ANALYSIS OF DISPARITIES BETWEEN VIRGINIA HIGH SCHOOLS

A PREPRINT

Matthew C. Haid
wsd6vn@virginia.edu

Katie E. Tsai
nmu3tr@virginia.edu

Kayla L. Nguyen
ten4vp@virginia.edu

1 Abstract

For this project, we examine the educational disparities among high schools in Virginia through the application of K-means clustering. We will employ K-means clustering to group high schools in Virginia based on various features from the Department of Education's School Quality Profiles, attempting to reveal underlying patterns and correlations within the educational landscape. Leveraging a comprehensive dataset, we are motivated by the imperative to address inequities in funding, resource allocation, and academic outcomes, which disproportionately affect students based on their school's location and socioeconomic factors. As we navigate the complex landscape of Virginia's high schools, our mission is clear: to equip policymakers, educators, and communities with a lens through which they can shape the future of education in the state.

2 Introduction

Educational disparities exist across many education systems, ranging from the school to district and regional level. On a larger scale, Virginia's education system is indicative of one where a student's access to quality education is influenced by location and other socioeconomic issues. The root of these disparities date back decades and have a variety of causes. Because of these factors, students across the state do not have equal opportunities or consistent levels of education that sufficiently prepare them for their future endeavors. In the past 20 years alone, the number of schools separated by income has raised by more than 60% (1). Recently, in July of 2023, it was found that new released formulas decreased the average funding per student by $1,900, inadequately covering student needs, classroom resources, and access to faculty such as a counselor (4). The Virginia Department of Education (VDOE) has released their own plans to best support learners and teachers post-pandemic, but there is still much to achieve before equity is reached across the education system. With this project, the goal is to minimize these disparities across Virginia high schools by attempting to group schools together based on the multitude of factors that can impact an education: spending per student, available resources, overall academic achievement, and more. By achieving this goal, schools may understand correlations across different academic and socioeconomic factors to create a more equitable plan for funding and equal opportunities for students across the state.

3 Method

This method relies on picking the right features from various datasets that directly relate to our goal of clustering schools based on disparities. These features need to be strong enough to draw meaningful insights into what distinguishes academically-excellent schools from those that face challenges. To assess if a feature fits the needs for this experiment, we examine its standard deviation across all clusters. If it is consistently high across clusters, it does not have a strong impact in assigning a school to a specific cluster. We are also exploring other techniques such as feature importance scores, mutual information, and correlation analysis to spot influential features.

For the clustering, we will be trying out a variety of techniques: specifically K-means, Gaussian Mixture Models (GMM), and DBSCAN. The final choice ultimately depends on how well they perform in terms of Cluster Validation Metrics, including separation and connectivity of clusters. Clusters will be compared using metrics such as the silhouette score, Davies-Bouldin index, and sum of squares error. This will help us choose the algorithm that best fits our data and goals.

To fine-tune our clustering models, we plan to utilize an approach that encompasses both cross-validation techniques and hyperparameter optimization. Cross-validation will enable us to assess the performance of different hyperparameter configurations, including the choice of distance metric, for each algorithm (K-means, DBSCAN, and GMM). The number of clusters (K) for K-means will be chosen based on domain knowledge and specific evaluation metrics such as the silhouette score. For DBSCAN, the number of clusters is determined inherently by the algorithm, and GMM will use other criteria, such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC), to select the number of clusters. This comprehensive analysis will help us select the most appropriate settings for each clustering method in our dataset.

This particular data analysis goes beyond clustering. In addition to examining which features may attribute to the performance of different schools, we are looking into geospatial patterns to see if certain areas are dealing with educational disparities more acutely. We are also diving into the characteristics and statistics of each cluster, visualizing their typical properties through cluster centroids. This all ties back to our motivation to understand what drives academic success or struggles and how we can better allocate resources to schools that need them.

While our approach is robust, there are still limitations in our study design, including variations in the year from which the most recent data of features are from and missing records. There is also the risk of potential biases in reported datasets used throughout our analysis, so we have been sure to be cautious and strive for a thorough and well-rounded research process.

4 Experiments

The design of our preliminary experiment represents a novel approach to addressing educational disparities among Virginia high schools. We have identified independent variables as the various educational features, such as absenteeism and teacher quality, and framed our dependent variable as the clustering of schools into distinct groups. We encountered challenges when merging features due to differing data years, data segregation by race, and missing records from many high schools. Despite these hurdles, we constructed a refined dataset that encompassed seven critical features, involving 334 schools. While our initial dataset may have contained fewer features than we aspired to include, this preliminary experiment allowed us to investigate disparities effectively.

