ML4VA: PREDICTING STUDENT SUCCESS RATE OF VIRGINIA PUBLIC SCHOOLS ($Category : Education$)

Anjali Pagidi
University of Virginia
mdp3ka

Pravallika Kullampalle
University of Virginia
qkm3zr

Ananyashri Sai
University of Virginia
xqr8dj

May 9, 2025

1 Abstract

In order to identify the factors most strongly linked to student achievement as measured by SOL (Standards of Learning) pass rates, we looked at data from Virginia Public Schools. Using a Kaggle dataset of 19 interconnected tables, we combined information on school-level behaviors, teacher qualifications, funding, economic factors, and student demographics. We developed a comprehensive Python data pipeline that includes preprocessing, feature engineering, and model creation using scikit-learn. After training and evaluating three different regression models—random forest, decision tree, and linear regression—the random forest model had the lowest RMSE. We used K-means clustering to investigate natural groupings among schools, and we employed the elbow approach number of clusters which came out to k=10. We then used the correlation coefficient and the centroid of each cluster to compute a weighted sum by multiplying the correlation coefficient with the corresponding value within each cluster. We then ordered the clusters with the highest weighted sum cluster receiving a rating of 10 and the lowest weighted sum cluster receiving a score of 1. A classification algorithm was used to predict a school's rating given new data. Lastly, an Anvil Front end was developed to enter data for a new school and view centroid data for each of the 10 ratings. In addition to highlighting important factors that contribute to academic performance, our work attempts to establish the framework for an understandable school rating system that parents, educators, and legislators may find helpful.

2 Motivation

As former students of Virginia Public Schools, we want to understand what factors contribute to student success. Our goal is to find patterns in the data that show which aspects of a school experience matter most in predicting student achievement. We are also interested in how schools can improve these factors to help students do better. Since many parents choose where to live based on school ratings, we plan to create a rating system that reflects our findings.

3 Dataset

URL: `https://www.kaggle.com/datasets/zsetash/virginia-public-schools/data`

The Virginia Public Schools dataset, which has data collected from the 2021-2022 school year, has information about public schools across Virginia. The dataset consists of multiple sub-datasets, each focusing on specific topics such as SOL scores, qualifications of the educators, or attendance. Together, they provide comprehensive information on Virginia's public schools, including district names, school names, addresses, enrollment figures, and key performance indicators. It was found that one of the variables that affected the student achievement was teacher quality, which is also one of the features used in our study of Virginia public schools.

4 Related Work

A previous study published by the peer reviewed journal Education Resources Information Center (ERIC) examined secondary student achievement in large and small high schools in Virginia ([1]). This research study looked into how various predictor variables such as Virginia Standards of Learning assessments (SOLs), socioeconomic status, student attendance, minority population, and teacher quality contributed to student success of eleventh graders across Virginia public schools. This study used multiple regression analysis as its main measure of each predictor variables and the student achievement. The model controlled individual features at a time to measure how each variable affected student success. The study found that teacher quality, one of the features we will be analyzing as well, was one of the variables strongly correlated with student success.

In regards to using machine learning to study and predict student success, a prior study used KNN as well as SVM to predict student success based on features such as student environment or absences ([2]). Their model was able to make predictions with an accuracy of approximately 87%.

5 Methods

Data was obtained from the Virginia Public Schools dataset on Kaggle, made of 19 tables covering demographics, economic factors, student behaviors, teacher information, and SOL (Standards of Learning) testing data. SOL pass rates were used as the measure of student success, while features such as chronic absenteeism rates, free and reduced lunch eligibility, state and federal funding, teacher quality and education, and demographic information were analyzed.

During the pre-processing phase, data from the various tables was consolidated using Google Sheets. A function was applied to match tables by school district and school name, enabling the columns to be efficiently merged into a master dataset. This consolidated file was then imported into Google Colab and the data was analyzed using the scikit-learn library of python. The code and data can be found here: https://drive.google.com/drive/folders/1MpY9pL2OK1RC7bT5WaAc0jWEXWnEwrgW?usp=drive_link.

