. Data points are mapped out in a high-dimension space, and a decision surface is constructed to classify the input data.

Decision Trees is a classification model that makes decisions based on possible consequences from a tree-like flow model [10]. Each node in the flowchart has a criterion that tests the input data.

RFC is an ensemble learning method that can deal with classification problems [11]. It is formed by several decision trees and predicts with other models such as bagging.

GNB is a supervised classification algorithm based on the Bayes Theorem, specifically assuming a Gaussian distribution [12]. Predictions are made after calculating the mean and standard deviation of data.

GBC is another ensemble method, which builds on many predictive algorithms. It modifies weaker methods and creates a new decision mechanism.

Voting Classifier is also an ensemble method and allows customized machine learning algorithms in the parameter for prediction. The decision is the highest probability output from the input algorithms chosen by the majority vote. In this paper, Voting Classifier consists of the aforementioned algorithms.

In training and testing each model, a 10% testing ratio is applied, and accuracies along with confusion matrixes are generated. Accuracy is calculated by $(TP + TN)/(TP + TN + FP + FN) \times 100\%$, where TP and TN are True Positive and True Negative, and FP and FN are False Positive and False Negative.

In this paper, the two principal components account for 88.7% of the variance, as shown in **Figure 10**. In addition, as shown in **Figure 11**, two factors—ethnicity and parental education level—are the most influential in students' learning since their vectors are the longest along the first principal component axes. Further, parental education level and ethnicity negatively correlate; so do gender and lunch while ethnicity and test preparation course participation positively correlate.

In theory, PCA would compensate accuracy for simplicity. However, in my work, PCA, applied with 90% of the variance, does not have much lower accuracies and even, in some cases, has higher accuracies. Thus, PCA is useful tool in providing insights and predictions.

As shown in **Table 1** and **Table 2**, the models' accuracy scores indicate that GBC suits for best subject prediction, and Decision Trees suits for writing outcome prediction. On the other hand, GNB suits for math and reading outcome prediction (not shown), and SVC suits for overall learning outcome prediction (not shown). Besides, PCA minimally affects the prediction accuracy. The main reason is most likely that the demographic factors analyzed in this work are not closely correlated to any other. As shown in **Figure 12**, the highest correlation is only 0.017, which is well below the threshold correlation, 0.3 [13]. The PCA does not reduce the data, and thus, the results are similar to ones without PCA.

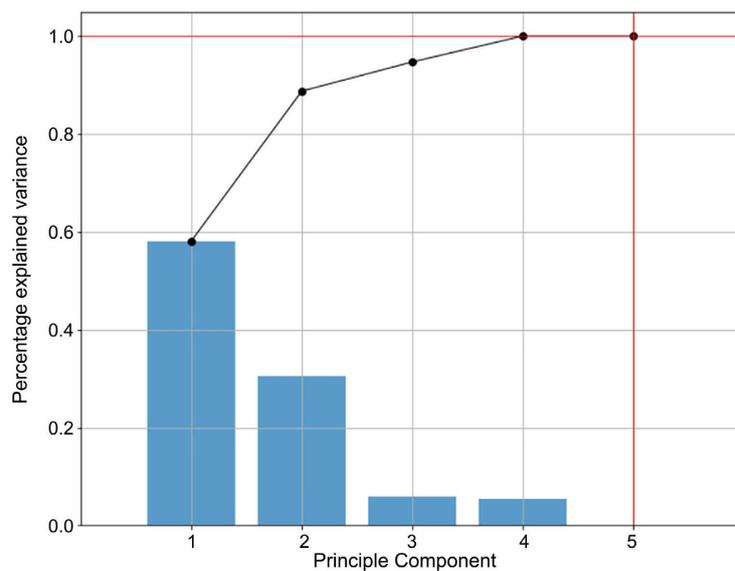


Figure 10. PCA cumulative explained variance graph.

Interestingly, the confusion matrixes of the same model with and without PCA complement each other; the same model with PCA may predict more TP in one class than it without PCA may.

