Boston University
Questrom School of Business
BA305 B1

Academic Performance Analysis Report

Team 7

[redacted]

Deanna Soukhaseum

Table Of Contents

Problem

With the declining academic performance across the nation in recent years, potential initiatives and actions to improve students' test scores is a major area of interest for the government and other educational institutions.[1]

These departments or institutions have the ability to propose policies or implement measures in order to provide more support to struggling students. However, they first need to understand which students will likely require the most support as these institutions have limited resources available.

Our dataset, which looks at grades, along with a combination of factors that may impact these grades, can be a good knowledge base to understand how to minimize the disparities in academic performance due to uncontrollable factors, such as a student's gender or educational setting. As a result of this analysis, these institutions will be able to predict which students are most likely to underperform, so that they can figure out where to focus their efforts to make a significant difference.

Data Composition and Cleaning

Data Overview

Our dataset shows the academic results of students based on test scores gathered from 2,133 students who took the same exam. The original dataset contained 11 columns with details about each student, including their school, class size, gender, lunch qualifications, grades a week before taking the standardized exam (or the "pretest" scores), grades after taking the standardized exam (or the "posttest" scores), etc.

The students' grades after taking the exam (the "posttest" column) were determined to be the outcome variable of our analysis. Out of the data collected from 2,133 students, there were no missing values in any of the columns.

Data Cleaning

We decided to drop 3 features ("school", "classroom", and "student_id") that did not provide enough information to be directly relevant to predicting students' grades.

The "student_id" variable is unique to each student and, as a result, lacks any predictive power. Next, there were no inconsistencies between the "school" and "classroom" variables, indicating that the "school" and "classroom" variables provide us with the exact same information and that we would only need one of these columns at most. However, unlike the "school_setting" or "n_student" (the number of students in each class) variables, "school" and "classroom" on their own fail to provide any meaningful information on the students' learning environment that is not

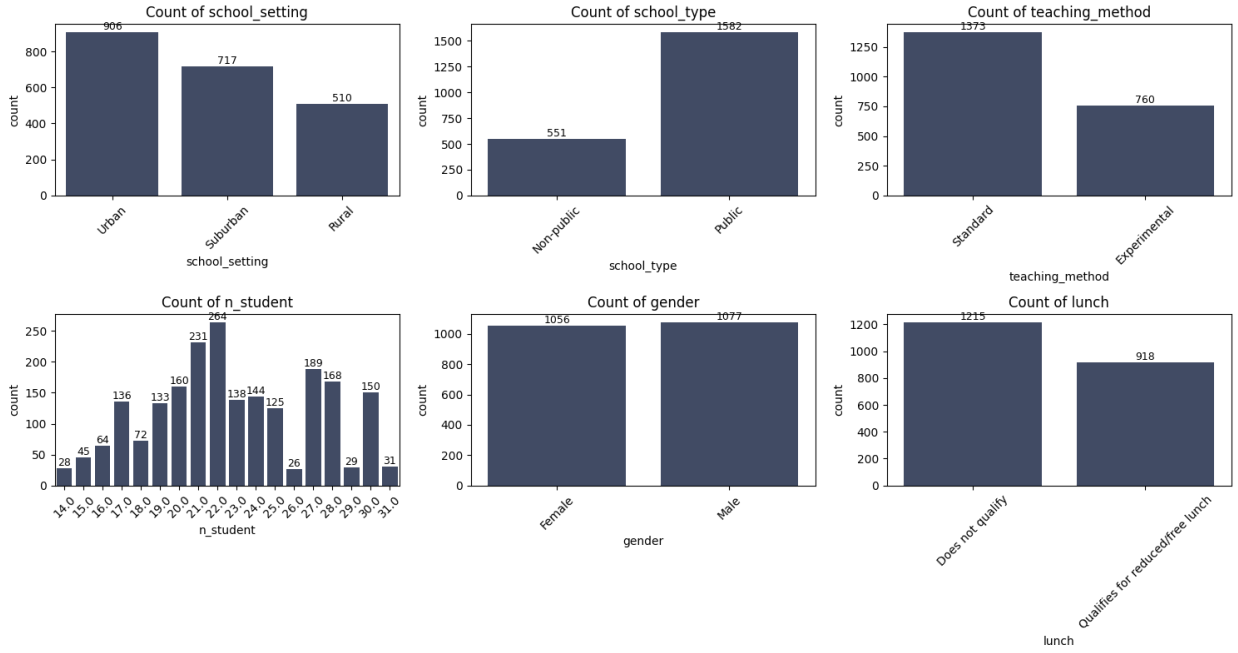
[1] https://www.brookings.edu/articles/the-alarming-state-of-the-american-student-in-2022/

already present in the remaining variables. Because of this, we determined that both “school” and “classroom” would not be useful for predicting the grades of students from schools or classrooms outside of the ones listed in our dataset.