The data was split into 20% training data and 80% testing data using a random state set to 42. A correlation matrix was calculated to get an initial understanding of the data. Additionally, correlations between the features and SOL pass rate were graphed to observe a visual representation of the data. Additionally, plots were generated on top of a map of Virginia to map schools based on their longitude and latitude, with the size and color of each points represented SOL pass rate.

A data processing pipeline was implemented, and additional features were engineered. To prepare the dataset for analysis, we first cleaned and integrated 19 tables, resulting in a single structured dataset with 1,712 schools and 31 important attributes. Beyond standard preprocessing, we generated new variables to capture more complex interactions. These included percentages of teacher qualifications (e.g., percentage of master's or PhD holders), student-teacher ratios, funding distribution ratios (comparing federal and state spending), and gender and absence ratios. These additional characteristics were designed to better reflect school surroundings and potential determinants of performance. We put in place a strong data processing pipeline that scaled numerical variables and one-hot encoded categorical variables to facilitate future deployment and replicability. This pipeline expedited the modeling process for all tasks related to classification and regression.

The first step was to predict SOL pass rates using various features in order to validate the existence of a relationship between SOL pass rate and the other variables in the data. Initial models including linear regression, a decision tree regressor, and a random forest regressor were implemented using the LinearRegression, DecisionTreeRegressor, and RandomForestRegressor library of scikit-learn respectively. 10-fold cross validation was used for all of these models. Hyperparameter tuning was then performed on the random forest model using the GridSearchCV library of scikit-learn. The root mean squared error, mean squared error, and mean absolute error libraries were used to calculate metrics to measure the models performance.

The next step was to then use K-means clustering to find natural clusters within the data. To implement this, we developed a separate data pre-processing pipeline to handle the entire dataset (without dropping any columns). To find the best value of k, we tested values from 2-19 and created an 'elbow' plot.

Based on this, a value of k=10 was selected and k-means was performed using the KMeans library of scikit-learn. From these natural clusters observed in the data, we then developed a school rating system, assigning a rating from 1-10 (one value per each of the 10 clusters). In order to assign this rating, the average value of each feature was calculated for each cluster, and then a weighted total was found, with the weights corresponding to the correlation between each feature and the SOL pass rate. Each cluster was given a meaningful score as a result, and we ranked them to give them values ranging from 1 to 10. In order to reflect each school's rating according to the cluster to which it was given, a new "Rating" column was lastly added to the dataset. Using this "Rating" column, we conducted an additional analysis of trends of different variables across rating groups.

After this, we created and tested classification models to be used to predict the rating of a school based on its features. Specifically, we implemented Logistic Regression, Decision Tree, SVM Classifiers, and Gradient Boosting.

Finally, we used the Anvil framework to create a web application that would allow instructors and users to access it in real time. Using the anvil.server module, this application establishes a direct connection to our Google Colab backend, enabling users to enter school parameters and obtain immediate rating predictions. In addition, the interface allows users to view the average statistics for any of the rating groups. Without requiring technological know-how or direct model interaction, this integration enables school leaders to get feedback on student achievement. It is important to note that the current implementation of the user interface is a simple working prototype that does not check how user input is entered (no input validation), and therefore may result in errors if nuances in data entry are not considered. A demonstration of how data can be entered may be found in our project demo, which can be found here: https://youtu.be/wMsCJockrus?si=HXu1yFw5D2egF8-p.

6 Experiments

We began by understanding basic relationships between the variables in our dataset. We created a linear correlation matrix to find the variables with the highest linear correlation with the response variable, the SOL pass rate. In Table 1 below, we list the variables we believed could have more significant impacts on our model based on the correlation coefficient found.

Table 1: Correlation matrix of highly correlated numerical variables with SOL pass rate.

Variable	Correlation
teacher_MA	0.200615
free_reduced_lunch_eligible	-0.6889290
percent_econ_disadvantaged	-0.667541
rate_chronic_absence	-0.524526

We also plotted scatter matrices of each numerical variable against the SOL pass rate. Next, we plotted each school's latitude, longitude, and their SOL pass rate, as pictured below in Figure 1.

