4/19/2024

A Machine Learning Approach to Analyzing Tennessee Public School Letter Grades

Oluwasegun Adegboye, Mark Constantino, Jason Horne, Fuyima Inoue

April 19, 2024

East Tennessee State University

Course: CSCI 5270 – Machine Learning

Table of Contents

List of Figures

List of Tables

Background

Introduction

Despite the Tennessee Legislature having passed the law in 2017 (Tennessee Code Annotated, 2024), the Tennessee Department of Education did not create a system for assigning letter grades to schools until the Fall of 2023 (Tennessee Department of Education a, 2024). The purpose behind the letter grades, according to the Commissioner of Education, is to provide parents with a clear view of how their local school is performing while detractors claim that it is insulting to distill the work of a school down to one letter grade weighted heavily on standardized tests (West, 2023). The calculation of letter grades for Tennessee schools is based on student performance on the Tennessee Comprehensive Assessment Program (TCAP) tests (Tennessee Department of Education a, 2024). This calculation encompasses several metrics:

- Achievement: This metric represents the percentage of students achieving proficiency on the exam.
- Growth: A statistical measure that evaluates students' performance against their previous scores or predicted outcomes.
- Growth25: Specifically tracks the progress of the bottom 25% of performers within a school, focusing on those in particular subgroups.
- For high schools, an additional metric, College and Career Readiness (CCR), is included. This is assessed through a combination of ACT scores, industry-recognized credentials, ASVAB scores, and enrollment in postsecondary courses.

The Tennessee Department of Education scores each category according to established criteria which then they weight to produce a final grade on a 5-point scale that corresponds to a traditional letter grade, A-F. The weighting for elementary and middle schools is as follows: Achievement (50%), Growth (40%), and Growth25 (10%). In high schools, the weights are adjusted to: Achievement (50%), Growth (30%), Growth25 (10%), and CCR (10%) (Tennessee Department of Education a, 2024).

The significance of this study will be to illuminate variables in school data that impact these letter grade scores beyond how students score on the test. We aim to identify and quantify the influence of demographic factors, school resources, and other non-test score-related variables on the overall grading system by employing a machine learning approach. This analysis will not only provide a deeper understanding of the factors contributing to the assigned letter grades but also offer insights into potential areas for improvement within the schools. Ultimately, the findings could inform policymakers and educational leaders, leading to more nuanced and equitable educational strategies that recognize the diverse challenges and strengths of Tennessee Schools.

Problem Statement

Critics of standardized testing have long contended that it simplifies the complexities of education too much (Stake, 1991), positing that students from more advantageous demographics tend to fare better on these tests (Green & Celkan, 2011). Preliminary analyses of the data from Tennessee's Letter Grade Scores seem to corroborate this, revealing discernible demographic effects on school grades (Horne, 2024). In light of these concerns, our study adopts a machine learning approach to meticulously examine which demographic factors and scores serve as the most significant predictors of a school's letter grade. Furthermore, we aim to analyze the weight of these predictors' coefficients, offering a nuanced understanding of their impact. This approach seeks not only to identify patterns within the grading data but also to contribute to a more equitable assessment framework by highlighting underlying biases and inequalities.

Motivation

The motivation behind this study is multifaceted, driven by a commitment to educational equity, the quest for a deeper understanding of assessment practices, and the potential for data-driven decision-making to enhance educational outcomes. Our research involves the critical examination of standardized testing and the resultant school letter grading system implemented in Tennessee. The central concern is the equitable assessment of school performance. Standardized testing, while a uniform measure, often fails to account for the diverse socio-economic, cultural, and environmental contexts in which students learn. This oversight can lead to the perpetuation of educational disparities, with schools in underprivileged areas unfairly penalized by a system that does not fully capture their challenges or achievements.

By exploring beyond test scores to include demographic and other influential factors, our study aims to shed light on these disparities, advocating for a more just and inclusive approach to school evaluation. The introduction of a letter grade system to evaluate schools, based on standardized test scores among other factors, raises significant questions about the criteria used for these assessments. Our investigation into the variables that impact letter grades is motivated by a desire to unpack these practices, critically analyzing their validity and the implications for schools and communities. This understanding is crucial for educators, policymakers, and stakeholders to evaluate the effectiveness and fairness of the grading system.

In an era where data analytics has transformed numerous sectors, applying a machine learning approach to analyze school performance data represents an innovative step toward evidence-based education policy.

By identifying key predictors of school grades, our study aims to offer actionable insights that can guide interventions, resource allocation, and policy reforms. The potential to use data to highlight areas of need, predict outcomes, and tailor educational strategies is a powerful motivator, promising to enhance the educational experience for all students. Ultimately, the motivation for this research is to contribute to the ongoing dialogue around educational assessment and policy. By providing a nuanced analysis of the factors influencing school letter grades, we hope to inform policy debates and inspire changes that recognize the complexity of education. Our goal is to ensure that assessment practices are fair, transparent, and conducive to the growth and development of every student, regardless of their background.

Research Questions

RQ1: What demographic factors are most predictive of school letter grades in Tennessee?

RQ2: How do school resources and other non-test score-related variables influence the overall grading system?

RQ3: Can machine learning models accurately predict school performance grades and identify areas for improvement?

RQ4: What inherent patterns and clusters can be identified among Tennessee public schools based on performance indicators and demographic variables using the K-Means clustering algorithm, and how do these groupings inform potential educational interventions and resource allocations?

Literature Review:

Previous research has consistently demonstrated a correlation between socioeconomic status, demographic factors, and academic performance. For instance, Hirn et al. (2018) analyzed the extent to which socioeconomic disparities influence student learning outcomes, concluding that lower socioeconomic status is significantly linked to poorer academic performance. Similarly, Latour and Tissington (2011) examined the role of demographic factors such as race and income level, finding that these elements are crucial predictors of educational success and often correlate with disparities in school resources and student support systems.

Further compounding these findings, Rutkowski et al. (2018) explored the impact of social determinants on academic achievements across multiple regions, highlighting that poverty, ethnicity, and access to educational resources play pivotal roles in shaping student performance. Their work specifically pointed out that schools serving higher concentrations of disadvantaged students often receive lower evaluation scores, which can perpetuate a cycle of underfunding and underperformance.

Building on these studies, this project hopes to delve deeper into the Tennessee education system to analyze how these factors affect the assignment of letter grades to schools. By employing a machine learning approach, we intend to uncover more nuanced insights into how socioeconomic and demographic variables influence school grades. This analysis is particularly pertinent given the recent implementation of a standardized letter grading system by the Tennessee Department of Education, which aims to provide clear and accessible evaluations of school performance to parents and stakeholders.

This project will also consider the criticism that such grading systems oversimplify the complexities of educational environments and may fail to account for the broader socio-economic challenges faced by schools. In light of these concerns, our study seeks to not only identify the determinants of school grades but also propose ways to enhance the fairness and effectiveness of these evaluations. Through this research, we hope to contribute valuable insights to policymakers and educational leaders, aiding in the development of more equitable educational strategies that recognize and address the diverse challenges of Tennessee schools.

In summary, by integrating and building upon the existing body of literature on the impact of socio-economic and demographic factors on school performance, this study aims to provide a comprehensive analysis of the newly implemented letter grading system in Tennessee. Our goal is to inform ongoing educational debates and support the creation of policies that ensure all students have the opportunity to succeed, regardless of their background.

Analysis

Dataset

The dataset used was combined from two publicly available downloads on the TN Department of Education's Data Downloads and Requests page (Tennessee Department of Education b, 2024). The files used were the 2022-23 A-F Letter Grade File (Tennessee Department of Education c, 2024) and the 2021-2022 Membership File (Tennessee Department of Education d, 2024). We merged the two datasets using a primary key made up of a concatenation of the district number and school number (Horne, Jason Horne's Github, 2024).

The A-F Letter Grade Dataset (Tennessee Department of Education c, 2024) has 1900 rows and 74 columns while the 2021-2022 Membership Dataset (Tennessee Department of Education d, 2024) has 1823 rows and 23 columns. Before merging, we filtered out any schools that were ineligible for letter grades. This left us with a dataset containing 1670 rows and 89 columns.

Features

An overview of the features in the dataset is as follows:

Core Information

- Year: The academic year for which the data was collected.
- System: A numerical identifier for the school district.
- System Name: The name of the school district.
- School: A numerical identifier for individual schools within a district.
- School Name: The name of the individual school.

School Classification

- Lg_Ineligible: Indicates whether a school is ineligible for a letter grade (e.g., due to not serving the grade levels assessed by standardized tests).
- School Pool: Categorizes schools based on their level (e.g., Elementary, Middle, High, K-8).
- Grade Band 3-5, 6-8, 9-12: Indicates the grade bands served by the school, essential for understanding the scope of students' education levels.

Performance Scores and Weights

- Ach_Score: Achievement score based on standardized test results.
- Growth_Score: Measures student growth over time.
- Growth25_Score: Growth score specifically for the bottom 25% of performers.

- Ccr_Score: College and Career Readiness score, evaluating how well students are prepared for post-secondary success.
- Ach_Score_Weighted, Growth_Score_Weighted, etc.: Weighted scores that factor into the final grading, reflecting the relative importance of each metric.
- Ach_Weight, Growth_Weight, etc.: The weight assigned to each scoring category in the overall evaluation.