Out of the remaining features, we had 5 categorical features, which we then converted into dummy variables to be used in our models. The following categorical variables were converted: “school_setting,” “school_type,” “teaching_method,” “gender,” and “lunch.”

Visualization

Before building our models, we visualized our data in order to discover potential findings that may impact our analysis. We first decided to investigate the distribution of our predictor variables using multiple bar charts as seen below.

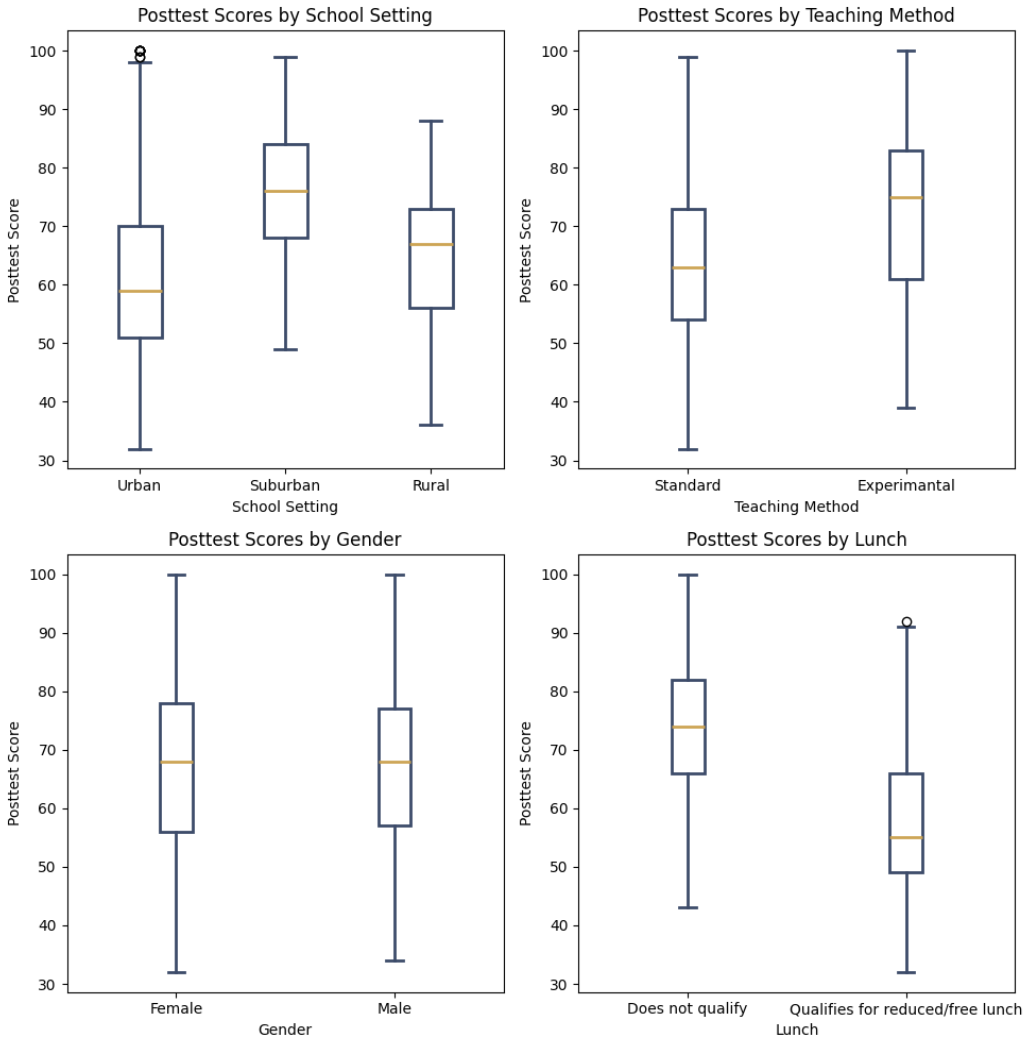


By using these bar charts of our predictors (excluding “pretest” scores), we discovered the following:

