Background

Chronic absenteeism is described as students missing at least ten percent of their enrolled school days and remains a concern in New York City's public schools. In September of 2023, the Department of Education revealed that 36% of students were chronically absent throughout the 2022-2023 school year (Chalkbeat New York). Despite this being a slight improvement from the previous academic year, chronic absence highlights the loss of valuable instruction time, which can significantly impact students' academic performance and educational outcomes (Kids Data). Furthermore, disparities based on race, gender and economic status often intersect with chronic absenteeism, emphasizing the need to address systemic issues and ensure equitable access to education for all students. Using a nuanced approach, such as machine learning, can allow educators and school officials to better understand how social factors affect student attendance and achievement. Thus, it raises the question: **How can machine learning uncover and address disparities in academic attendance and achievement related to race, gender, and economic status?**

This research aims to use machine learning to address school absenteeism and disparities in academic performance among students in New York City public schools. By developing a machine learning model, we can identify patterns and trends to improve attendance rates and learning outcomes for students.

Datasets

A total of four datasets from the New York City Public Schools InfoHub and NYC Open Data were used in this research. The datasets are listed as follows:

- https://infohub.nyced.org/reports/students-and-schools/school-quality/information-and-data-overview/end-of-year-attendance-and-chronic-absenteeism-data (End-of-Year Attendance and Chronic Absenteeism Data)
- https://infohub.nyced.org/reports/academics/test-results (ELA Test Results 2013 to 2023, Math Test Results 2013 to 2023)
- https://infohub.nyced.org/reports/students-and-schools/school-quality/information-and-data-overview (Demographic Snapshot 2018-19 to 2022-23)

The first dataset contains attendance data between 2017 and 2023 from students in every grade level. Per the New York City Department of Education (NYC DOE), attendance is attributed to the school the student attended at the time. If a student attends multiple schools in a school year, then the student will contribute data towards multiple schools. Starting in 2020-21, the NYC DOE transitioned to the New York State Education Department's (NYSED) definition of chronic absenteeism. This means that students are considered chronically absent if they have an attendance of 90% or less.

The second and third datasets were also collected by the NYC DOE and includes English Language Arts (ELA) and math test scores from 2013 to 2023 from students in 3rd grade to 8th grade at every school. This data is updated yearly, however does not include test scores from 2020 and 2021 due to the COVID-19 pandemic.

The final dataset came from the same data sources and provides demographic information such as disability status, English language learner (ELL status), race/ethnicity, gender and economic need.

All data was provided at the city, borough, district and school levels, however, for this research, school-level data was used because it provided a sufficient amount of data when merging the four datasets. The final dataset consists of 3354 rows and 27 columns with variables such as economic need index, number of days absent, percentage chronically absent, etc.

Approach

The data was imported by downloading the four separate datasets and reading them into RStudio. To clean the datasets, school years of interest were filtered from the datasets, which were 2018-19 to 2022-23. Then, each academic year was converted to a singular year to match year values in the math and ELA datasets. For example, 2018-19 was changed to 2019. Next, missing values were imputed into the math and ELA datasets using the mean test scores. Lastly, test scores were scaled to values between 0 and 1. Once the data cleaning process was complete, the datasets were merged and written to a CSV file.

Through exploratory data analysis, it was discovered that there were no strong correlations between race, gender and income with respect to attendance and academic performance. This was strange to see because research studies, such as Gee et al. (2023) and Bowen et. al (2022), suggest that chronic absence disparities are driven by socio ecological factors. The lack of strong linear relationships does not imply that there is no interaction between variables; rather it indicates the variables interact in a different manner.

Given the nature of the dataset, a neural network was used. This machine learning model is incredibly useful because it can capture linear and non-linear relationships and predict y factors from a combination of all x factors. In this case, the predictor variables were % Female, % Male, % Asian, % Hispanic, % Multi-Racial, % Native American, % White, % Students with Disabilities, % English Language Learners, % Poverty and Economic Need Index, while the outcome variables were % Attendance, % Chronically Absent, Mean Scale Score_e and Mean Scale Score_m.

Limitations/Challenges

Originally, I wanted to investigate how the timing of an absence can affect student academic performance, however the dataset did not provide any useful information such as date/time, type of absence or reason for absence. The data was numeric, thus, the absence of qualitative data hindered the exploration of potential relationships within the dataset, such as type of absence and gender. Having qualitative data to identify reasons for absence would have been beneficial in understanding the factors contributing to attendance patterns.

Standardized testing is also not the best measure of academic achievement/academic performance. This was used interchangeably in the analysis of the data, due to the lack of a large scale, quantitative and qualitative dataset that is necessary to paint the full picture of a student’s academic performance.

Conclusion

Using the weights from the neural network, we can conclude that racial, gender and economic disparities exist in academic achievement and school attendance. Male students reduce chronically absent predictions more than female students, thus female students are learned to be more chronically absent. Economic need index also contributes positively as the third most feature for chronic absence. Validation loss and training loss were both relatively small, approximately 0.15, and converge to the same value, thus the model generalizes the data well.

Future Works

In the future, I hope to uncover more racial, economic and gender disparities as it relates to school attendance across the United States. Leveraging a qualitative and descriptive dataset could aid in identifying reasons for absence and establishing support networks for students.

Works Cited

Bowen, Francis, et al. “Revealing Underlying Factors of Absenteeism: A Machine Learning Approach.” *Frontiers*, Frontiers, 7 Nov. 2022, www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.958748/full.

“Demographic Snapshot.” *Information and Data Overview*, infohub.nyced.org/reports/students-and-schools/school-quality/information-and-data-overview. Accessed 19 Apr. 2024.

“End-of-Year Attendance and Chronic Absenteeism Data .” *End-of-Year Attendance and Chronic Absenteeism Data*, infohub.nyced.org/reports/students-and-schools/school-quality/information-and-data-overview/end-of-year-attendance-and-chronic-absenteeism-data. Accessed 19 Apr. 2024.

“English Language Arts and Math State Tests.” *Test Results*, infohub.nyced.org/reports/academics/test-results. Accessed 19 Apr. 2024.

Gee, Kevin A, et al. “Explaining Disparities in Absenteeism between Kindergarteners with and without Disabilities: A Decomposition Approach.” *Science Direct*, Elsevier, 1 Feb. 2024, www.sciencedirect.com/science/article/pii/S0885200624000024#:~:text=Additionally%2C%20racial%20and%20ethnic%20disparities,peers%20(Gee%2C%202018).

“Reasons for School Absence in Past Month, by Gender and Grade Level.” *Kids Data*, PRB, www.kidsdata.org/table/159/new-haven-unified/2096/reasons-for-absence-gender. Accessed 19 Apr. 2024.

Zimmerman, Alex. “36% of NYC Public School Students Were Chronically Absent Last School Year.” *Chalkbeat*, Chalkbeat, 6 Sept. 2023, www.chalkbeat.org/newyork/2023/9/6/23862246/nyc-public-school-chronic-absenteeism-pandemic/.