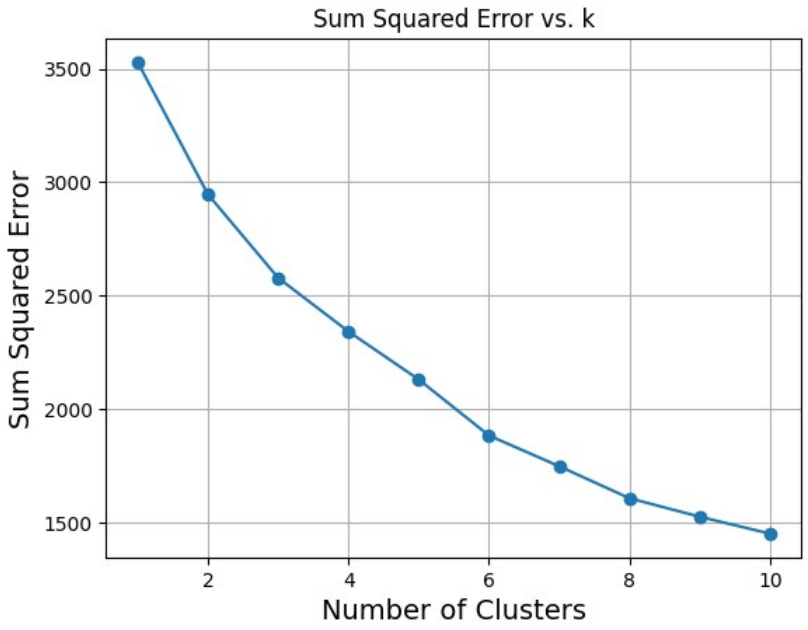

Figure 1: Number of Clusters v. Sum Squared Error.

In our first experiment, we conducted an analysis of the optimal number of clusters (K) by incrementing K and plotting the sum of squared errors (Figure 1). This technique, commonly known as the "elbow method," revealed that the ideal K for our dataset was 3. Subsequently, we examined cluster means and intra-cluster standard deviations to discern the unique characteristics of each cluster. The first cluster exhibited low standard deviation in "Number of Students Enrolled for Half the Year or More" and a low mean, signifying high enrollment rates throughout the year. The second

cluster displayed a low standard deviation in "Percent of Out-of-Field Teachers" and a mean close to 0, indicating a "middle-of-the-pack" performance compared to other schools. Lastly, the third cluster demonstrated low standard deviations in "High Poverty" and "Percent of Inexperienced Teachers" while exhibiting high means for both variables. This provided invaluable insights into what set each cluster apart, thereby offering a novel perspective on the educational landscape in Virginia.

In our progression, we recognized the need to enhance our dataset to ensure a comprehensive analysis of educational disparities among Virginia high schools. The decision was prompted by the realization that our initial dataset lacked sufficient features for precise school categorization. Our aim was to identify additional characteristics that would contribute to a more nuanced understanding of what constitutes academic excellence in high schools. To address this, we delved into more datasets, consolidating information into a comprehensive dataset that expanded upon the original. Despite encountering challenges such as the prevalence of identical high school names in Virginia, causing complications in data processing, we managed to compile a dataset featuring 259 high schools.

While the number of schools fell short of our initial target, we successfully augmented the dataset with 7 additional features. This substantial increase in features significantly enhanced the accuracy of our experiment and facilitated a more in-depth analysis of the clusters. The culmination of these efforts led to the creation of a finalized dataset, which served as the foundation for our exploration of different clustering algorithms.

The subsequent step in the experiments involved selecting and applying clustering models to categorize high schools. Following an exploration of various clustering methods, we identified four primary categories: centroid-based clustering, density-based clustering, distribution-based clustering, and hierarchical clustering. Given the mismatch between our dataset and hierarchical structures, we chose not to pursue models from the hierarchical category.

Distribution-based clustering operates on the assumption that data follows distributions like Gaussian distributions. To assess this, we opted to test the Gaussian Mixture model, considering it a suitable candidate for comparison with other selected models.

Density-based clustering identifies clusters by connecting regions of high-density areas. Despite challenges associated with varying data densities and high dimensions, we decided to test the DBSCAN model, acknowledging potential performance differences compared to other chosen algorithms.