Letter Grade and Overall Success Rates

- Lg_Score: The calculated score that determines the letter grade.
- Lg_Grade: The letter grade assigned to the school based on the lg_score.
- Overall_Success_Rate_All_Students: The overall success rate of all students across various metrics.
- Overall_Success_Rate_*: These features (e.g., ed, el, swd, aian, asian, etc.) represent the overall success rates for specific demographic groups (economically disadvantaged, English learners, students with disabilities, American Indian or Alaska Native, Asian, etc.).

Success Rates by Grade and Subject

- Success_Rate_G3-5_Ela, Math, Science, etc.: Success rates in English Language Arts, Math, and Science for specific grade bands, offering detailed insight into academic performance across subjects.

Growth Scores by Subject and Demographics

- Growth_Numeracy_Score, Growth_Literacy_Score, etc.: Subject-specific growth scores.

- Growth_Ela_Math_Score_*: Growth scores in ELA and Math for specific demographic groups, providing a granular look at academic progress.

College and Career Readiness Rates

- Ccr_Rate: Overall rate of college and career readiness.
- Ccr_Act_Rate, Ccr_Postsec_Rate, etc.: Specific metrics contributing to the CCR rate, such as ACT performance and post-secondary enrollment rates.
- Ccr_Rate_*: CCR rates for specific demographic groups, highlighting disparities or successes in preparing students for their future careers.

Data Cleaning

In the process of preparing our dataset for analysis, a critical step involved streamlining the data to ensure its relevance and effectiveness in addressing our research questions. While our logistic regression model is capable of handling multiple features, we opted for a strategic reduction of variables to enhance the clarity and focus of our study. Specifically, we decided to eliminate the college and career readiness (CCR) data. This decision was informed by our intention to analyze data across all school levels—elementary, middle, and high schools—collectively, necessitating a dataset that uniformly applies across these tiers without the specialization inherent to CCR metrics. In addition, there were 1539 schools in the K8 pool versus 359 in the high school pool.

Consequently, the refined dataset retains a focused set of columns that are pertinent across the educational spectrum, facilitating a comprehensive analysis. The preserved features include:

- School Identifiers: 'system_name' and 'school_name' provide essential context about the educational institutions.
- Performance Scores: Metrics such as 'ach_score' (achievement score), 'growth_score' (measuring academic progress), and 'growth25_score' (growth of the bottom 25% of students) offer insights into school performance.
- Letter Grade Data: 'lg_score' and 'lg_grade' directly relate to the overarching study focus, translating performance metrics into evaluative grades.
- Overall Success Rates: This includes 'overall_success_rate_all_students' and success rates segmented by demographic categories like economically disadvantaged ('overall_success_rate_ed'), English learners ('overall_success_rate_el'), and students with disabilities ('overall_success_rate_swd').
- Subject-Specific Growth Scores: These detail growth in numeracy, literacy, science, and social studies, alongside growth scores for various demographic groups, providing a nuanced view of academic progress.
- Demographic Percentages: Including 'african_american_pct', 'asian_pct', 'economically_disadvantaged_pct', among others, these variables allow for the examination of demographic influences on school performance.

By focusing on these variables, our study aims to distill the complex dynamics of school performance into actionable insights, bridging academic achievements with demographic factors to unearth patterns that may inform educational strategies and policies. This selective approach ensures our analysis remains robust, relevant, and aligned with our research objectives, offering a meaningful exploration of the factors contributing to the letter grades assigned to Tennessee schools. In addition to streamlining our dataset by selecting relevant columns, we implemented a series of data cleaning steps to ensure the integrity and uniformity of the dataset, facilitating accurate analysis and interpretation. These steps are crucial in preparing the data for regression models and include the following adjustments:

1. Standardizing Percentage Representations**:** In our dataset, percentage values represented as "Less than 5%" and "More than 95%" were standardized to numerical values to enable quantitative analysis. Specifically, "Less than 5%" was replaced with 2.5, reflecting an estimate within the indicated range, while "More than 95%" was substituted with 97.5. This conversion was applied across all columns where such representations were found, ensuring consistency in data interpretation.
2. Converting Data Types**:** We identified columns represented as percentage data and ensured that it was formatted as numeric data types to facilitate accurate statistical analysis. This involved converting columns such as 'african_american_pct', 'asian_pct', 'economically_disadvantaged_pct', among others, to numeric values, accounting for potential non-numeric entries that could impede analysis. We used the pd.to_numeric function with error handling to coerce incompatible values into a manageable format during the conversion process to avoid data loss or distortion.

By executing these data cleaning steps, we improved the dataset's analytical readiness, ensuring that percentage values are accurately represented and uniformly processed across all relevant columns. This preparation is critical for the subsequent analysis, as it guarantees the reliability of the data inputs into our models, thus bolstering the validity of our findings. These adjustments, along with the feature selection process, form the comprehensive data cleaning section of our methodology, underscoring our commitment to rigorous data analysis practices. None of the data had to be encoded or imputed. Resulting in one less step in the data cleaning process.

Descriptive Statistics

In our attempt to discover the complexities behind the letter grades assigned to Tennessee public schools, descriptive statistics serve as our initial guide. We aim to provide a snapshot of the current educational landscape across Tennessee schools by analyzing the distribution, central tendencies, and variability of our collected data. These statistics offer the first glimpse into the dataset's patterns and outliers, painting a broad picture of school performance, demographic compositions, and other crucial factors influencing educational outcomes. Before delving into complex predictive models and inferential statistics, understanding the basic structure and trends within our data is essential. This approach not only increases the clarity of our analyses but also ensures that our interpretations and conclusions are grounded in reality.

In analyzing the central measures of tendency for pivotal statistics within our educational data, we observe notable patterns. For the 'Overall Success Rate All Students', the standard deviation is 17.40, indicating variability around a mean score of 36.97 on a 100-point scale. On the other hand, the 'LG Score', 'Growth Score', and 'Growth25 Score' are all gauged on a 5-point scale. The 'LG Score' demonstrates a mean of 3.24, suggesting that on average, schools are scoring slightly above the midpoint. The 'Growth Score' has a mean of 3.14, indicating a similar trend, while the 'Growth25 Score' averages slightly higher at 3.37, which may point to relatively more positive growth among the lower-performing quartile of students.

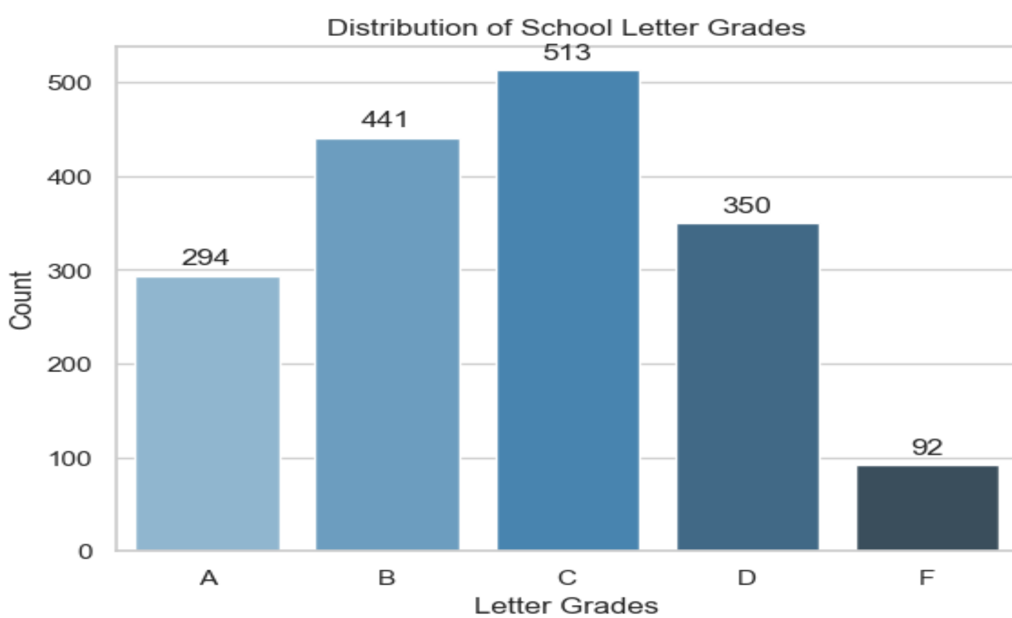
Table 1:- Central tendency and measure of variations of the score of school grades.

	overall_success_rate_all_students	***lg_score***	***growth_score***	***growth25_score***
count	1653.000000	1653.000000	1653.000000	1606.000000
mean	36.967877	3.235269	3.136116	3.372354
std	17.395222	1.097657	1.583888	1.185933
min	5.100000	1.000000	1.000000	1.000000
25%	24.700000	2.400000	1.000000	3.000000
50%	35.700000	3.200000	3.000000	3.000000
75%	47.200000	4.100000	5.000000	4.000000
max	94.900000	5.000000	5.000000	5.000000

The data presented in Table 1 reflects the central tendency and variation within the scores, offering a clear perspective of where schools stand on these scales. The measures reveal that, generally, schools are performing around the midpoint or above on the LG scale, exhibiting growth in both overall and specific subgroups as identified by the 'Growth Score' and 'Growth25 Score'.