A. An even balance of male and female students took this exam.
B. Students who took this exam came from a variety of class sizes, ranging from 14 students to 31 students per classroom, with most students from classes with 22 students each.
C. Students who took this exam mainly attend school in an urban environment, followed by students from suburban schools as the second largest group before students from rural schools.
D. More of the students who took this exam likely come from a higher socioeconomic background as more than half do not qualify for free or subsidized school lunches.

E. Most of the students who took this exam attend a public school rather than private school.

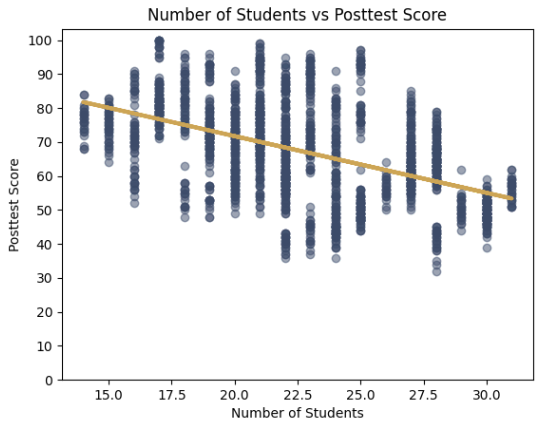
Next, we wanted to compare the distribution of "posttest" scores between our categorical variables by creating several boxplots as seen below.

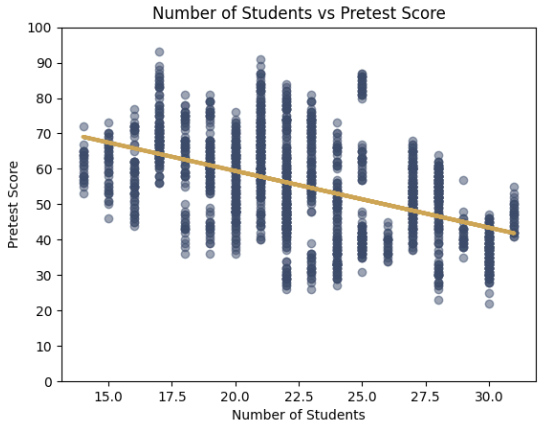


Through these boxplots of our categorical predictors, we discovered the following:

A. On average, students who attend suburban schools are more likely to receive higher "posttest" scores than students who attend rural or urban schools.
B. On average, experimental teaching methods are more likely to result in higher "posttest" scores than standard teaching methods.
C. Female students and male students have a similar distribution of "posttest" scores with average scores in the high 60s.
D. On average, students who qualify for reduced or free school lunch are more likely to receive lower "posttest" scores than students who must pay full price for school lunch.

Finally, we investigated the distribution of "posttest" scores across different values of our numeric predictor "n_students" (the number of students per classroom) by creating a scatter plot as seen below.