Centroid-based clustering, known for organizing data into non-hierarchical clusters, is efficient yet sensitive to initial conditions and outliers. For this category, we selected the K-means model based on our preliminary experiments. Thus, we proceeded with three new models - K-means, DBSCAN, and Gaussian Mixture — and executed them on our dataset, examining centroid means and intra-standard deviations to extract valuable insights into the clustering patterns among Virginia high schools.

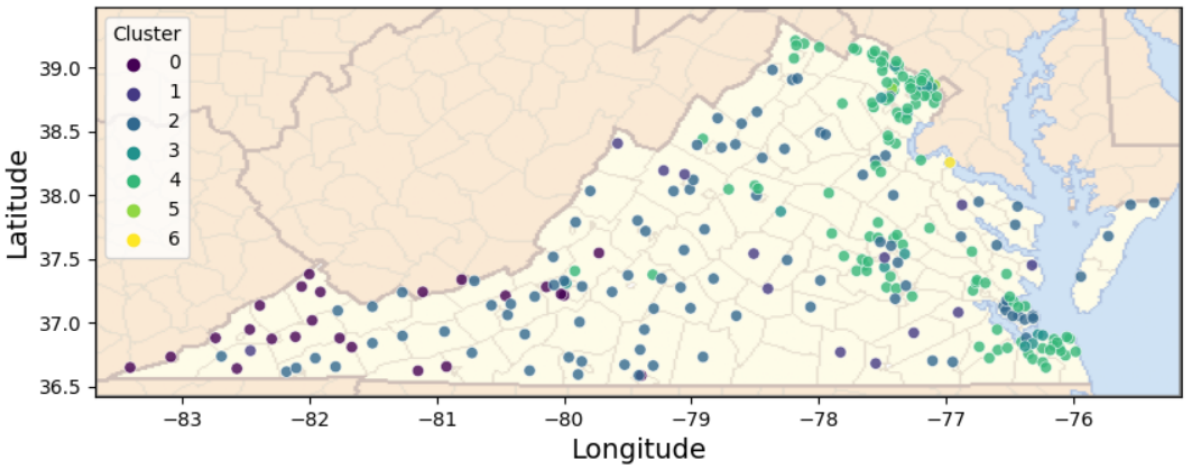

Figure 2: Geographical Locations of GMM Cluster Labels.

In our Gaussian Mixture Model (GMM) experiments, we searched for the optimal number of components by constructing a Log-Likelihood vs. K graph (Figure 3). Executing the GMM model multiple times on our data with varying numbers of components allowed us to accumulate log-likelihood scores in an array. The subsequent visualization of this curve guided our understanding of the model's behavior. Simultaneously, we sought the best Bayesian Information Criterion (BIC) to determine the ideal number of clusters. Repeated runs of the model, comparing the BIC values, pinpointed the

iteration with the lowest BIC, defining our total number of components needed to run this model. Despite our initial attempt to map clusters onto the Virginia map, we found the results unsatisfactory, particularly with some clusters containing only a handful of schools (Figure 2). These clusters closely resembled the outcomes of the original K-means experiment, leading us to abstain from further exploration down this path.

In our exploration of the DBSCAN algorithm, precision required fine-tuning two key hyperparameters: Minimum Samples and Epsilon. Epsilon, denoting the neighborhood radius, signifies the maximum distance between two data points for one to be considered in the neighborhood of another. We set Minimum Samples to 24, aligning with the general recommendation of 2 times the number of features. Determining the appropriate Epsilon involved employing sklearn's "NearestNeighbors" on our dataset to ascertain the neighbor distance for each point (Figure 4). Sorting this data enabled the creation of a graph plotting K-Nearest Neighbor Distance on the y-axis against the Sorted Observations of the 24th nearest neighbor. Our analysis led us to identify a critical point, around 250, where the graph exhibited an exponential increase. However, implementing this high Epsilon resulted in a single cluster upon prediction. As we decreased Epsilon, the number of outliers surged until all data points were classified as outliers. Recognizing the sparse nature of our data, incompatible with the dense data requirements of DBSCAN, we opted not to pursue further experimentation in this domain.