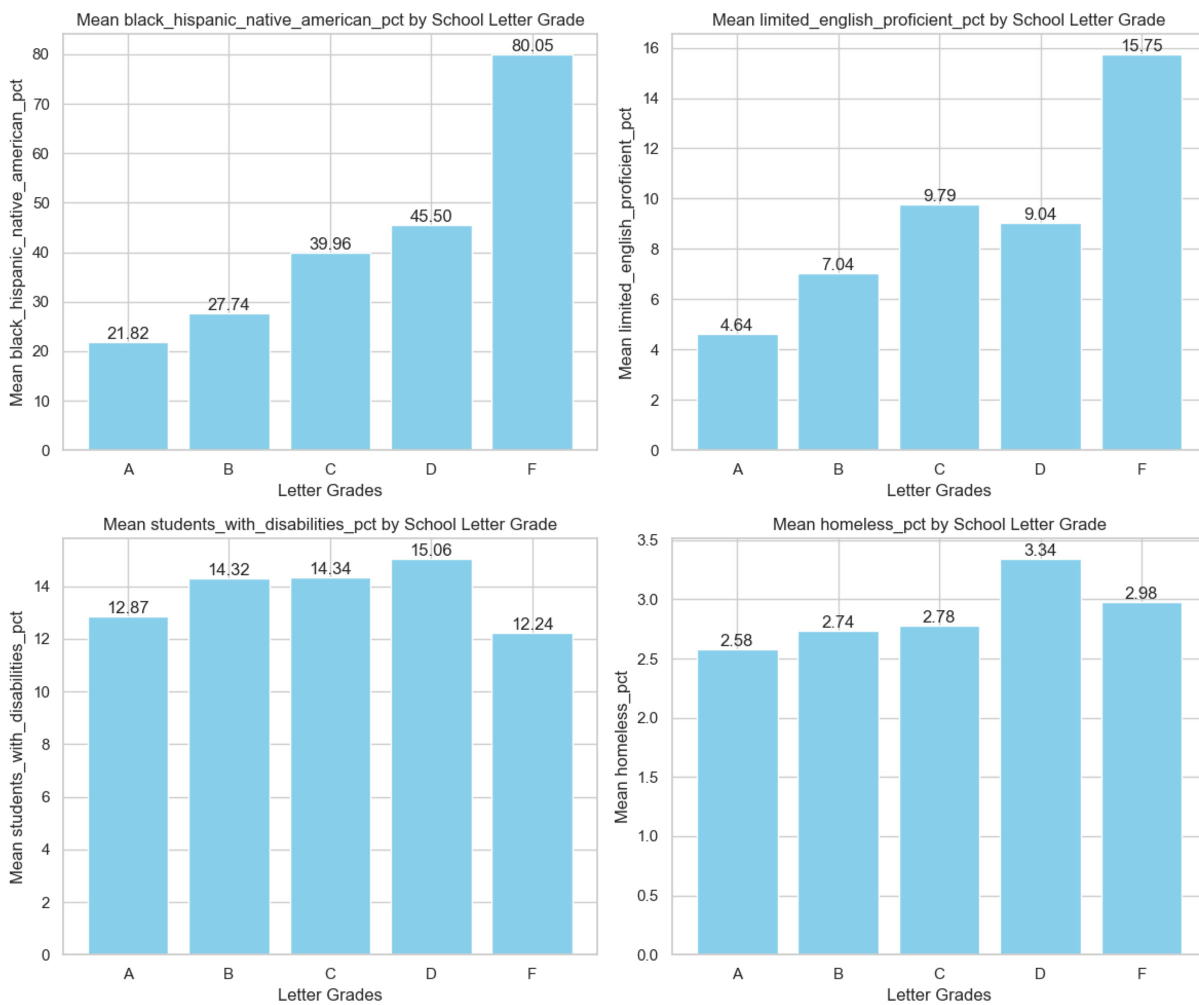
Figure 1 (below) illustrates the distribution of school letter grades within a dataset. It's evident that the majority of schools fall within the 'B' and 'C' grade categories, with counts of 441 (26.09%) and 513 (33.36%) respectively, indicating a central tendency around these grades. Grades 'A' and 'D' have lower frequencies, with 'A' grades occurring in 294 (17.39%) schools and 'D' in 350 (20.71%) schools. Notably, 'F' grades are the least common with only 92 (5.44%) occurrences. This distribution suggests that while a significant number of schools perform at an average level ('B' and 'C'), there is a smaller but substantial group achieving high performance ('A') and another set of schools that are notably underperforming ('D' and 'F').

Figure 1:- Distribution of school letter grades



Looking at the percentage of students in subgroups across the letter grades, we created a bar chart (Figure 2) for each major subgroup. In the top-left chart, we observe the mean percentage for 'Black, Hispanic, Native American' students across different school grades. Schools with an A grade have the lowest mean percentage of around 21%, which steadily increases across the letter grades, peaking at schools with a D grade at nearly 46% before slightly decreasing at F grade schools.

Figure 2:- Chart of letter grades vs other determinants of school letter grades.



The top-right chart displays the mean percentage of 'Limited English Proficient' students. A-grade schools have a mean percentage just below 5%, with this figure rising as the letter grades descend, reaching the highest mean of over 9% for schools with a D grade. The bottom-left chart highlights the mean percentage of 'Students with Disabilities'. The trend increases from A to C grade schools, fluctuating around 14%, and then jumps to over 15% for schools graded D, before decreasing in schools with an F grade.

Lastly, the bottom-right chart provides the mean percentages for 'Homeless' students. Unlike the other charts, this one show less variability across the letter grades, with percentages hovering around 2-3%. There is a slight increase for schools with a D grade and another marginal rise for those with an F grade. A general trend is apparent across all charts where schools with a D grade tend to have higher mean percentages of these demographic groups, possibly suggesting a correlation between these demographics and lower school letter grades. Conversely, A-grade schools consistently have the lowest mean percentages in these categories. This indicates a potential area for further investigation into the factors influencing school performance and the effectiveness of the support structures in place for these student groups.

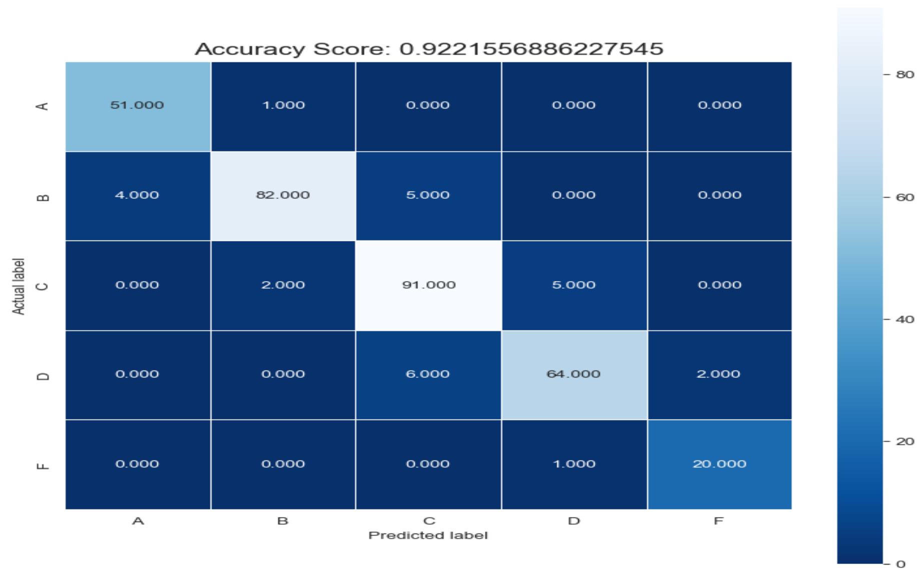
Regression Models

Logistic Regression

We applied logistic regression to predict school performance grades, denoted by letter grades A through F. Then we evaluated the model's performance using a confusion matrix and Receiver Operating Characteristic (ROC) curves. The confusion matrix (see: Figure 3) for the logistic regression model reveals a nuanced capability to distinguish between the different grades. The majority of predictions align with the true labels for each class, with particularly high accuracy in predicting grades B (Class 1) and C (Class 2).

Conversely, some confusion exists, however, between adjacent grades because of the similarity between the features of contiguous performance categories.

Figure 3: Logistic-Regression Confusion-Matrix



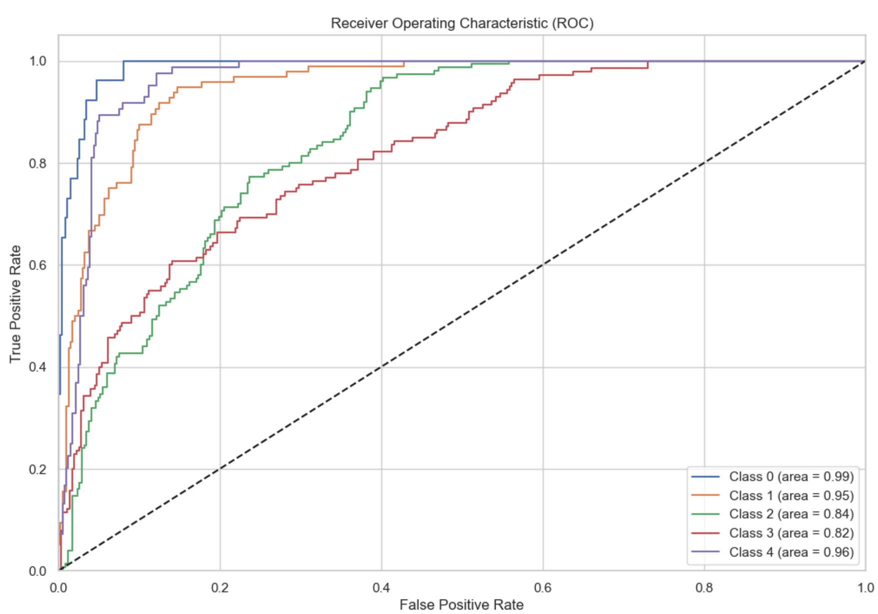
The precision and recall for each grade are as follows:

- Grade A (Class 0) shows high precision and recall, with a notable ability to correctly identify true instances of this grade.
- Grade B (Class 1) demonstrates solid precision, reflecting the model's capacity to accurately predict this grade when it does.
- Grade C (Class 2) exhibits a balanced precision and recall, indicative of the model's overall proficiency in classifying instances into this common category.
- Grade D (Class 3) and Grade F (Class 4) present competitive precision and recall values, suggesting reliable identification despite smaller sample sizes.