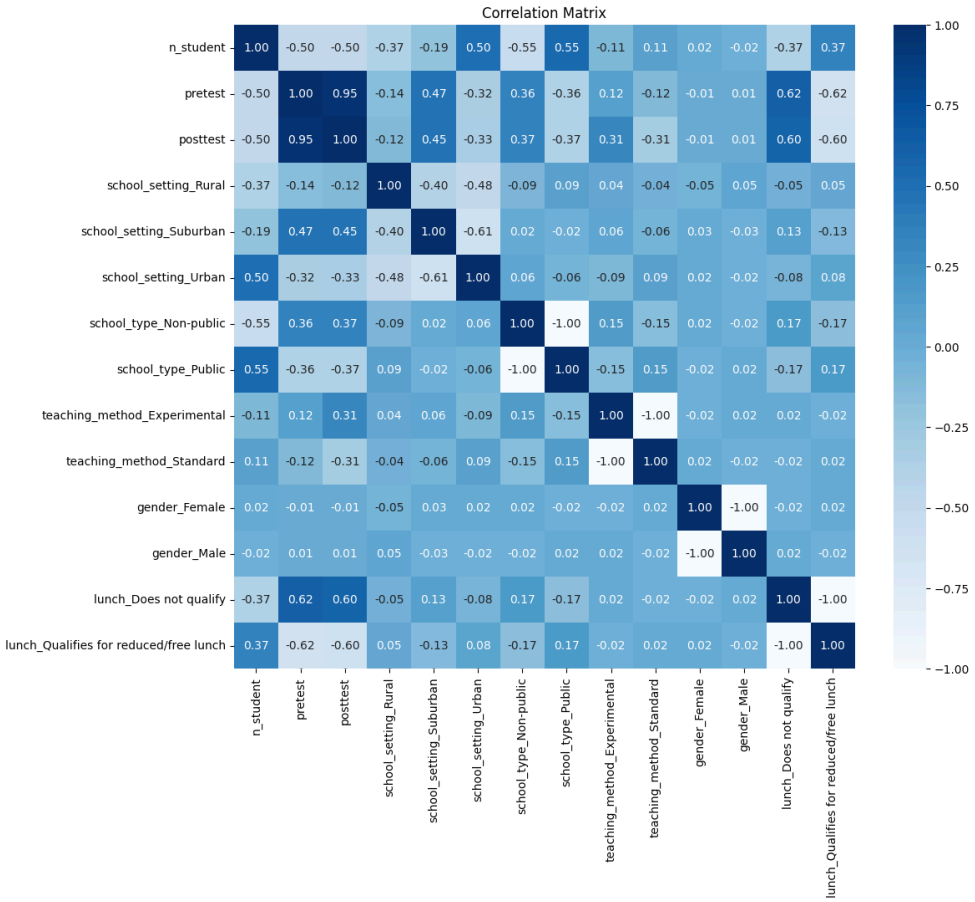
Through this scatter plot displaying the distribution of "posttest"and "pretest" scores against our numeric predictor, we discovered the following:

A. On average, students' "posttest" scores are more likely to decline as their class size increases.
B. Students from mid-sized classes have a greater range of "posttest" scores compared to students with the largest and smallest class sizes.
C. On average, there was less of a disparity across students from different class sizes when comparing the academic performance of students before taking this exam to their performance after taking the exam.

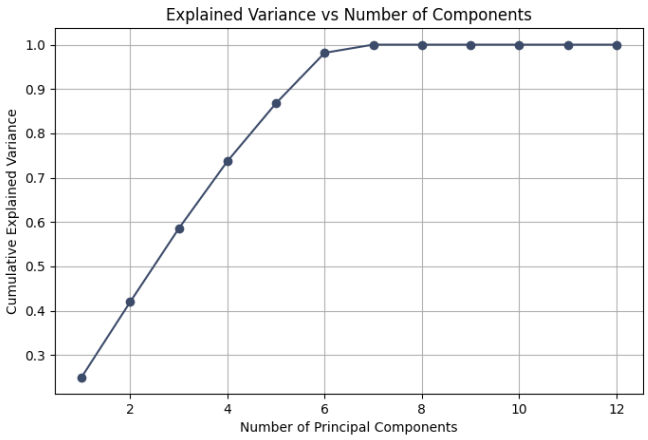
Multicollinearity & PCA

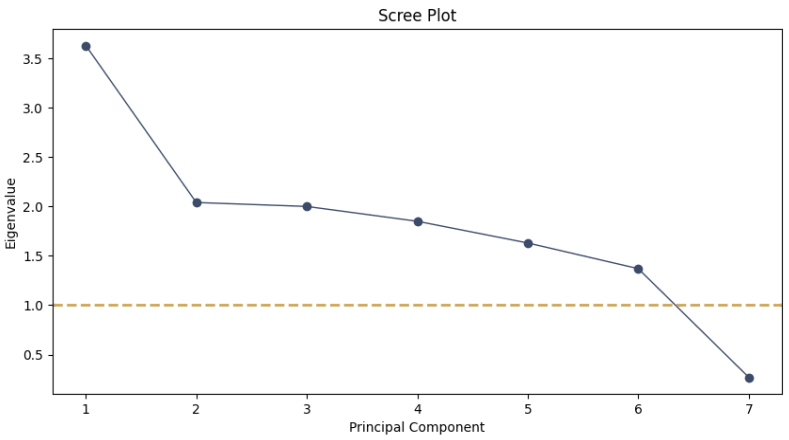
Using a correlation matrix, we decided to check for multicollinearity between our predictors in order to prevent any overlapping information between variables.