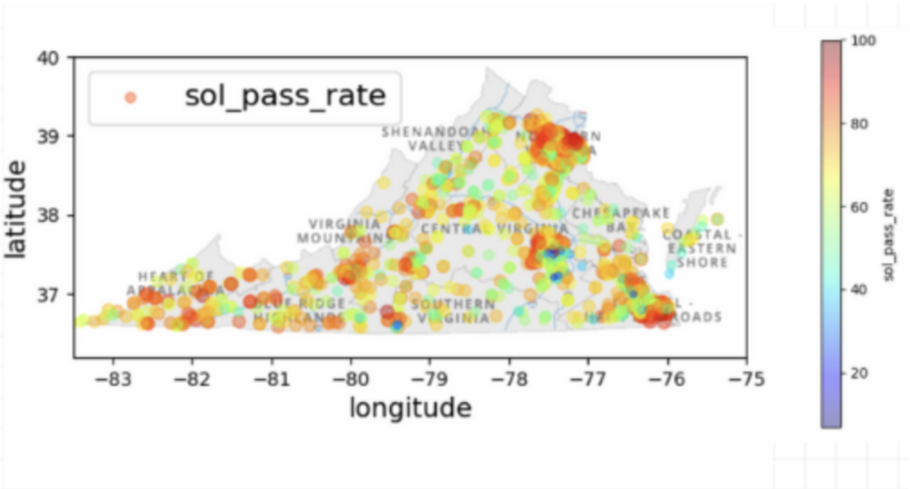


Figure 1: Map of latitude, longitude, and SOL pass rate of schools in Virginia

Finally, the following are the results of the baseline models we created. The root mean square error (RMSE) for the linear regression, decision tree, and random forest models was 9.388, 12.4677, and 8.28 (in %), respectively.This outcome showed how well ensemble models captured intricate, non-linear interactions in the educational data.

The following elbow plot was used to decide that 10 clusters needed to be used for the development of the rating system.

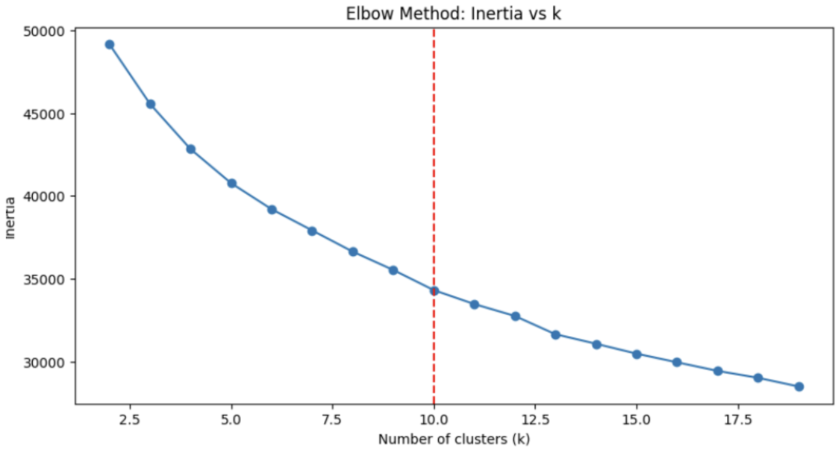
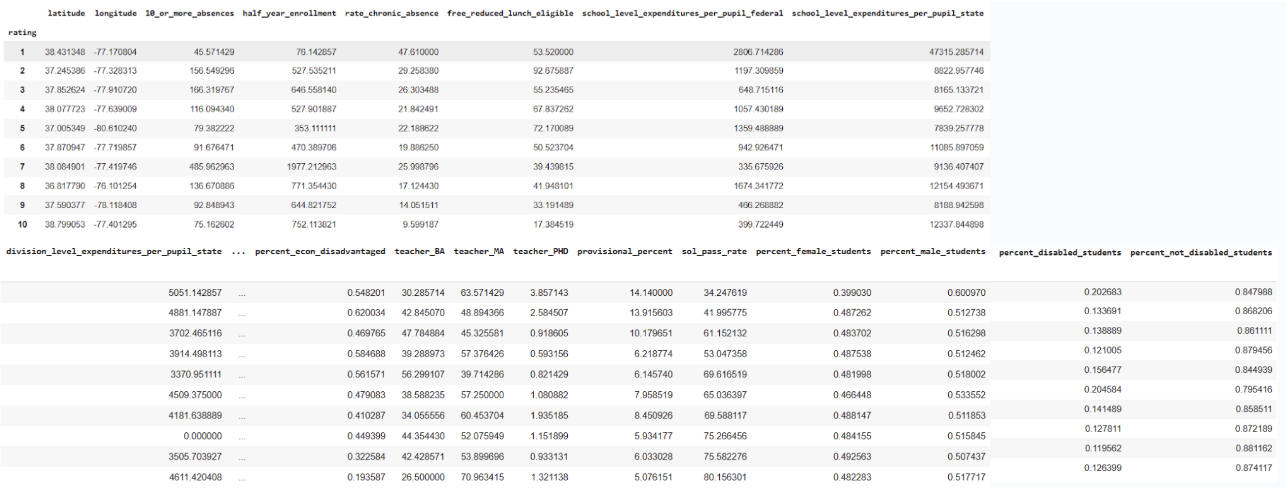