The ROC analysis further corroborates these findings, with AUC values providing a single measure of model performance across various decision thresholds:

- Grade A (Class 0): AUC of 0.99, indicating an excellent capability to discriminate between high-performing schools and others.
- Grade B (Class 1): AUC of 0.95, showing the model's effectiveness in identifying above-average performance.
- Grade C (Class 2): AUC of 0.84, pointing to a good discriminative power for average-performance schools.
- Grade D (Class 3): AUC of 0.82, which denotes a solid predictive ability for below-average performance.
- Grade F (Class 4): AUC of 0.96, reflecting a strong performance in distinguishing schools with the lowest grades.

Figure 4: Logistic Regression ROC Curve



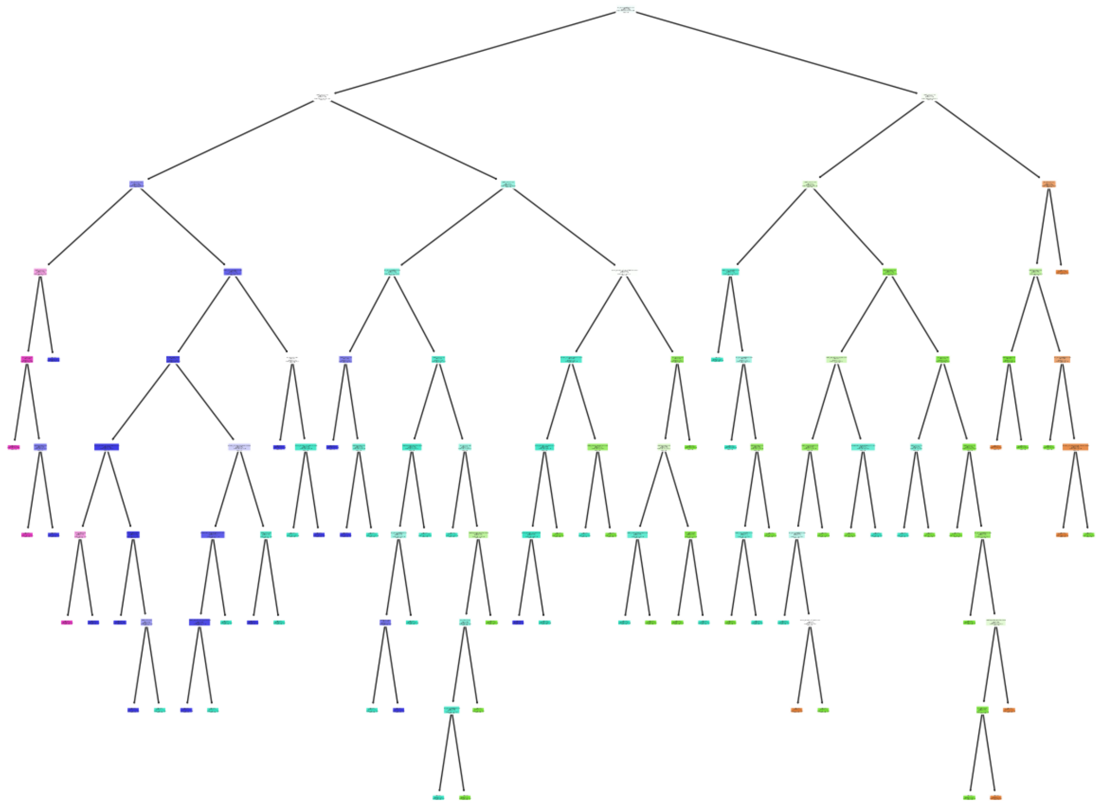
These ROC and AUC results underscore the model's competency in discriminating between the performance levels of schools across the entire spectrum of grades, from high achievers to those who are struggling. The logistic regression model provides a dependable analytical framework for understanding and predicting educational outcomes, thereby supporting educational administration and policy-making with actionable insights.

Classifier Models

Decision Tree

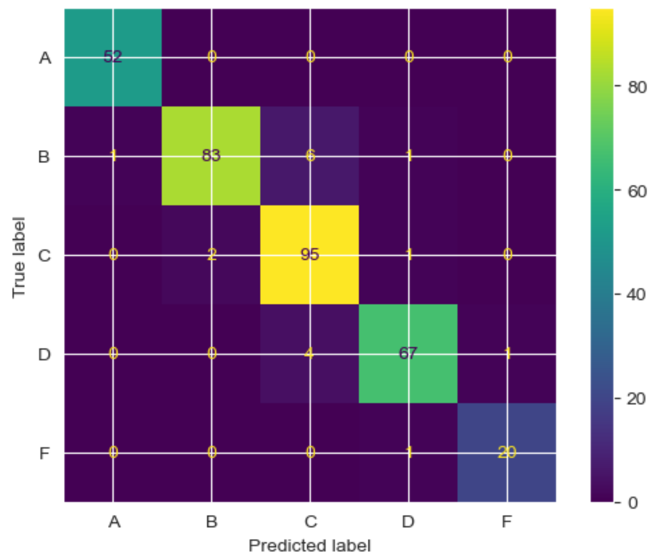
Our decision tree model, as visualized in Figure 3, reveals an intricate structure of decision-making pathways, indicative of the complex factors at play in determining school letter grades. This depth and breadth highlight the model's consideration of a multitude of variables, but they also underscore a complexity that could challenge straightforward interpretation. The extensive branching mirrors the multifaceted nature of educational performance assessment, considering diverse metrics from academic achievements to demographic factors.

Figure 5: Decision Tree Plot Tree



A pivotal component of our classifier model evaluation is the confusion matrix (see Figure 4), a visualization that illuminates the model's predictive precision across the letter grade spectrum. Notably, the matrix illustrates a commendable predictive accuracy, particularly in classifying schools with a 'C' grade, which could suggest a central tendency in the data. However, the matrix also reveals less accuracy in distinguishing 'B' graded schools, inviting further scrutiny into the model's classification criteria and the underlying characteristics of these institutions.

Figure 6: Decision Tree's Confusion Matrix



The balance between a model's ability to learn from the training set and its performance on unseen data is a crucial aspect of machine learning. This balance is depicted in our 'Accuracy vs. Alpha' plot, which demonstrates the model's performance at varying levels of alpha regularization. There is a plateau indicating an optimal alpha range that maximizes accuracy while mitigating the risks of overfitting. This optimal point maintains stable accuracy levels across the training and testing datasets, a testament to the model's robustness. Our Decision-Tree model exhibits exceptional performance, as reflected by an accuracy score of approximately 95%. The detailed metrics — precision, recall, f1-score — are consistently high across all letter grades, with particularly noteworthy precision in the prediction of 'A' grades. This consistency indicates not only the model's capability to accurately generalize across different grades but also its potential utility in practical applications for educational assessment and planning.

The results from our Decision-Tree model (see Table 2) offer profound insights into the factors influencing school letter grades. The high degree of accuracy suggests that the model successfully captured the complexities of the grading system. This reinforces the notion that school performance, as encapsulated by letter grades, can be systematically analyzed, and understood through a machine learning lens, which may yield meaningful implications for policy formulation and educational strategy development.

Table 2: Decision Tree Classifier Results

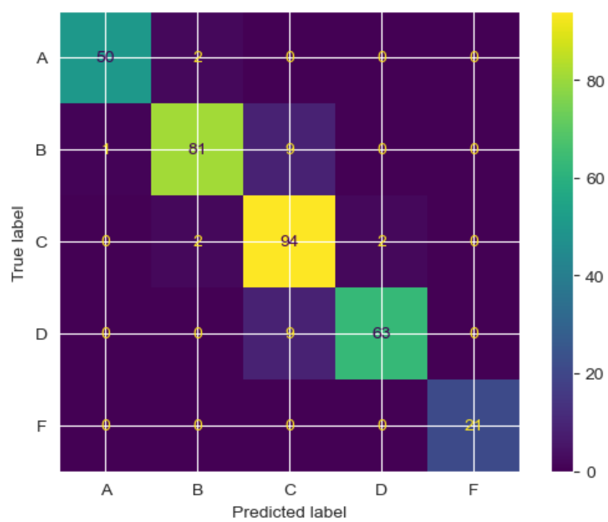
	Precision	**Recall**	**F1-Score**	**Support**
A	0.98	1.00	0.99	52
B	0.98	0.91	0.94	91
C	0.90	0.97	0.94	98
D	0.96	0.93	0.94	72
F	0.95	0.95	0.95	21
Accuracy			0.95	334
Macro Avg.	0.95	0.95	0.95	334
Weighted Avg.	0.95	0.95	0.95	334

The overall accuracy score for the Decision-Tree model is 0.9401. These results pave the way for a rich discussion on the implications of machine learning in educational evaluation. We will delve into how our findings correlate with current educational policies and practices, explore the potential for data-informed decision-making, and consider the broader impact on the quest for equitable and effective education.