As seen in the matrix below, many of our predictors are highly correlated with our target variable and their associated dummy variables, but none of our predictors suffer from multicollinearity as none of them had strong correlations (their correlations did not exceed +/- 0.65). As a result, we decided not to drop any variables from our initial analysis of the correlation matrix.



We then decided to conduct a Principal Component Analysis to evaluate whether we would be able to further remove any information overlap between variables, allowing us to focus on fewer dimensions while preserving the most important student information.





As shown by the graph above, in order to retain 95% of the variance from our dataset, we would have to retain 6 of these new principal components to use in our predictive models. However, we

decided against using these principal components as we did not have much of an issue with multicollinearity, as seen by the correlation matrix, and we decided that using the new components would not have enough of an effect on our models to include it.

Modeling

We ran a total of seven prediction models in our dataset. Because our target variable is continuous and quantitative, we decided to use RMSE to measure how the model performs compared to our naive baseline.

Model	Naive Baseline	Linear Regression	KNN Regressor	Decision Tree - Full	Pruned Tree	Random Forest	XGBoost	Neural Network
RMSE	14.209	3.206	3.909	4.283	3.371	3.41	3.28	3.204
R^2	-	0.947	0.924	0.909	0.944	0.942	0.947	0.906

Naive Baseline

Before we began building our predictive models, we first used the mean of posttest scores in our training data as our benchmark for our naive baseline model. Our baseline achieved an RMSE of 14.209, indicating that, on average, our baseline predictions deviate from the actual results by around 14.209 points on a 100-point scale.

Multiple Linear Regression

We first ran a multiple linear regression model, as we believed that this would be the simplest prediction method due to there being less opportunities to tune the model. After running the model, we saw a significant improvement from our baseline model of around 11 points, with the RMSE of our linear model being 3.206 (See Appendix 1 for model outputs).

For our categorical variables, we used rural school settings, attending a private school, using an experimental teaching method, identifying as a female student, and not qualifying for subsidized lunch as the baseline variables.

Looking at the coefficients of the model, most of our predictors were statistically significant. However, the urban school setting, public school type, and male gender variables were not statistically significant, as their p-values exceeded the threshold of 0.05. As a result, the impact of these 3 variables on this prediction model is not strong enough to be considered meaningful based on the available data in our dataset.

Out of the remaining variables, having a greater number of students in each class, using a standard teaching method over an experimental one, and qualifying for subsidized lunch has a

negative impact on predicted test scores. On the other hand, having a higher pretest score positively impacts predicted test scores in this model.

KNN Regressor

Our k-nearest neighbors regressor model performed worse than our linear model, since our best KNN model had an RMSE of 3.909, which is around half a point higher than the RMSE of our linear model.

After running KNN models for k values between 1 and 30, we found that the optimal value for k, which is the number of nearest neighbors considered when predicting a data point, was 4 for our dataset, since it resulted in the lowest RMSE (See Appendix 2 for the graph of each model's RMSE).

Decision Trees

Our full decision tree resulted in a high RMSE of 4.283 and had a high complexity with 1775 nodes, a depth of 18 layers, and 888 leaves (See Appendix 3 for the tree visualization). Because of its complex structure, we were concerned about the model potentially resulting in high variance and issues of overfitting that can cause potential biases. Due to this complexity, we then pruned the decision tree by experimenting with parameter limitations.

We first tried pruning the tree by individually changing 3 parameters in our model: maximum tree depth, minimum tree size, and minimum samples per leaf node. We found that the optimal values that resulted in the lowest RMSE for each of these parameters individually were a maximum tree depth of 6 layers, a minimum tree size of 22 leaf nodes, and a minimum of 12 samples per leaf node (See Appendix 4 for RMSE graphs). The model that performed best out of these 3 models was the one where we limited the tree size to a minimum of 22 leaf nodes with a RMSE of 3.371 (See Appendix 5 for the tree visualization), compared to a RMSE of 3.548 for a maximum tree depth of 6 layers and a RMSE of 3.416 for a minimum of 12 samples per leaf node.

