STEM Fellowship High School Big Data and AI Challenge 2024-2025

Literacy Rates Factor Analysis

Cathy Zhang[1], Elliot Fang[1], Jamie Seoh[1], and Aarnav Thite[1]

[1]Iroquois Ridge High School

January 11, 2025

Abstract

The purpose of this study was to identify and analyze the most important factors influencing global literacy rates. As a critical skill for personal and societal development, understanding its determining factors allows policymakers and educators to implement more effective strategies aimed at improving educational outcomes. Data for this study was obtained from multiple credible sources, including the UNESCO Institute for Statistics, Our World in Data, OECD, and Kaggle databases. Following the processing and merging of the datasets, several statistical methods were carried out to analyze the relationship between literacy rates and various factors. A Random Forest Regressor model was used to calculate relative importance of factors, while a Chi-Square test and correlation matrix were applied to analyze the significance of various relationships. The analysis found that the pupil-teacher ratio is the most influential factor affecting literacy rates, with lower ratios significantly contributing to better educational outcomes. The Gross Enrollment ratio was the second most important factor, and other factors such as primary and secondary school enrollment and completion rates demonstrated moderate influence. Government expenditure on education showed minimal direct impact, which suggests that efficient resource allocation plays a more critical role than the amount of resources. These findings emphasize the need for targeted policies to improve pupil-teacher ratios and quality of education. Future research could explore the interaction between education quality and access across socio-economic backgrounds to further optimize literacy improvement strategies.

Keywords
Literacy, Education, Correlation

1 Introduction

Literacy is the ability to understand, assess, use, and create written text to contribute to society and develop one's knowledge to achieve both personal and broader-ranged goals [1]. It is crucial for both personal and societal development, providing individuals a foundation for further learning and critical thinking. Despite global efforts to improve literacy rates, challenges arise in attempts to sustainably alter the large-scale education system with continuous changes to the approach to match feedback [2].

With the many difficulties in managing and following through with an extensive effort such as altering the world's education systems, analyzing the factors that influence the dependent variable, literacy rates, becomes essential.

One of the factors that influence literacy rates is the pupil-teacher ratio. Areas with higher ratios are found to have lower literacy rates [3]. The higher ratios necessitate the larger class sizes, preventing teachers from spending time with individual students and directly influencing their academic performance [4]. Teachers could get to know the students better and provide them with personalized support [5]. Larger numbers of students packed into one classroom also greatly hinder quality learning [6].

Government expenditure on the education department is also an influential factor, though it is extremely limited in developing countries due to financial difficulties [7]. The expenditure directly affects the growth of the education sector, thereby affecting the literacy rate [8].

A country's gross domestic product (GDP) also has a significant impact on its literacy rate. A higher GDP yields higher literacy rates [9], as the increased prosperity allows for increased access to education.

Enrollment and completion rates are also found to be directly correlated with literacy rates [10][11]. Higher enrollment rates lead to a

greater portion of the population being educated and becoming literate, and higher completion rates means that a greater portion of the population has at least begun to develop literacy.

The purpose of this study is to identify and analyze the factors that most significantly impact the world's literacy rate. It merged data on different factors found from the UNESCO Institute of Statistics database [12], the Our World in Data database [13], the OECD database [14], and the Kaggle database [15]. Several analysis tools such as Random Forest, the Chi-Square Test, and a correlation matrix were utilized to identify whether correlations between literacy rates and the factors existed, and to determine the factor with the strongest impact on literacy rates.

With more insight on the key factors, educators and policymakers can design better-informed strategies to increase literacy, contributing to educational policies and development. More citizens will be able to bring wealth to their countries, improving economies all over the world [16].

2 Materials & Methods

All analyses were performed using Python 3.10.7.

2.1 Reformatting and Merging the Raw Datasets

All CSV datasets were reformatted using the Pandas library so that each factor would have its own column for effective analysis later on. The data was then merged and reorganized into one main dataset that would then be analyzed through various methods.

2.2 Correlation Analysis

2.2.1 Random Forest Regressor model

The Pandas, train_test_split, and RandomForestRegressor Python libraries were used to split the main dataset into data to train and test the Random Forest model. The data was split into a 3:7 ratio of training to testing. The model assigned an importance value to each factor that was inputted based on how much it impacted the literacy rate factor. The result was then plotted using the matplotlib.pyplot and seaborn Python libraries.