Random Forrest

The Random Forest classifier achieved a commendable accuracy score of approximately 92.5%, indicating a strong predictive performance across the different letter grades A through F. The confusion matrix (see: Figure 5) further illustrates this point, showcasing a high number of correct predictions across the diagonal, where true labels and predicted labels align.

Figure 7: Random Forests Confusion-Matrix



In terms of precision, which measures the correctness of the predicted positives, the model excelled with near-perfect precision for grades A and F (0.98 and 0.95, respectively), and impressively high precision for grades B, C, and D. Recall, which evaluates how well the model captures actual positives, was particularly high for grades C and F, indicating that the model has a strong ability to identify these grades among the test samples. The F1-score, a balanced measure of precision and recall, was notably high for all grades, with the top scores for grades A and F (0.98 each), displaying the model's balanced classification capabilities for these categories. Grades B, C, and D also received robust F1-scores, further underscoring the model's overall effective performance. The support, or the number of true instances for each grade, varied across the grades, but the model managed to maintain high levels of accuracy regardless of group size, indicating its effectiveness across varied sample sizes. In summary, the Random Forest model displayed a high level of accuracy, with substantial precision and recall across all letter grades, solidifying its potential for practical application in educational assessment contexts.

Feedforward Neural Network with Keras

The performance of our Feedforward Neural Network, constructed and refined via the Keras framework, has been thoroughly evaluated. The confusion matrix delineates a comprehensive picture of the model's predictive capabilities. With an accuracy score of approximately 92.5%, the model displays a commendable predictive performance.

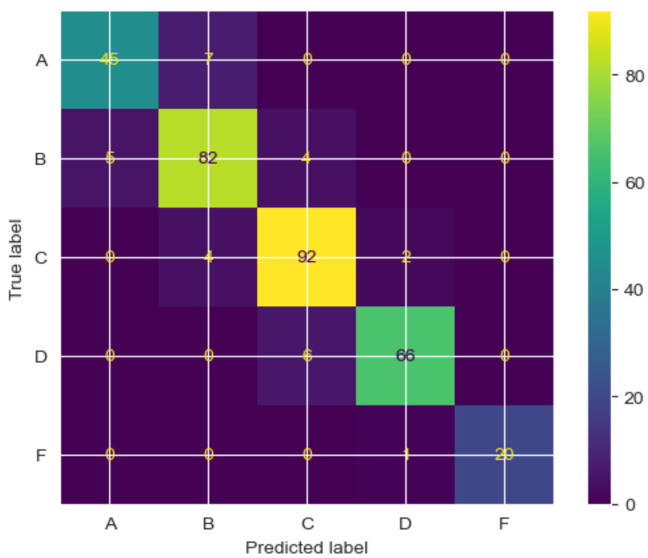


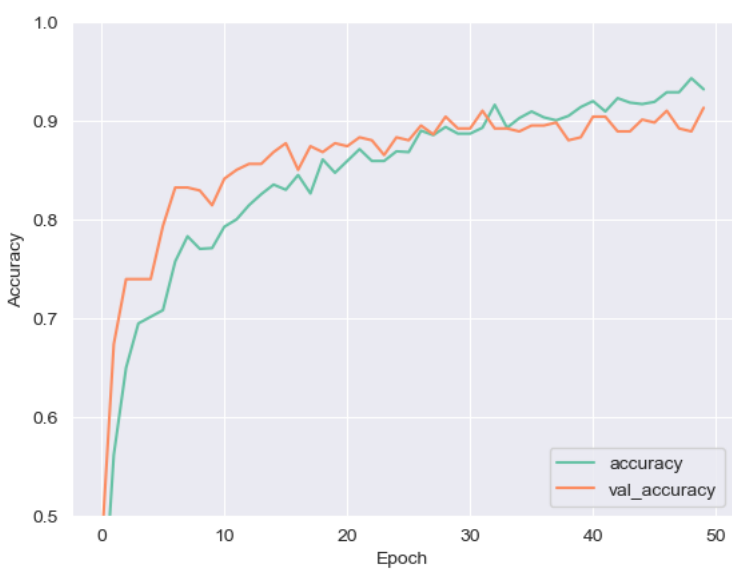
Figure 8: Feedforward Neural Network Confusion-Matrix

Precision metrics across the different letter grades are robust, with grades A and F reaching a precision as high as 0.98 and 0.95, respectively. Recall for grade C stands out at 0.95, indicating the model's strong ability to identify this category accurately. This is consistent with the confusion matrix, where the highest number of correct predictions appears for grade C, capturing 92 out of 98 instances correctly.

The F1-scores mirror this high level of precision and recall, highlighting the model's balanced approach in categorizing schools with respective grades. Notably, the model presents exceptional identification of schools with grades A and F, as evidenced by the F1-scores of 0.98.

The learning curves for the model (see: Figure 7), reflected through the progression of training and validation accuracy, exhibit a positive trend. The model reaches a plateau near a high accuracy rate, indicating a convergence and suggesting that the model is well-tuned with a minimal gap between learning from the training set and generalizing to the validation set.

Figure 9: Feedforward Neural Network Training Accuracy



In summary, our Neural Network's training process culminates in a high degree of accuracy, demonstrating effectiveness and reliability in predicting educational outcomes, as visualized in the confusion-matrix and learning curves.

Extreme Gradient Boosting

Utilizing the XGBoost (Extreme Gradient Boosting) ensemble learning method, the model's performance has been meticulously optimized through an extensive hyperparameter tuning process, involving 324 fits across three folds for each of the 108 candidates. The chosen hyperparameters include a column sample by tree of 0.7, learning rate (eta) of 0.1, a learning rate of 0.01, a maximum depth of 7, 300 estimators, with an objective set to 'multi:softmax' for a multiclass classification of 5 classes, and a subsample rate of 1. The tuning process yielded a model that significantly improved the classification accuracy, achieving an impressive accuracy score of 95.8%. This level of accuracy indicates the model's exceptional capability to generalize from the training data to unseen data.

The corresponding confusion-matrix (see: Figure 8) provides further insight into the model's predictive performance, revealing the number of correct and incorrect predictions for each class. Notably, the model shows exceptional strength in predicting categories B, C, and D, with high numbers of correct classifications, as evidenced by the bright shades in the confusion matrix.

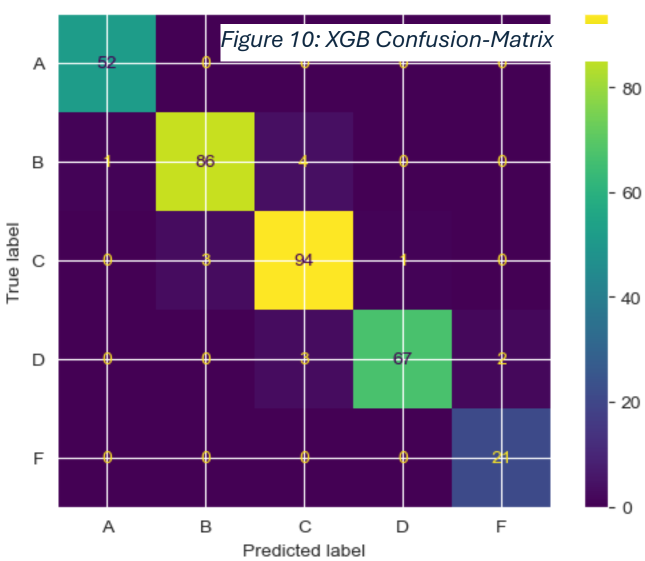


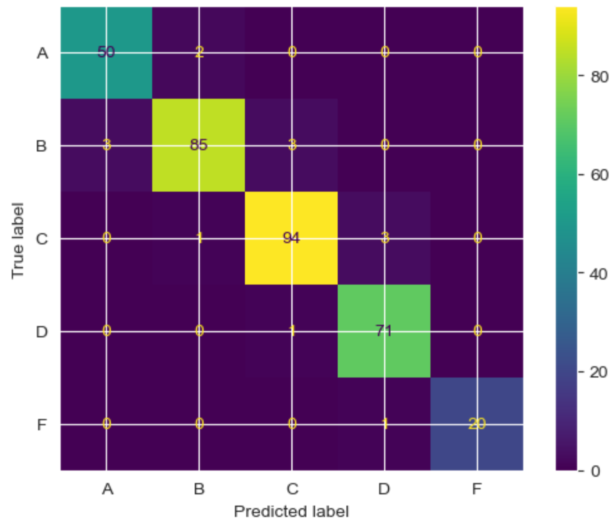
Figure 10: XGB Confusion-Matrix

The XGBoost model's ability to discern between the different letter grades suggests a strong predictive power, with high precision and recall across the board expressed in the F1-scores. This robust performance underscores the effectiveness of the model in educational outcome predictions and supports its potential application in identifying and addressing the nuanced needs of various educational institutions.