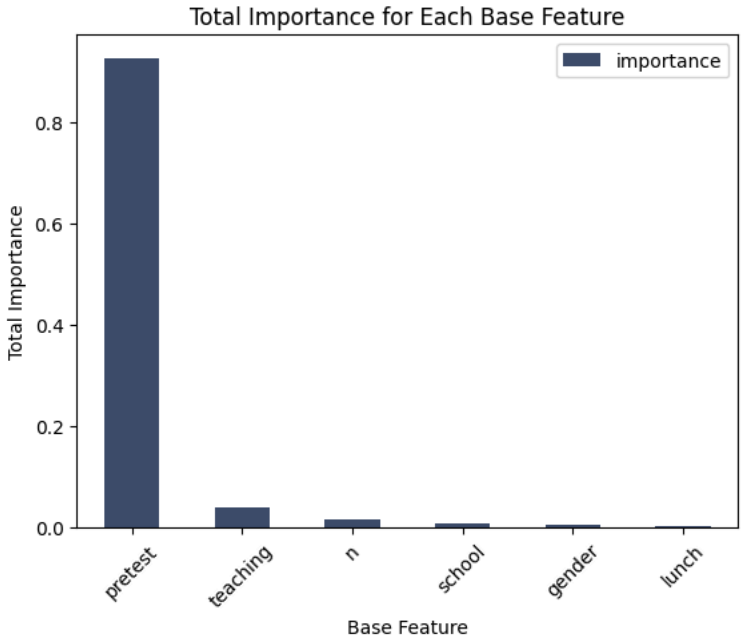
Next, we tried pruning the tree by limiting a combination of the following parameters: maximum tree depth, minimum tree size, minimum samples per leaf node, and minimum impurity decrease. We found that the optimal combination of parameters was a maximum depth of 5 layers, a minimum tree size of 50 leaf nodes, a minimum of 10 samples per leaf node, and a minimum impurity decrease of 0 (See Appendix 6 for the tree visualization). However, this combination resulted in a lower RMSE of 4.016, leaving the pruned tree with a minimum tree size of 22 leaf nodes as our best performing decision tree model so far.

Finally, we tried 2 additional methods to improve our decision tree models, which were 5-fold cross validation and XGBoost. The 5-fold cross validation method splits our data into five parts,

trains a model on four of them, tests it on the fifth, and repeats this process five times to ensure reliable performance. XGBoost is an algorithm that uses gradient boosting that learns from the mistakes of previous models to improve each subsequent model. While 5-fold cross validation did not yield better results with a higher RMSE of 4.22, XGBoost had the lowest RMSE out of all of the decision tree models with a RMSE of 3.28 (See Appendix 7 for tree visualization).

Random Forest

After building individual decision tree models, we created a random forest model, which combines multiple decision trees to improve accuracy and prevent overfitting. Our random forest model consisted of 1000 decision trees with a RMSE of 3.41. This model showed an improvement over our full tree, along with some of our pruned trees, but was still higher than our best pruned tree and our model that used XGBoost.



After building the random forest model, we found that the most important feature used to predict posttest scores was, by far, the students' previous academic performance, which was represented by the "pretest" variable. Following "pretest," other important features included whether an experimental or standard teaching method was used ("teaching") and the number of students in each class ("n").

Neural Network

As seen in the table below, we iterated through various numbers of nodes for up to 2 hidden layers. We found that the model that performed the best was 2 hidden layers with the first layer having 10 nodes and the second layer having 5 nodes. This model resulted in a RMSE of 3.224, which performed the second best with the lowest RMSE after our linear regression model.