2.2.2 Chi-Square Test

By utilizing the Pandas and chi2_contingency (from SciPy) libraries to load the datasets, categorical bins for relevant variables and definitions are created. They are then used to produce a contingency table to perform the Chi-Square test; a few variables are calculated as a result. The Chi-Square Statistic measures the magnitude of the difference between observed and expected frequencies. The P-value is the probability that the observed differences occurred by chance. The hypothesis that is made, is that the variables that are being compared are independent. If the P-value is smaller than the Chi-Square Statistic, this hypothesis is rejected, proving the significance in the correlation between the two variables [17].

2.2.3 Correlation Matrix

Using the Pandas library to generate a correlation matrix and Pygal, this dot chart is returned. The Pandas .corr() method was used to calculate pairwise Pearson correlation coefficients between variables. The resulting correlation matrix was visualized as a heatmap using the Pygal library, which employed a color gradient to indicate positive correlations. Brighter shades of red indicated stronger correlations, and likewise for blue shades and negative correlations.

3 Results

3.1 The pupil-teacher ratio is the feature that has the most influence on the literacy rates

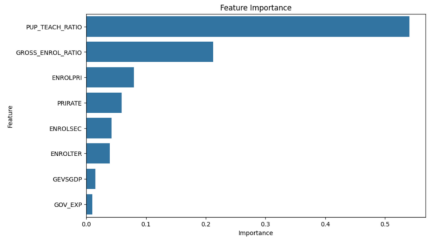

Figure 1: Bar graph depicting the importance value assigned to each feature using the Random Forest Regressor model when determining their impact on literacy rates.

3.2 The pupil-teacher ratio is significant for literacy rates

[h]

```
Contingency Table:
 PUP_TEACH_CAT  Low  Medium  High
LITRATE_CAT
Low              0       4     9
Medium           1       6     4
High            39      43     3

Chi-Square Statistic: 47.08581649759785
P-Value: 1.4634263933691644e-09
Degrees of Freedom: 4
```

Figure 2: Formatted result from a written script from a Chi-Square test that proves the pupil-teacher ratio is significant to literacy rate using the X2 contingency model.

3.3 Correlation matrix depicting strongest correlation between pupil-teacher ratio and literacy rate

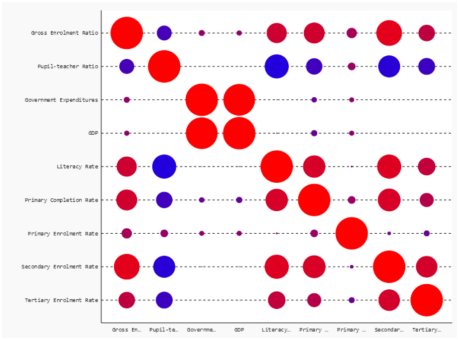

Figure 3: Dot chart depicting the correlation between various factors of literacy rates. The larger the circle, the stronger the correlation, and the redder the circle, the more positive the correlation is. It is notable that there is a strong negative correlation between the literacy rate and the pupil-teacher ratio with an R value of approximately 0.736.

4 Discussion

The results of this study highlight the importance of various factors influencing literacy rates. As illustrated in Figure 1, the pupil-teacher ratio (PUP_TEACH_RATIO) emerged as the most influential feature in predicting literacy rates, with an importance value significantly greater than other features. In Figure 2, the Chi-Square Statistics value is much greater than the P-value, which proves its significance in determining the literacy rate. This is then further backed up by a strong negative correlation between the literacy rate and pupil-teacher ratio in Figure 3. A lower ratio likely enhances the quality of education by enabling more personalized instruction and improved classroom management, thus contributing to an increased literacy rate.

The second most significant factor was the Gross Enrollment Ratio (GROSS_ENROL_RATIO). This variable reflects the percentage of eligible children enrolled in educational institutions and serves as a key indicator of access to education. High enrollment rates often correlate with better literacy rates, as they signify greater participation in learning facilities.

Other factors, such as the Enrollment Rate in Primary (ENROLPRI), Secondary (ENROLSEC), Tertiary Education (ENROLTER), and Primary School Completion Rate (PRIRATE), displayed moderate importance. These findings suggest that while access to and completion of different education levels contribute to literacy, their influence is secondary to factors like pupil-teacher interaction.

Interestingly, government spending on education (GOV_EXP) and the percentage of GDP allocated to education (GEVSGDP) were found to have minimal impact on literacy rates. This may indicate that the mere allocation of financial resources is insufficient to ensure improved literacy rates. Instead, the method by which resources are distributed and utilized may have a more substantial impact. For instance, hiring qualified teachers and reducing class sizes would likely result in a non-negligible change to literacy outcomes. This finding contrasts with earlier studies, such as those by Baldacci et al. (2005), which highlights a positive correlation between public education spending and human capital development [18]. However, our results support more recent critiques of large-scale literacy improvement, including Pritchett's (2013) work, which suggests that the effectiveness of spending depends on how resources are managed and allocated [**?**].