Neural Network MLPClassifier

An MLPClassifier, a multi-layer perceptron model which is a type of feedforward artificial neural network, was employed to predict categorical school grades. The model's classification report indicates exemplary precision and recall scores across all grade categories, with a particularly perfect precision score of 1.00 for grade 'F'. The confusion-matrix visualizes the distribution of predictions against the true labels. High counts on the matrix's diagonal, particularly for grades 'B', 'C', and 'D', demonstrate the model's accuracy and its ability to distinguish effectively between the classes. This is consistent with the high f1-scores reported, which provide a harmonious mean of precision and recall, indicating the model's balanced capability in both the identification and prediction of the correct classes.

Figure 11: MLPClassifier Confusion-Matrix



Overall, the MLPClassifier has achieved a micro, macro, and weighted average score of 0.96 across precision, recall, and the f1-score, reflecting an exceptionally consistent performance across all metrics. The high scores across these averages highlight the model's robustness and its potential for implementation in educational assessment tools. The model was most successful in accurately classifying the 'C' grade, correctly identifying 94 out of 98 instances, as well as demonstrating near-perfect classification of the 'D' grade with 71 correct out of 72 instances. These results underscore the Neural Network's nuanced

understanding of the feature space and its ability to make highly accurate predictions.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.94	0.96	0.95	52
1	0.97	0.93	0.95	91
2	0.96	0.97	0.96	98
3	0.95	0.99	0.97	72
4	1.00	0.95	0.98	21
Micro Avg.	0.96	0.96	0.96	334

Macro Avg.	0.96	0.96	0.96	334
Weighted Avg.	0.96	0.96	0.96	334
Samples Avg.	0.96	0.96	0.96	334

Table 3: MLP Classifier Results

In conclusion, the MLPClassifier's performance, as evidenced by the confusion matrix and classification report, suggests that it is a highly effective tool for predicting school performance grades, providing valuable insights into educational outcomes.

Unsupervised Learning

K-Means Clustering

In exploring the dataset through unsupervised learning, we applied the K-Means clustering algorithm to identify inherent groupings within the school data based on a variety of performance indicators and demographics. The optimal number of clusters was determined using the elbow method, which suggested a clear inflection point at k=5, indicating that five distinct groups exist within the dataset. The 3D scatter plot (see: Figure 10) illustrates the spatial distribution of the schools across these five clusters, demonstrating meaningful differentiation among them. A pie chart was utilized to provide a clear representation of the relative size of each cluster, showing a relatively balanced distribution with Cluster 2 being the largest and Cluster 1 being the smallest.

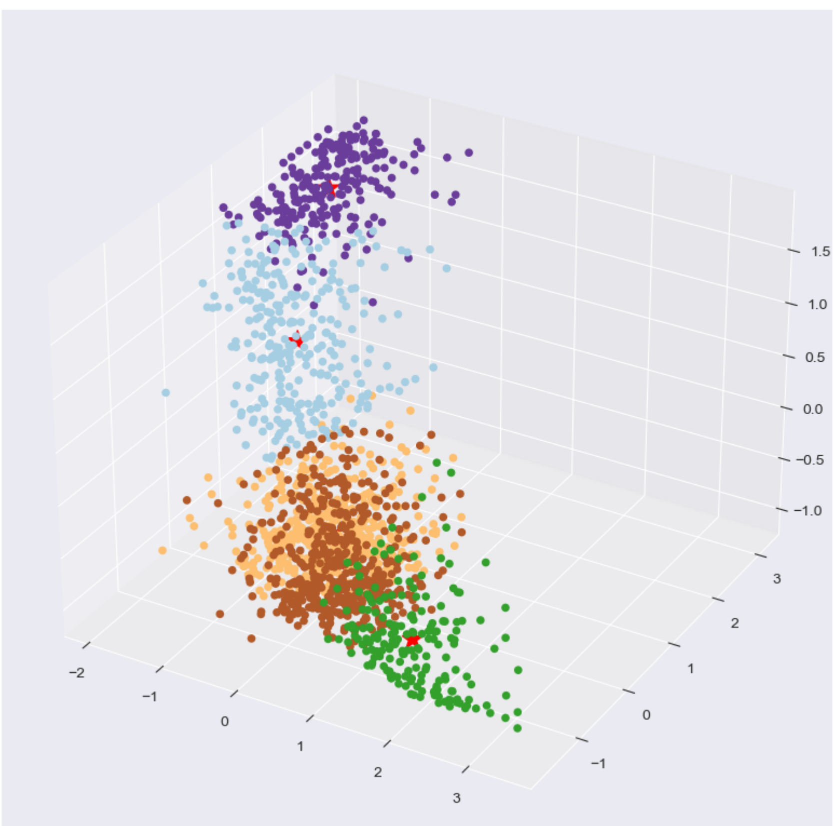


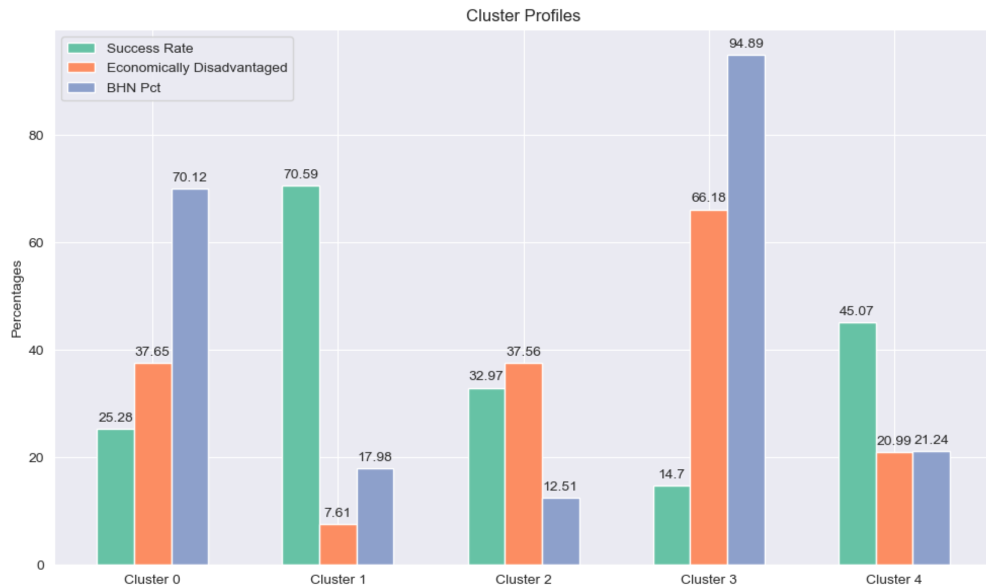
Figure 12: 3D Scatterplot

Each cluster's profile consists of three key features: overall success rate, percentage of economically disadvantaged students, and the combined percentage of Black, Hispanic, and Native American students. The profiles are as follows:

- **Cluster 0** is defined by the lowest overall success rates, a moderate level of economic disadvantage, and the highest representation of Black, Hispanic, and Native American students.
- **Cluster 1** represents schools with high success rates, the least economic disadvantage, and lower representation of Black, Hispanic, and Native American students.
- **Cluster 2** shows moderate success rates, moderate economic disadvantage, and a lower-than-average representation of Black, Hispanic, and Native American students.
- **Cluster 3** contains schools with the lowest success rates, high economic disadvantage, and the highest representation of Black, Hispanic, and Native American students.
- **Cluster 4** includes schools with above-average success rates, a below-average level of economic disadvantage, and a relatively low representation of Black, Hispanic, and Native American students.

These profiles suggest distinct patterns in school performance related to demographics and socioeconomic status. Cluster 3, in particular, may require focused attention due to its profile of high economic disadvantage and low success rates. In our analysis bar charts helped visualize the detailed percentages for the three defining features of each cluster, emphasizing the contrast between clusters in terms of success rate, economic disadvantage, and racial and ethnic representation.

Figure 13: Clustering Bar-Chart



In conclusion, the K-Means clustering provided insightful groupings of schools, unveiling patterns that could be pivotal for targeted interventions and resource allocation, aiming to support schools in need and improve educational outcomes.

Answers to Research Questions

RQ1: How do socioeconomic status, minority group representation, and English proficiency levels predict school letter grades in Tennessee?

Our analysis indicates that these demographic predictors have a significant impact on school letter grades. By quantifying the influence of each demographic factor, we aim to uncover deeper insights into how these elements contribute to educational outcomes.

RQ2: What is the relationship between access to educational resources and school letter grades?

This question explores how the availability of modern facilities and resources correlates with the letter grades schools receive, emphasizing the role of material support in achieving higher educational outcomes.

RQ3: How effectively can machine learning models predict school grades based on existing performance data and demographics?

Our models have demonstrated high accuracy in predicting school grades, indicating that machine learning could be instrumental in educational assessments and policy-making. This analysis seeks to validate the robustness of these models in a real-world educational setting.

RQ4: What patterns emerge from clustering Tennessee schools based on demographic and resource variables, and how can these insights inform educational policy?

By applying the K-Means clustering algorithm, our study identifies distinct groupings of schools sharing similar socioeconomic statuses, demographic profiles, and resource access. This question delves into the specific characteristics of each cluster, assessing how these factors correlate with school performance. The findings aim to highlight disparities and inform targeted interventions that could support equitable educational improvements.