Hidden Layer #	# Of Nodes	RMSE
1	9	3.246
	10	**3.213**
	11	3.224
2	4	3.271
	5	**3.224**
	6	3.268

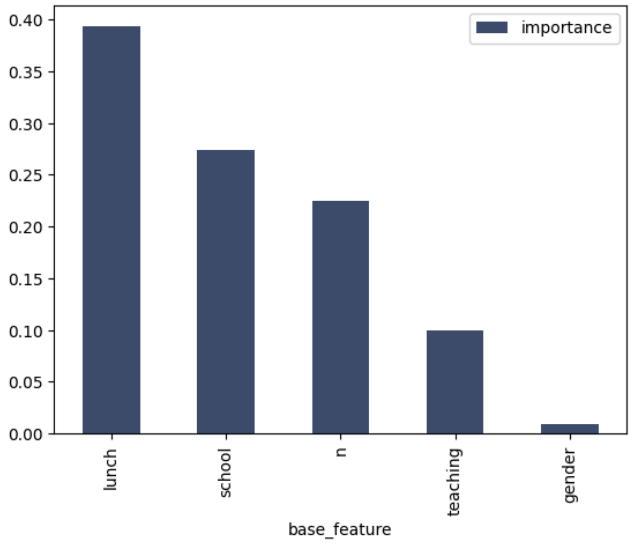
Similar to our decision tree models, we looked to further improve our neural network model by implementing 5-fold cross validation. By using 5-fold cross validation, we were able to lower our RMSE to 3.204, which made this our best performing model out of all the prediction models we tried.

Models Without The Pretest Variable

Because the "pretest" variable had by far the most impact on all of our predictions, we decided to try running all of our models without the "pretest" variable, as we were concerned that it was overshadowing the effects of our other variables. However, doing so increased the RMSE for all of our models, as shown in the table above, indicating that including "pretest" scores in our dataset is optimal for predicting students' "posttest" scores.

Model	Naive Baseline	Linear Regression	KNN Regressor	Decision Tree - Full	Pruned Tree	Random Forest	XGBoost	Neural Network
Optimal Parameters	-	-	k = 5	-	max_depth = 11	n_est = 1000	-	1 hidden layer, 2 neurons
RMSE	14.209	8.270	6.403	5.359	5.357	5.462	5.30	7.763
R^2	-	0.653	0.797	0.858	0.858	0.852	0.861	0.702

Looking at the feature importance of our prediction models that exclude the "pretest" variable, the most important features were slightly different. For example, when considering the "pretest" variable, students' socioeconomic status (represented by "lunch") had the smallest influence on academic success, but this variable had the largest impact on academic success when removing the "pretest" variable from our analysis. The type of school (public or private) and the number of students per classroom were the next most important features when excluding students' prior academic performance.



Challenges

Because of the lack of missing values in our dataset, the data cleaning and visualization were pretty straightforward and we didn't run into any problems in those areas. However, we still encountered a few obstacles when running our prediction models.

RMSE Issues

All of our initial prediction models' RMSE were higher than the RMSE of our naive baseline model due to mistakes in how we calculated our naive baseline. Once we resolved these mistakes, the RMSE of our naive baseline was much lower than it was originally.

Tree Overfitting

Before pruning, our decision tree created over 1,000 nodes, which caused us to suspect that the model was overfitting (See Appendix 3 for the full decision tree). Because of this, we made adjustments to our original decision tree model by pruning our tree, along with applying algorithms like XGBoost and k-fold cross-validation to enhance our models.

Lack Of Variables

Due to having few variables in our dataset, we were limited in the number of factors available for us to analyze. In terms of future improvements, it would be ideal to have additional numeric variables as most of our features were categorical. In addition, the "pretest" variable had significantly more predictive power for determining the "posttest" scores than any of the other variables in our dataset. This suggests that expanding our dataset to include additional variables in the future can provide deeper insights into key factors that determine academic success.

Conclusion

Our best model for predicting students' academic performance was our neural network model, which was enhanced by 5-fold cross-validation, with an RMSE of 3.204. Students' prior academic performance, represented by their "pretest" scores, was by far the most important predictor of academic success. Following this, the teaching method used, the number of students

per classroom, and the type of school (public or private) were the next most important features, though this was slightly different when excluding the “pretest” variable.