Figure 2: Elbow Plot of k-values 2 to 19

The following were the values of each of the centroids for each of the 10 clusters created:



rating	latitude	longitude	10_or_more_absences	half_year_enrollment	rate_chronic_absence	free_reduced_lunch_eligible	school_level_expenditures_per_pupil_federal	school_level_expenditures_per_pupil_state
1	38.431348	-77.170804	45.571429	76.142857	47.610000	53.520000	2806.714286	47315.285714
2	37.245386	-77.328313	156.549296	527.535211	29.258380	92.675887	1197.309859	8822.957746
3	37.852624	-77.910720	166.319767	646.558140	26.303488	55.235465	648.715116	8165.133721
4	38.077723	-77.639009	116.094340	527.901887	21.842491	67.837262	1057.430189	9652.728302
5	37.005349	-80.610240	79.382222	353.111111	22.188622	72.170089	1359.488889	7839.257778
6	37.870947	-77.719857	91.676471	470.389706	19.886250	50.523704	942.926471	11085.897059
7	38.084901	-77.419746	485.962963	1977.212963	25.998796	39.439815	335.675926	9136.407407
8	36.817790	-76.101254	136.670886	771.354430	17.124430	41.948101	1674.341772	12154.493671
9	37.590377	-78.118408	92.848943	644.821752	14.051511	33.191489	466.268882	8188.942598
10	38.799053	-77.401295	75.162602	752.113821	9.599187	17.384519	399.722449	12337.844898

division_level_expenditures_per_pupil_state	...	percent_econ_disadvantaged	teacher_BA	teacher_MA	teacher_PHD	provisional_percent	sol_pass_rate	percent_female_students	percent_male_students	percent_disabled_students	percent_not_disabled_students
5051.142857	...	0.548201	30.285714	63.571429	3.857143	14.140000	34.247619	0.399030	0.600970	0.202683	0.847988
4881.147887	...	0.620034	42.845070	48.894366	2.584507	13.915603	41.995775	0.487262	0.512738	0.133691	0.868206
3702.465116	...	0.469765	47.784884	45.325581	0.918605	10.179651	61.152132	0.483702	0.516298	0.138889	0.861111
3914.498113	...	0.584688	39.288973	57.376426	0.593156	6.218774	53.047358	0.487538	0.512462	0.121005	0.879456
3370.951111	...	0.561571	56.299107	39.714286	0.821429	6.145740	69.616519	0.481998	0.518002	0.156477	0.844939
4509.375000	...	0.479083	38.588235	57.250000	1.080882	7.958519	65.036397	0.466448	0.533552	0.204584	0.795416
4181.638889	...	0.410287	34.055556	60.453704	1.935185	8.450926	69.588117	0.488147	0.511853	0.141489	0.858511
0.000000	...	0.449399	44.354430	52.075949	1.151899	5.934177	75.266456	0.484155	0.515845	0.127811	0.872189
3505.703927	...	0.322584	42.428571	53.899696	0.933131	6.033028	75.582276	0.492563	0.507437	0.119562	0.881162
4611.420408	...	0.193587	26.500000	70.963415	1.321138	5.076151	80.156301	0.482283	0.517717	0.126399	0.874117

Figure 3: Centroid values for some variables for clusters 1-10

7 Results

With an RMSE of roughly 8.28, Random Forest Regression outperformed other models in our regression models, demonstrating good predictive accuracy for projecting SOL pass rates based on school-level characteristics. The Linear Regression Model had an RMSE of 9.523. The Decision Tree Regressor had an RMSE of 12.283. The percentage of instructors with master's degrees (positive association), the rate of chronic absenteeism (negative correlation), and the state funding ratio (positive correlation) were important prognostic factors.