Conclusions

Across multiple models — Random Forest, Feedforward Neural Network, XGBoost, and logistic regression — we observed a generally high level of accuracy in predicting school letter grades. Each model demonstrated distinct strengths in identifying various performance levels, with particularly impressive results from the Random Forest and XGBoost classifiers. Feature Importance: The Random Forest and XGBoost models provided insights into feature importance, revealing which factors beyond traditional academic metrics play a significant role in predicting school performance grades. Demographics, economic status, and school-level characteristics emerged as influential predictors, suggesting that a holistic approach is necessary to understand school performance.

The application of K-Means clustering uncovered natural groupings within the schools, revealing patterns that associate economic and demographic factors with school performance. Notably, clusters with high levels of economic disadvantage and higher percentages of minority students tended to have lower overall success rates. The Receiver Operating Characteristic (ROC) analysis of the logistic regression model demonstrated the model's strong discriminative ability, especially for distinguishing between the highest (Grade A) and lowest (Grade F) performing schools. The AUC values indicate that the model can reliably differentiate between schools across the spectrum of performance categories. The predictive models deployed in this analysis have demonstrated not just academic merit but profound practical utility for education administrators.

By pinpointing schools likely to underperform, these models serve as advanced warning systems, allowing educational policymakers to preemptively allocate resources, support, and interventions tailored to the unique needs of these institutions. The clustering analysis, in particular, provides a nuanced categorization of schools, revealing underlying patterns that may contribute to school performance. By examining the common characteristics of schools within each cluster, policymakers can identify successful schools that serve as benchmarks and develop replicable strategies. For instance, schools in high-achieving clusters can be studied to distill the factors contributing to their success, which could be implemented in lower-performing clusters.

These clusters could also assist in customizing improvement plans for each group of schools. For example, clusters with high levels of economic disadvantage may require interventions that address socioeconomic barriers to education, whereas clusters with high success rates might be models for curriculum development, teacher training programs, or student engagement strategies. Moreover, the insights gained from these models could inform more equitable resource distribution, ensuring that schools with the greatest need receive targeted assistance. This could take the form of financial investments, support programs for students and teachers, or infrastructure enhancements. While the models performed well, there is room for further research. For instance, exploring the impact of additional variables such as teacher qualifications, student-teacher ratios, and school programs could potentially enhance the models' predictive power. Moreover, investigating the causes of misclassifications, particularly among contiguous performance categories, could yield insights for improving models further.

Decision makers inside education can analyze these findings to develop targeted interventions for schools. For instance, schools in clusters with high economic disadvantage may benefit from programs aimed at equity in education, while also trying to replicate best practices from high-performing clusters. In conclusion, the application of machine learning techniques in educational data analysis offers a proactive approach to educational improvement. It empowers stakeholders to make data-driven decisions that are equitable, efficient, and effective, ultimately fostering an educational environment where every school has the opportunity to succeed.

Future Research

Future studies could incorporate additional demographic variables such as parental education levels, student mobility rates, and access to extracurricular activities. These factors could provide further insights into their impact on school performance and help refine predictive models. Studies such as conducting a longitudinal study to track changes in school letter grades over time would help determine trends and the long-term effects of educational policies. This could also assist in understanding how interventions and resources impact schools differently based on their demographic and socioeconomic contexts. Additionally, comparing the Tennessee school grading system with those of other states or countries could reveal insights about different educational assessment practices. This comparative analysis would help identify best practices and potentially more equitable approaches to grading schools. Another approach might be investigating the impact of teacher qualifications, such as certification levels and years of experience, on school letter grades would be valuable. Understanding how teacher proficiency affects outcomes could guide professional development and recruitment policies.

On the research methodology side, future research could explore more advanced machine learning techniques such as deep learning or ensemble models that may provide better accuracy and insights. Testing different algorithms and feature sets could optimize predictive performance. Also, incorporating qualitative research methods, such as interviews or focus groups with educators, parents, and students, could provide contextual understanding that quantitative data alone cannot. This would add depth to the findings and help tailor interventions more effectively.

From a policy analysis, analyzing the impact of specific educational policies on school letter grades would offer valuable feedback on current legislative measures. This could help policymakers refine or develop strategies that more effectively address the challenges identified through the machine learning analysis. And indeed, a focused study on how resources are allocated within schools and their direct impact on educational outcomes could help optimize budgeting and resource distribution to maximize student achievement.

Team Member Roles

- Jason Horne: Managed the project, contributed to data analysis and model validation, and wrote the drafts.
- Oluwasegun Adegboye: Focused on data collection and preprocessing, and he also oversaw the final drafts of the paper.
- Mark Constantino: Led the statistical analysis and model development, and he oversaw final drafts of the presentation and paper.
- Fuyima Inoue: Managed documentation and prepared the final presentation in addition to working on a lot of the modeling and coding.

What did we learn?

The project highlighted how external factors like socioeconomic status disproportionately affect educational outcomes. We also learned the importance of data-driven decision-making in improving school performance and policy development. For all of us, this really enhanced our skills in machine learning and data analysis, particularly in the context of educational data. We had to build pipelines and do data processing on an unfamiliar dataset, and it was a great experience for everyone. Also, working in a team environment like this was also a great learning experience with all four students having different professional, academic, and ethnic backgrounds. It was a great experience and shows the power of diversity.

References

Green, L., & Celkan, G. (2011). Student demographic characteristics and how they relate to student achievement. *Procedia - Social and Behavioral Sciences*, 341-345.

Hirn, R. G. (2018). Exploring instructional differences and school performance in high-poverty elementary schools. *Preventing School Failure: Alternative Education for Children and Youth*, 37-48.

Horne, J. B. (2024, January 12). *A preliminary look at TN school letter grades*. Retrieved from Jason Horne: https://www.jasonhorne.org/blog/2024/1/12/a-preliminary-look-at-tn-school-letter-grade-data

Horne, J. B. (2024, March 16). *Jason Horne's Github*. Retrieved from GitHub: https://github.com/jasonbhorne/tnpublicschoolstuff/blob/main/merged_data.xlsx

Lacour, M., & Tissington, L. D. (2011). The effects of poverty on academic achievement. *Educational Research and Reviews*, *6*(7), 522-527.

Rutkowski, D., Rutkowski, L., Wild, J., & Burroughs, N. (2018). Poverty and educational achievement in the US: A less-biased estimate using PISA 2012 data. *Journal of Children and Poverty*, *24*(1), 47-67.

Stake, R. E. (1991). The Teacher, Standardized Testing, and Prospects of Revolution. *The Phi Delta Kappan*, 243-247.

Tennessee Code Annotated. (2024, March 16). *Tenn. Code Ann. § 49-1-228.* Retrieved from Lexis Nexis: https://advance.lexis.com/documentpage/?pdmfid=1000516&crid=e68d98f9-1e8a-45cd-bb1c-37298c8bebce&nodeid=ABXAABAACABC&nodepath=%2FROOT%2FABX%2FABXAAB%2FABXAABAAC%2FABXAABAACABC&level=4&haschildren=&populated=false&title=49-1-228.+School+grading+system+%E2%

Tennessee Department of Education a. (2024, March 16). *School Letter Grades*. Retrieved from TN Department of Education: https://www.tn.gov/education/schoollettergrades.html

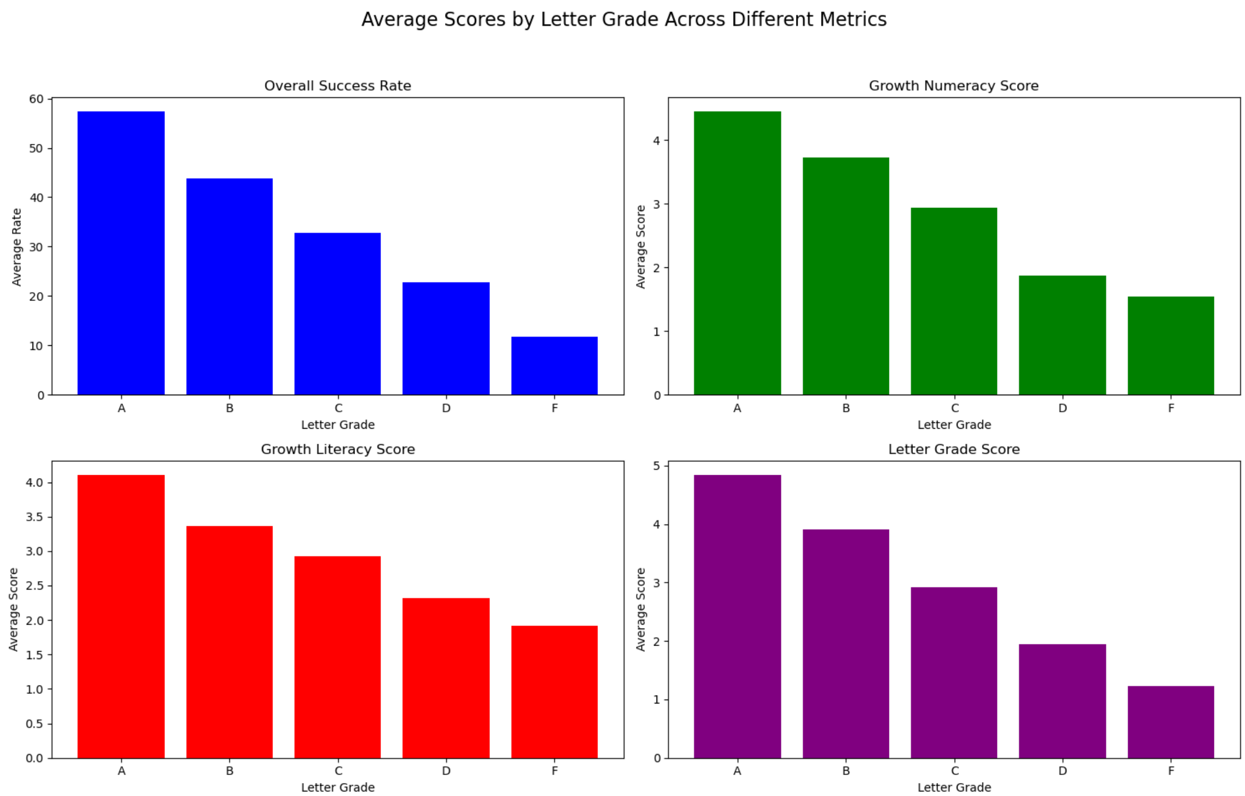
Tennessee Department of Education b. (2024, March 16). *Data Downloads and Requests.* Retrieved from TN Department of Education: https://www.tn.gov/education/districts/federal-programs-and-oversight/data/data-downloads.html