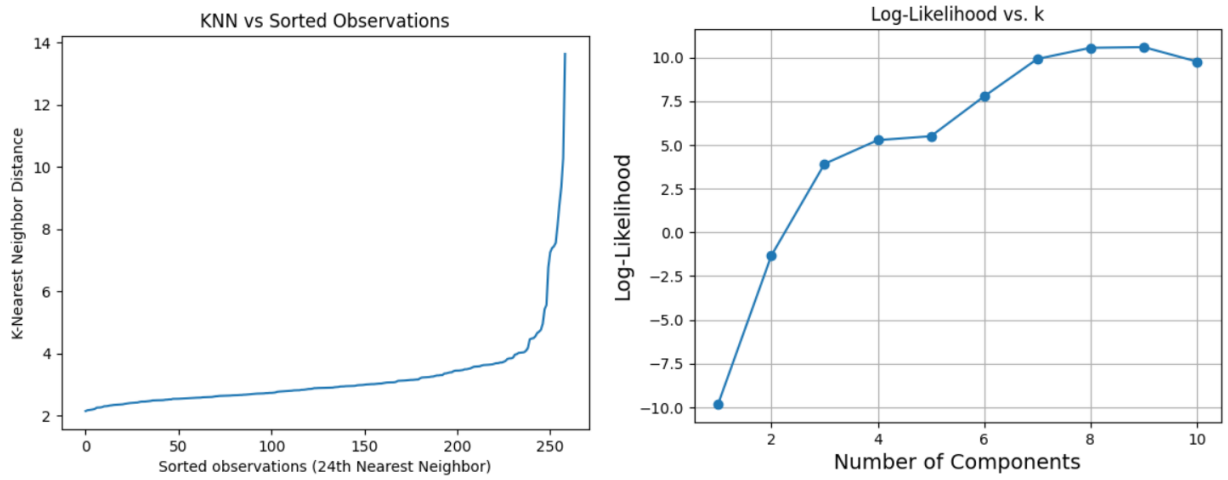

Figure 3: Log-Likelihood v. K. and Sorted Observations (Nearest Neighbor) v. Nearest Distance.

Our experiences with K-means closely paralleled our preliminary experiments on the original dataset. The number of clusters and geographical distributions remained largely consistent, with a few high schools switching places (Figure 5). Detailed results will be expounded upon in the results section, but for now, it suffices to note that our decision to proceed with K-means brought us closer to finalizing our results.

5 Results

When running the K-Means algorithm with 3 clusters, we were able to gain valuable insights into the learning landscape that exists within Virginia, as when visualizing our results, the clusters showed a clear depiction of the disparities that exist within Virginia. As seen in Figure 4, Cluster 1 was mainly in the Northern Virginia, Richmond, and Norfolk areas. When analyzing notable features, the mean for this cluster in "Number of Students Missing 10% or More of the Day Enrolled" was the lowest compared to the other clusters at 0.38. Additionally, it had one of the lower "Student-Teacher Ratio" mean at -0.31, the lowest "Dropout Rate" of 0.06, and the highest mean for "Virginia On-Time Graduation Rate" at 0.57. Across all the features, there was a low standard deviation, meaning that there was not much variance that existed between the schools in this cluster. Based on the region these clusters were centralized around and respective data, it can be concluded that a cluster appeared in these areas due to them being large, prominent cities, with a large population. It can be further concluded when comparing their data against the other clusters, that the schools in this cluster receive ample amount of resources for their students to succeed, which aligns with the cluster being in large, populated cities.

Cluster 2 was mainly in the Southern and Western regions of Virginia. The mean for this cluster was low across most of the features which is a positive sign for features like "Chronic Absenteeism Rate" and "Number of Students Missing 10% or More of the Days Enrolled," but is a negative sign when looking at features like "Virginia On-Time Graduation