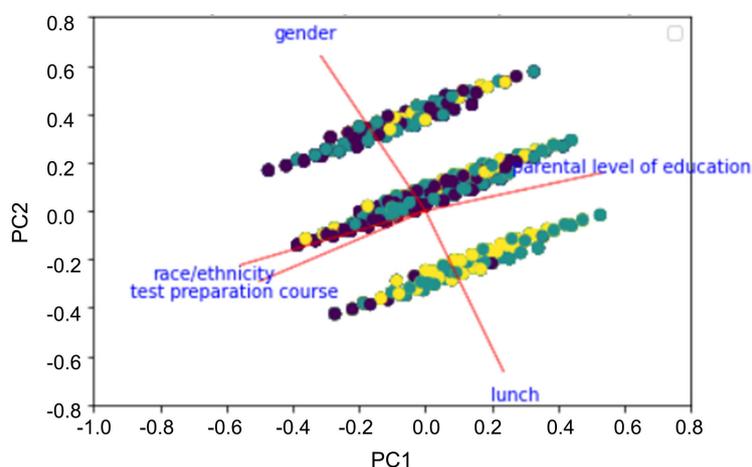


Figure 11. PCA biplot.

Table 1. Best subject prediction results for all algorithms.

Algorithms	Without PCA		With PCA	
	Accuracy	Confusion Matrix	Accuracy	Confusion Matrix
KNN	59.6%	$\begin{bmatrix} 281 & 43 & 21 \\ 105 & 223 & 67 \\ 42 & 126 & 92 \end{bmatrix}$	59.6%	$\begin{bmatrix} 281 & 43 & 21 \\ 105 & 223 & 67 \\ 42 & 126 & 92 \end{bmatrix}$
GNB	61%	$\begin{bmatrix} 307 & 30 & 8 \\ 133 & 199 & 63 \\ 47 & 109 & 104 \end{bmatrix}$	61%	$\begin{bmatrix} 307 & 30 & 8 \\ 133 & 199 & 63 \\ 47 & 109 & 104 \end{bmatrix}$
SVM	60.9%	$\begin{bmatrix} 305 & 31 & 9 \\ 132 & 199 & 64 \\ 46 & 109 & 105 \end{bmatrix}$	62%	$\begin{bmatrix} 294 & 43 & 8 \\ 109 & 226 & 61 \\ 40 & 119 & 101 \end{bmatrix}$
Voting Classifier	61.5%	$\begin{bmatrix} 298 & 39 & 8 \\ 119 & 222 & 54 \\ 42 & 124 & 94 \end{bmatrix}$	61.5%	$\begin{bmatrix} 298 & 39 & 8 \\ 119 & 222 & 54 \\ 42 & 124 & 94 \end{bmatrix}$
Decision Trees	61.8%	$\begin{bmatrix} 294 & 42 & 9 \\ 111 & 224 & 60 \\ 43 & 117 & 100 \end{bmatrix}$	61.8%	$\begin{bmatrix} 294 & 42 & 9 \\ 111 & 224 & 60 \\ 43 & 117 & 100 \end{bmatrix}$
RFC	61.9%	$\begin{bmatrix} 294 & 42 & 9 \\ 110 & 224 & 61 \\ 43 & 117 & 100 \end{bmatrix}$	62%	$\begin{bmatrix} 294 & 42 & 9 \\ 109 & 224 & 62 \\ 42 & 115 & 103 \end{bmatrix}$
GBC	62.2%	$\begin{bmatrix} 291 & 47 & 7 \\ 111 & 221 & 63 \\ 41 & 109 & 110 \end{bmatrix}$	62.2%	$\begin{bmatrix} 291 & 47 & 7 \\ 111 & 221 & 63 \\ 41 & 109 & 110 \end{bmatrix}$

Table 2. Writing learning outcome prediction results for all algorithms.

Algorithms	Without PCA		With PCA	
	Accuracy	Confusion Matrix	Accuracy	Confusion Matrix
KNN	71%	$\begin{bmatrix} 563 & 114 \\ 176 & 147 \end{bmatrix}$	71.4%	$\begin{bmatrix} 560 & 117 \\ 169 & 154 \end{bmatrix}$
GNB	71.1%	$\begin{bmatrix} 585 & 92 \\ 197 & 126 \end{bmatrix}$	71.1%	$\begin{bmatrix} 585 & 92 \\ 197 & 126 \end{bmatrix}$
SVM	72%	$\begin{bmatrix} 613 & 64 \\ 216 & 107 \end{bmatrix}$	71.3%	$\begin{bmatrix} 577 & 100 \\ 187 & 136 \end{bmatrix}$
GBC	72%	$\begin{bmatrix} 569 & 108 \\ 172 & 151 \end{bmatrix}$	72%	$\begin{bmatrix} 569 & 108 \\ 172 & 151 \end{bmatrix}$
Voting Classifier	71.6%	$\begin{bmatrix} 631 & 46 \\ 238 & 85 \end{bmatrix}$	72.1%	$\begin{bmatrix} 591 & 86 \\ 193 & 130 \end{bmatrix}$
RFC	72.7%	$\begin{bmatrix} 576 & 101 \\ 172 & 151 \end{bmatrix}$	72.6%	$\begin{bmatrix} 577 & 100 \\ 174 & 149 \end{bmatrix}$
Decision Trees	72.8%	$\begin{bmatrix} 579 & 98 \\ 174 & 149 \end{bmatrix}$	72.8%	$\begin{bmatrix} 579 & 98 \\ 174 & 149 \end{bmatrix}$