According to our cluster-based grading methodology, schools with better ratings were more likely to have graduate-qualified teachers, lower chronic absent rates, and more robust state-level funding per student. On the other hand, absenteeism and provisional teaching percentages were greater in lower-rated clusters. Scatter graphs comparing rating levels to specific parameters like teacherMA, ratechronicabsence, and statefundingratio were also used to illustrate these tendencies.

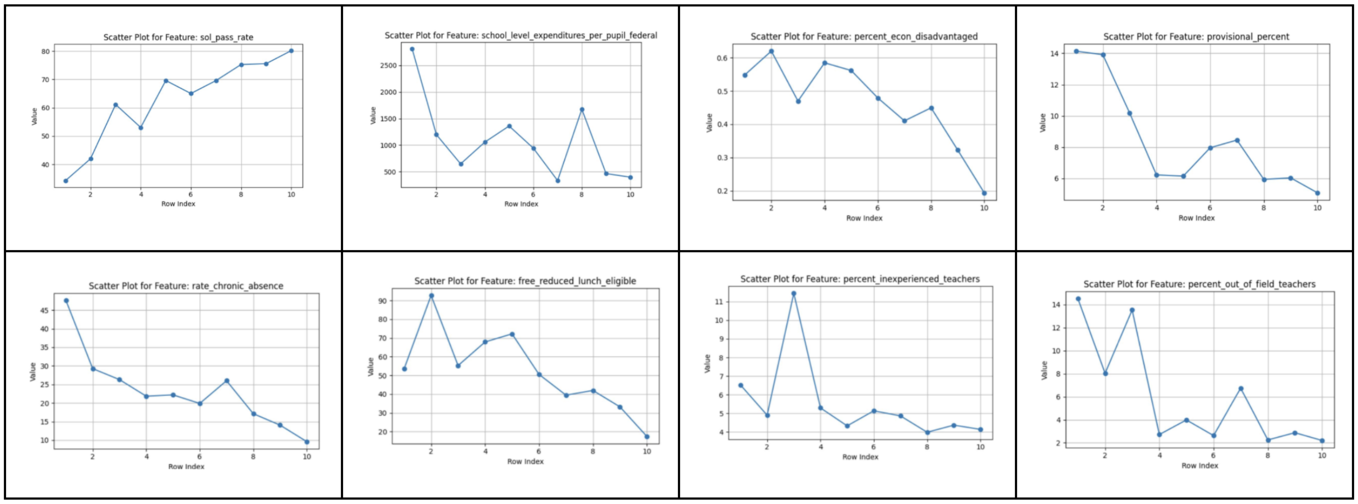
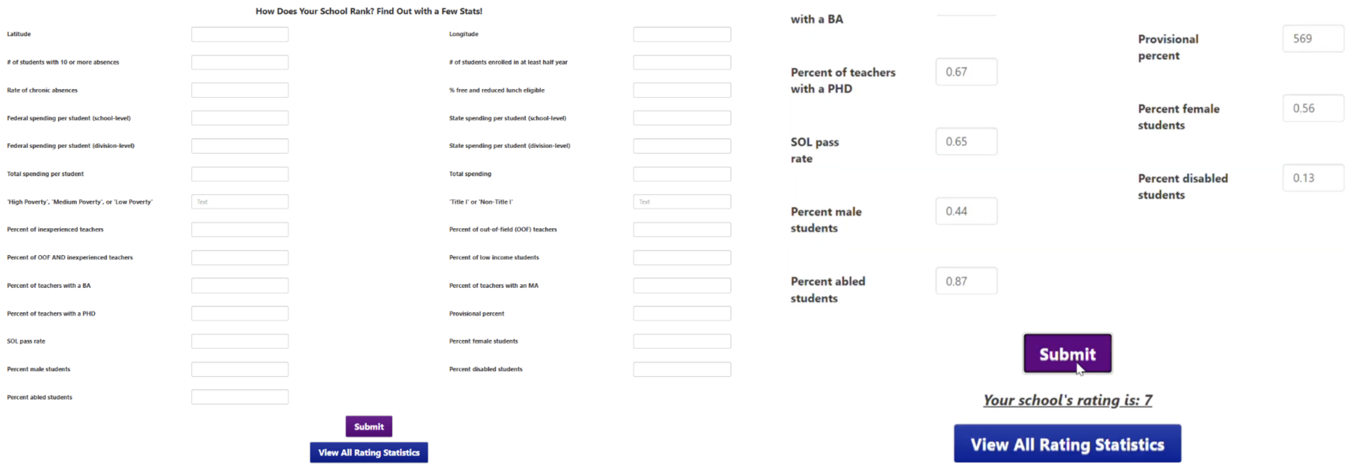