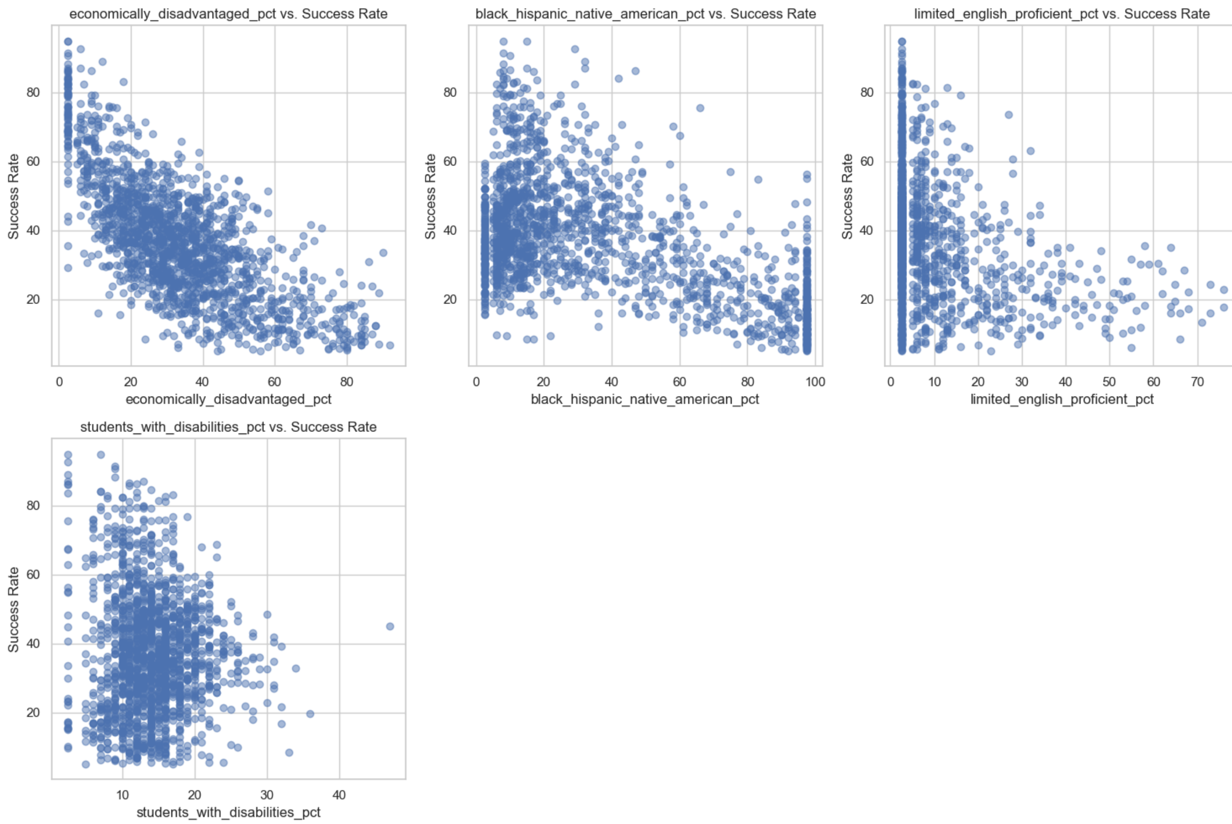
Tennessee Department of Education c. (2024, March 16). *2022-2023 A-F Letter Grade File.* Retrieved from TN Department of Education: https://www.tn.gov/content/dam/tn/education/data/2022-23_A-F_Letter_Grade_File.xlsx

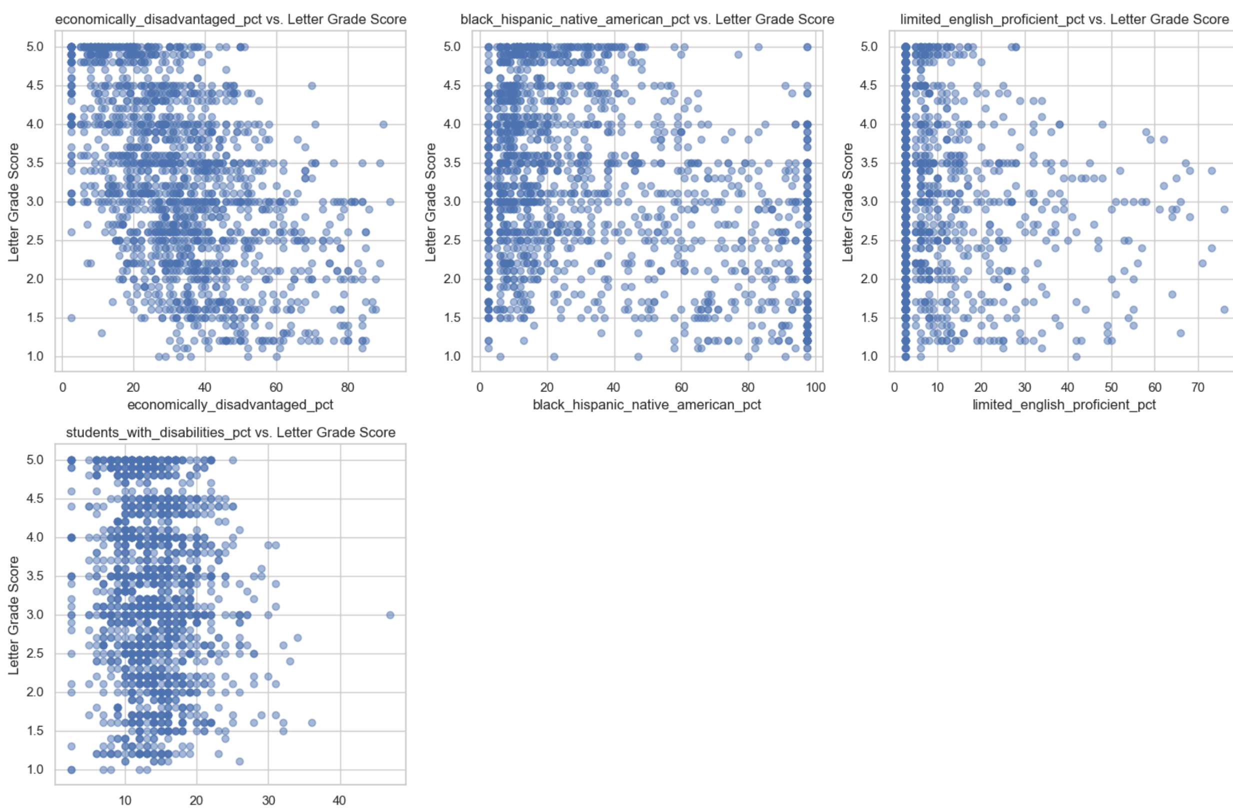
Tennessee Department of Education d. (2024, March 16). *2021-2022 Membership File.* Retrieved from TN Department of Education: https://www.tn.gov/content/dam/tn/education/data/school-profile-2021-2022-updated-2022-12-06.xlsx

West, E. R. (2023, December 21). *Letter grades for Tennessee schools have been released.* Retrieved from WCYB: Newschannel 5: https://www.newschannel5.com/news/letter-grades-for-tennessee-schools-have-been-released

Appendix


Average Scores by Letter Grade Across Different Metrics
Overall Success Rate
Average Rate
Letter Grade
Growth Numeracy Score
Average Score
Letter Grade
Growth Literacy Score
Average Score
Letter Grade
Letter Grade Score
Average Score
Letter Grade
A
B
C
D
F


economically_disadvantaged_pct vs. Success Rate
Success Rate
economically_disadvantaged_pct
black_hispanic_native_american_pct vs. Success Rate
Success Rate
black_hispanic_native_american_pct
limited_english_proficient_pct vs. Success Rate
Success Rate
limited_english_proficient_pct
students_with_disabilities_pct vs. Success Rate
Success Rate
students_with_disabilities_pct


economically_disadvantaged_pct vs. Letter Grade Score
Letter Grade Score
economically_disadvantaged_pct
black_hispanic_native_american_pct vs. Letter Grade Score
Letter Grade Score
black_hispanic_native_american_pct
limited_english_proficient_pct vs. Letter Grade Score
Letter Grade Score
limited_english_proficient_pct
students_with_disabilities_pct vs. Letter Grade Score
Letter Grade Score
students_with_disabilities_pct