These findings have important policy implications. Governments and education policymakers should prioritize strategies to reduce pupil-teacher ratios, particularly in regions with low literacy rates. Additionally, while ensuring high enrollment rates remains essential, complementary efforts should focus on the quality of instruction and resource allocation.

5 Conclusions

The purpose of this study was to identify and analyze factors that are significant contributors to the world's literacy rate. Based on the merged data, multiple analysis methods of Random Forest, the Chi-Square Test, and a correlation matrix were utilized. The Random Forest Regres-

sor model revealed that the pupil-teacher ratio influences literacy rates most, with an importance surpassing 0.5. This was further supported by the Chi-Square Test and the Correlation Matrix method. The Chi-Square Test proved its dependence of the two variables as the Chi-Square Statistic was greater than the P-Value. The correlation matrix depicts the pupil-teacher ratio as having the greatest size, underlying it is the most correlated relationship. As a result, this study leaves great potential for further pragmatic usage to education policymakers.

The significance of the pupil-teacher ratio is consistent with earlier studies highlighting its critical role in improving educational outcomes. For example, Graue et al. (2009) emphasizes that smaller class sizes and classroom quality synergize to produce learning opportunities for students [5]. These findings underscore the importance of teacher-student interaction in developing foundational skills.

Acknowledgements

This research has received no external funding. Thank you to the STEM Fellowship for organizing the National High School Big Data Challenge 2024 and making this paper possible.

References

[1] Organization for Economic Co-operation and Development. Skilled for life? key findings from the survey of adult skills. *Organization for Economic Co-operation and Development*, 2013.

[2] Ben Levin. The challenge of large-scale literacy improvement. *School Effectiveness and School Improvement*, 21(4):359–76, 2010.

[3] Shahadat Hossain Sujan Paul, Mohammad Rakibul Islam. The impact of student-teacher ratio on literacy rate of bangladesh: A study on 64 districts of bangladesh. *Journal of Business Analytics and Data Visualization*, 2(1):1–6, 2021.

[4] Beverly Carlson. Achieving educational quality. *Chile: Economic Commission for Latin America and the Caribbean, Restructuring and Competitiveness Network, Division of Production, Productivity and Management*, 2000.

[5] Kwok Chan Lai Peter Blatchford. Class size - arguments and evidence. *International Encyclopedia of Education*, (3):200–6, 2010.

[6] Melissa Sherfinski Elizabeth Graue, Erica Rauscher. The synergy of class size reduction and classroom quality. *The Elementary School Journal*, 110(2):178–201, 2009.

[7] Comfort Okpala Amon Okpala. The effects of public school expenditure and parental education on youth literacy in sub-saharan africa. *Journal of Third World Studies*, 23(2):203–12, 2006.

[8] D. E. Oriakhi Grace Ameh. Government expenditure and the development of the education sector in nigeria: An evaluation. *Review of Public Administration and Management*, 3(5):147–60, 2014.

[9] M. S. Rahman. Relationship among gdp, per capita gdp, literacy rate and unemployment rate. *British Journal of Arts and Social Sciences*, 14(2), 2020.

[10] Jia Qi Cheong Lianna Ooi Pei Wen. The relationship between literacy rate, primary school enrolment rate, death rate and gdp per capita. *International Journal of Academic Research in Business and Social Sciences*, 14(12), 2024.

[11] Sujan Paul. Literacy rate and some variables of primary education -a regression analysis. *The Bangladesh Accountant*, 2021.

[12] UNESCO Institute for Statistics. Gross enrolment ratio by level of education. 2024.

[13] Our World in Data. Pupils per qualified teacher in primary education. 2024.

[14] Organization for Economic Co-operation and Development. Expenditure on educational institutions as a percentage of gdp. 2024.

[15] Bushra Qurban. World education dataset.

[16] Geraldine E. Nzeribe Uju R. Ezenekwe Chiemezie D. Ukeje Maria C. Uzonwanne, Anoke E. Eze. Assessment of the efficiency of public education expenditure on literacy rate in nigeria. *Economic and Social Thought*, 7(1), 2020.

[17] JMP Statistical Discovery LLC. Chi-square test of independence.

[18] Benedict Clements Sanjeev Gupta Larry Q. Cui, Emanuele Baldacci. Social spending, human capital, and growth in developing countries: Implications for achieving the mdgs. *IMF Working Paper*, 2005.