Figure 4: Trend graphs of strongly correlated variables based on rating

Strong performance was demonstrated by the categorization models used to forecast ratings. Specifically, we implemented Logistic Regression, Decision Tree, SVM Classifiers, and Gradient Boosting. Additionally, fine tuning was performed on these models, with the accuracy of each of the models being 0.95, 0.72, 0.92, and 0.87 respectively. The classification algorithm that was chosen as a prediction model to be used was the Logistic Regression Model.

Lastly, school administrators can now see how their school stacks up against other schools in the same rating category in addition to their anticipated school rating through the Anvil application. This gives teachers useful information. For instance, administrators can pinpoint specific areas for improvement if a school in a cluster with a lower rating has a much higher chronic absenteeism rate and receives less state funding.



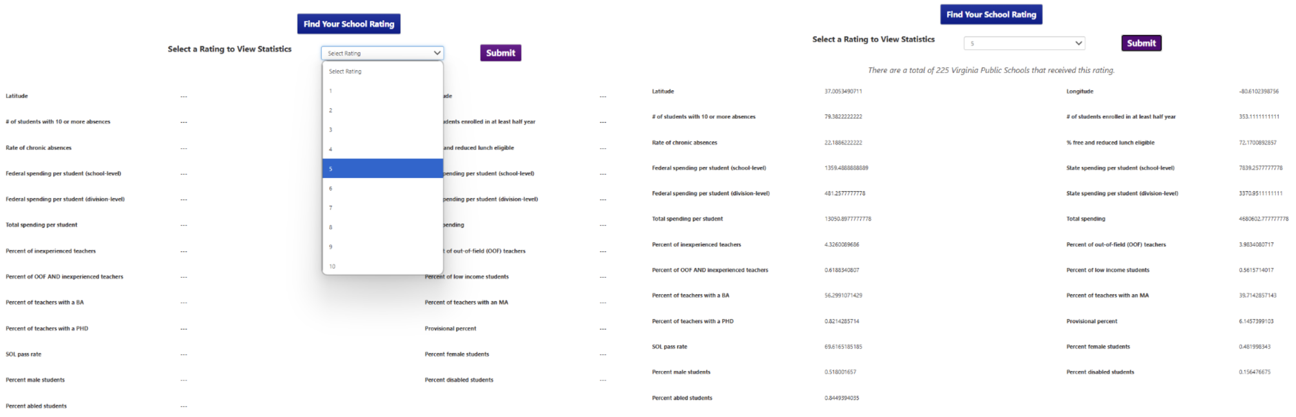
How Does Your School Rank? Find Out with a Few Stats!

Latitude		Longitude	
# of students with 10 or more absences		# of students enrolled in at least half year	
Rate of chronic absences		% free and reduced lunch eligible	
Federal spending per student (school-level)		State spending per student (school-level)	
Federal spending per student (division-level)		State spending per student (division-level)	
Total spending per student		Total spending	
'High Poverty', 'Medium Poverty', or 'Low Poverty'	Text	'Title I' or 'Non-Title I'	Text
Percent of inexperienced teachers		Percent of out-of-field (OOF) teachers	
Percent of OOF AND inexperienced teachers		Percent of low income students	
Percent of teachers with a BA		Percent of teachers with an MA	
Percent of teachers with a PHD		Provisional percent	
SOL pass rate		Percent female students	
Percent male students		Percent disabled students	
Percent abled students			

with a BA		Provisional percent	569
Percent of teachers with a PHD	0.67	Percent female students	0.56
SOL pass rate	0.65	Percent disabled students	0.13
Percent male students	0.44		
Percent abled students	0.87		