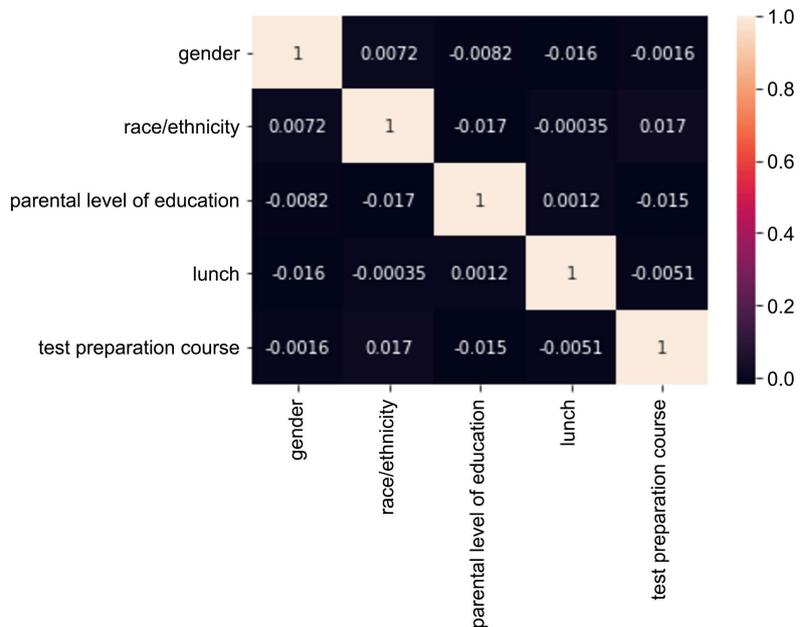


Figure 12. Correlation matrix.

Moreover, whereas some models overall better fit for some prediction areas, other models have their own advantages. For example, despite that GBC with PCA has higher accuracy than RFC with PCA (Table 1), GBC only leads in one of the subjects' TP prediction.

Result Discussion

After training and selecting the best models, predictions are made in five areas—students' best subject, learning outcome in math, reading, and writing, and overall aptitude level—for every student. As the outcomes of this project, the predicted status in each area acts as a reference for everyone since one can compare himself/herself to the one in my project. Everyone can find a place in the results because predictions are made for every possible, unique background, which is found by the permutation tool *itertools*.

The prediction results align with the findings identified in the previous section. If students' background is mixed with the more advantaged subgroups and the less advantaged ones, the larger number of advantaged subgroups there are, the more likely the student learns better.

Therefore, even though the prediction model is not perfect, students can have a general idea of their performance level and adjust accordingly. As a rule of thumb, all students should participate in test preparation courses and have standard lunch; in both aptitudes, those two qualities generally lead to better results. For the other three inherent factors, there are some controllable actions students can take. Students could learn from the better subgroups and mimic their behavior. For instance, Group E is well-performing in all subjects. If Group E has a certain study schedule, students in other groups should adapt themselves to that schedule as appropriate. Similarly, other characteristics can be observed and practiced accordingly to enhance learning.

5. Conclusions

In this paper, the influence of the students' demographic backgrounds on their math, reading, and writing—which reflect students' numeracy and literacy aptitudes—is examined through data exploration. The findings confirm the general trends in society. For example, females have better literacy aptitude while males have better numeracy aptitude. Students of the ethnicity of a higher letter with a higher level of parental education, lunch, and test preparation course participation are more likely to both pass and excel in all courses and thus, are more likely to have better numeracy and literacy aptitude. Those students as a whole also have less variation in their scores than other students do. More importantly, two of the inherent factors—ethnicity and parental education level, are the most principal in influencing one's learning, despite that the exact rationale behind the discovered phenomenon is not clear.