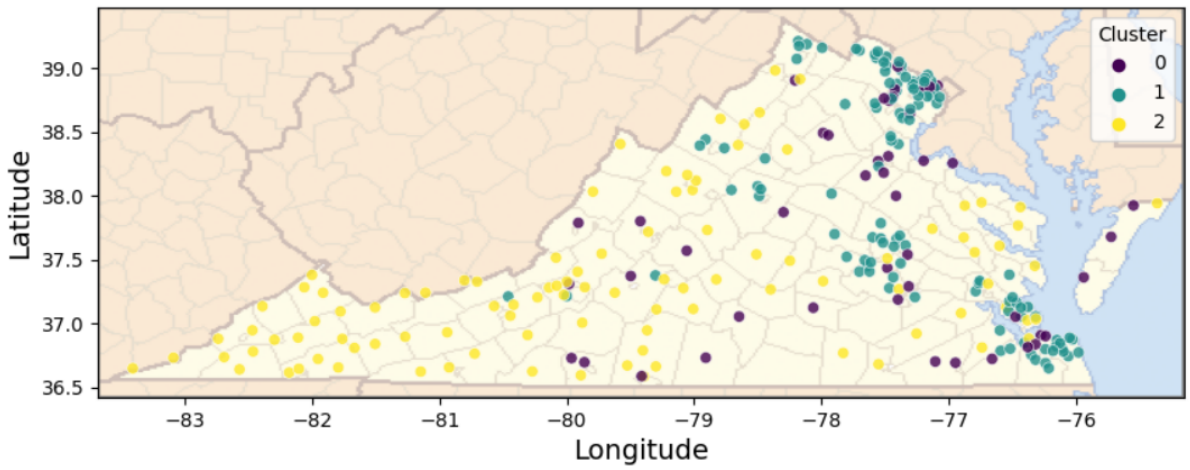

Figure 4: K-Means on High Schools in Virginia.

Rate" with a low mean of 0.07. It also had a relatively high "Dropout Rate" mean of 0.31. It had varying standard deviations, meaning that these features varied from school to school, with the highest one being 1.30 for "Percent of Inexperienced Teachers". The data for this cluster alongside geographical patterns unveils a narrative of students' high efforts within these regions, evident from the mean values of "Chronic Absenteeism Rate" and "Number of Students Missing 10% or More of the Days Enrolled." However, despite these efforts, the relatively low graduation rates suggest a potential scarcity of resources or deficiencies within the educational landscape.

Cluster 0 was spread out across the state. It had a high mean in notable features such as "Number of Students Missing 10% or More of the Days Enrolled" at 0.70, "Chronic Absenteeism Rate" at 1.12, "Student-Teacher Ratio" at 1.06, and "Poverty Level-High" at 1.08. In contrast, it had a low mean in "Number of Students Enrolled for Half the Year or More" at 0.02, "Title I" at 0.34, and "Virginia On-Time Graduation Rate" at 0.04. Its high standard deviation across all features aligns logically with the extensive geographical distribution that exists within this cluster. Based on the data, these schools reflect a poor performance level and suggest a need for targeted support to address the various challenges faced by both students and educators.

6 Conclusion

Education plays a pivotal role in shaping individuals and societies and is a huge priority for many residents. For students to reach their full potential and gain the education they deserve, there must be resources allocated equally to schools across the nation. We anticipated for disparities to exist within the educational environment in Virginia, and based on our findings, it is evident that these inequalities are real, and there must be change brought about. With our findings, residents of Virginia are able to easily understand where the disparities exist between the High Schools in Virginia. More specifically, policymakers, educators and communities must strive to ensure that all schools in Virginia are similar to the schools within Cluster 1. Access to resources should not be solely defined based on the geographical location or the popularity of a city; instead, there should be an equitable distribution of resources to create a fairer educational environment in Virginia. However, with such diversity, it is also crucial to tailor support to address specific regional needs while ensuring fairness and parity across the board. With a fairer educational landscape, it will ensure that every student has an equal opportunity to succeed, and their limits will not be defined by a lack of resources. In our findings, we strictly analyzed public high schools in Virginia, and many high schools were dropped due to disparities in the data. This shortcoming led to less comprehensive findings than we initially intended; however, with clustering, we were still able to identify trends among the data. For future work, we would widen our dataset to obtain a broader conclusion and devise a more general resource allocation plan for schools across the nation, not just Virginia. More specifically, we would analyze more features to determine where else disparities lie and broaden our data to more schools across the nation. We could also broaden our data collection to past years to see if there has been any improvement over the years or if our strides towards equality remain stagnant. By ensuring fair access to education, we pave the way for a brighter, more inclusive future where every student has the chance to thrive and contribute to a better society.

7 Contributions

Kayla and Katie worked on retrieving the data and pre-processing it to make it ready to use for the clustering algorithm. This included dropping the columns that were not of use, merging many different CSV files into a usable object, and transforming the data using a pipeline to ensure optimized performance. Furthermore, Matthew and Katie worked on implementing specific clustering algorithms and acquiring data to analyze for results. Kayla worked on editing the ML4VA video supplement, and all members contributed to completing to supporting scripts and the written report, including preparing figures and references.

References

[1] Commonwealth Institute for Fiscal Analysis. Increasingly separate and unequal in U.S. and Virginia schools. The Commonwealth Institute. https://thecommonwealthinstitute.org/research/increasingly-separate-and-unequal-in-u-s-and-virginia-schools/, 2021.

[2] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014.

[3] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

[4] Karina Elwood. Virginia underfunds K-12 education with flawed formula, report finds. Washington Post. https://www.washingtonpost.com/education/2023/07/16/virginia-jlarc-education-funding-report/, 2023.