Your school's rating is: 7

Figure 5: Find Your School Rating Form



Latitude	---	ude	---
# of students with 10 or more absences	---	udents enrolled in at least half year	---
Rate of chronic absences	---	and reduced lunch eligible	---
Federal spending per student (school-level)	---	pending per student (school-level)	---
Federal spending per student (division-level)	---	pending per student (division-level)	---
Total spending per student	---	pending	---
Percent of inexperienced teachers	---	t of out-of-field (OOF) teachers	---
Percent of OOF AND inexperienced teachers	---	Percent of low income students	---
Percent of teachers with a BA	---	Percent of teachers with an MA	---
Percent of teachers with a PHD	---	Provisional percent	---
SOL pass rate	---	Percent female students	---
Percent male students	---	Percent disabled students	---
Percent abled students	---		

There are a total of 225 Virginia Public Schools that received this rating.

Latitude	37.0053490711	Longitude	-80.6102398756
# of students with 10 or more absences	79.3822222222	# of students enrolled in at least half year	353.1111111111
Rate of chronic absences	22.1886222222	% free and reduced lunch eligible	72.1700892857
Federal spending per student (school-level)	1359.4888888889	State spending per student (school-level)	7839.2577777778
Federal spending per student (division-level)	481.2577777778	State spending per student (division-level)	3370.9511111111
Total spending per student	13050.8977777778	Total spending	4680602.777777778
Percent of inexperienced teachers	4.3260089686	Percent of out-of-field (OOF) teachers	3.9834080717
Percent of OOF AND inexperienced teachers	0.6188340807	Percent of low income students	0.5615714017
Percent of teachers with a BA	56.2991071429	Percent of teachers with an MA	39.7142857143
Percent of teachers with a PHD	0.8214285714	Provisional percent	6.1457399103
SOL pass rate	69.6165185185	Percent female students	0.481998343
Percent male students	0.518001657	Percent disabled students	0.156476675
Percent abled students	0.8449394035		

Figure 6: Find average feature value for any rating

8 Conclusion

The results of our project reveal how school administrators and users can better understand their school's standing compared to other schools in the Virginia Public Schools system. Our initial results revealed a relationship between SOL pass rates and other features. The low RMSE values for our regression models predicting SOL pass rate are indicators of a relationship between a school's characteristics and that school's student success rate. This allowed us to create a reliable (proven by the high accuracy rates of the rating classifiers we created) rating system based on all of these characteristics (weighing higher correlated ones more in the rating). The implications of this project are positive, as they contribute to the betterment and well-being of Virginia public schools. Ultimately, the model not only predicts a school's rating, but it also provides actionable feedback by showing how a school compares to others in its rating group. Educators can use this to identify areas for potential improvement—whether that's investing in teacher development, addressing attendance issues, or understanding how their funding compares. The goal of the tool is to empower schools to make data-informed decisions to support student success. Shortcomings of these models include the fact that the school data used was not from the most recent academic year. In addition, there is a chance that the rating system allowing only a rating from 1-10 may not be able to fully capture the intricacies and nuances of a school's characteristics. The future work would include developing an ML model based on more recent data and including more feature engineering that could allow for an even more reliable and accurate school rating system. Additionally, a fully fledged out web application focused on user experience could make the application easier for school administrators. Finally, future work can also focus on exploring more clustering algorithms to create more nuances rating groups.

9 Contributions

Pravallika (Code: Data pre-processing, regression models, rating system, classification models, Report: Abstract, Methods, Conclusion)
Anjali (Code: Data pre-processing, data visualization, rating system, classification models, Report: Experiments, Results, Conclusion)
Ananya (Code: Data pre-processing, K-means clustering, classification models, Report: Abstract, Results, Methods)

References

[1] Michael J. Brown and Glen I. Earthman. Examining secondary student achievement in large and small high schools in virginia. *Educational Planning*, 26(4):21–40, 2019.

[2] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane. Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning*, 17(12):236–251, June 15 2022.