To predict students' aptitudes, GBC, GNB, Decision Trees, and SVC are the best in their respective areas, achieving accuracies of near 70%.

Future works are essential to refine the methods for students to improve. One of the principal jobs is exploring the learning environments of students in better subgroups. This may involve social complexions that are beyond the mere factors and helps execute the actions for improvement. For example, certain subgroups are more privileged; females have more access to test preparation courses

than males do, as indicated by the positive correlation between gender and test preparation course in the PCA biplot. Similarly, students of parents with higher education levels are more likely to have standard quality lunches. If the education systems and societies that students in need are from can cater to their students by offering an environment privileged students enjoy, students could best eliminate their gaps.

In addition to intra-factor investigations, inter-factor research is also valuable. Indeed, some subgroups of a factor may be more advantaged in an aptitude, but what makes those subgroups more advantaged requires field experiments. For instance, students of parents with Master's degrees are more likely to succeed. However, it is not that degree but the parents' certain qualities that better the children's learning. Once the causal relationships are found, students can benefit from leveraging the conducive qualities of bigger factor subgroups even if they are not in the prime subgroups identified in this paper.

Lastly, the accuracy of the machine learning models can be further polished. While accuracies of around 60% - 70% can be acceptable, they should certainly be improved considering if used on a bigger population, the large number of people that may have a false benchmark to refer to. To modify the models, perhaps finding a dataset of more data points and social-economic information may be beneficial. A bigger dataset usually allows models to have more inputs to learn from, and more social-economic information helps models to make more definitive predictions. At the same time, experimenting with alternative models to current ones may help as certain models may be better fits to tackle the problem.

Acknowledgements

The research presented in this paper is done with guidance from Dr. Yang Lan, a computer science professor at California Polytechnic State University, and Dr. Kong Liang, a postdoctoral researcher from the University of Kentucky.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- [1] Miron, A. (2013) The Importance of Numerical Reasoning. <https://www.eskill.com/blog/numerical-reasoning/>
- [2] Wilson, R.T. (2016) Literacy and Its Significance in Modern Life. *Colleagues*, **13**, Article No. 5. <http://scholarworks.gvsu.edu/colleagues/vol13/iss1/5>
- [3] Kimmons, R. (2012) Exam Scores. http://roycekimmons.com/tools/generated_data/exams
- [4] (2021) Advantages and Disadvantages of Machine Learning Language, Data Flair. <https://data-flair.training/blogs/advantages-and-disadvantages-of-machine-learning/>
- [5] Sekeroglu, B., Dimililer, K. and Tuncal, K. (2019) Student Performance Prediction and Classification Using Machine Learning Algorithms. *Proceedings of the 2019 8th*

-
- International Conference on Educational and Information Technology*, New York, 2-4 March 2019, 7-11. <https://doi.org/10.1145/3318396.3318419>
- [6] Chaffai, A., el Abdouli, A., Anoun, H., Hassouni, L. and Khalid, R. (2015) Student Performance Analysis Using Large-Scale Machine Learning. *International Journal of Computer Science and Information Security*, **13**, No. 12.
- [7] Tomasevic, N., Gvozdenovic, N. and Vranes, S. (2019) An Overview and Comparison of Supervised Data Mining Techniques for Student Exam Performance Prediction. *Computers & Education*, **143**, Article ID: 103676. <https://doi.org/10.1016/j.compedu.2019.103676>
- [8] Jaadi, Z. (2021) A Step-by-Step Explanation of Principal Component Analysis (PCA). <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
- [9] Peterson, L.E. (2009) K-Nearest Neighbor. <https://doi.org/10.4249/scholarpedia.1883>
- [10] Kaminski, B., Jakubczyk, M. and Szufel, P. (2017) A Framework for Sensitivity Analysis of Decision Trees. *Central European Journal of Operations Research*, **26**, 135-159. <https://doi.org/10.1007/s10100-017-0479-6>
- [11] Mbaabu, O. (2020) Introduction to Random Forest in Machine Learning. <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>
- [12] Brownlee, J. (2016) Naive Bayes for Machine Learning. <https://machinelearningmastery.com/naive-bayes-for-machine-learning/>
- [13] Origin Lab. Principal Component Analysis. <https://www.originlab.com/doc/Origin-Help/PrincipleComp-Analysis>
<https://doi.org/10.1016/j.compedu.2019.103676>