Appendix

Appendix 1: Multiple Linear Regression Model Output

OLS Regression Results

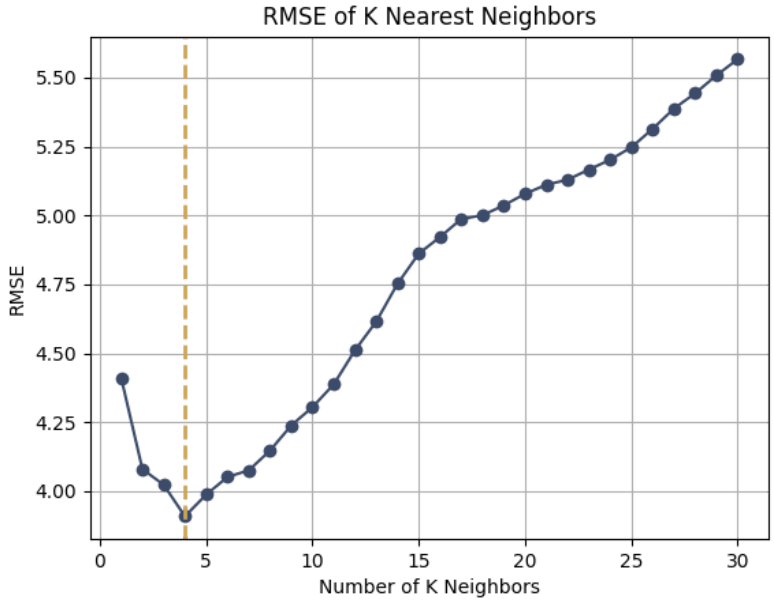
Dep. Variable:	posttest	R-squared:	0.947
Model:	OLS	Adj. R-squared:	0.947
Method:	Least Squares	F-statistic:	4785.
Date:	Fri, 02 May 2025	Prob (F-statistic):	0.00
Time:	19:37:50	Log-Likelihood:	-5511.6
No. Observations:	2133	AIC:	1.104e+04
Df Residuals:	2124	BIC:	1.109e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	23.2383	0.781	29.772	0.000	21.708	24.769
n_student	-0.0956	0.029	-3.258	0.001	-0.153	-0.038
pretest	0.9104	0.008	109.018	0.000	0.894	0.927
school_setting_Suburban	0.7680	0.220	3.490	0.000	0.336	1.200
school_setting_Urban	0.2115	0.245	0.862	0.389	-0.270	0.692
school_type_Public	-0.0361	0.230	-0.156	0.876	-0.488	0.416
teaching_method_Standard	-6.0332	0.148	-40.641	0.000	-6.324	-5.742
gender_Male	-0.0782	0.139	-0.561	0.575	-0.352	0.195
lunch_Qualifies for reduced/free lunch	-0.8976	0.192	-4.673	0.000	-1.274	-0.521

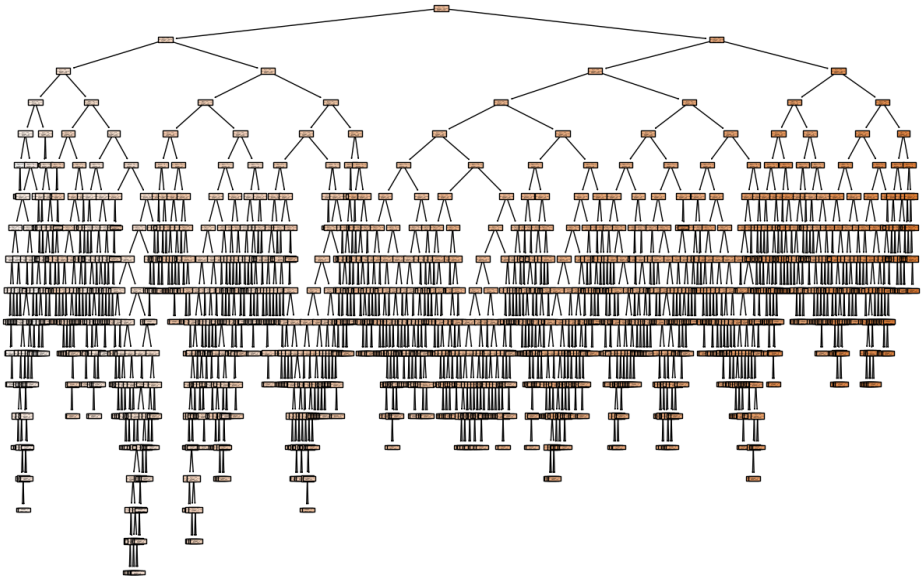
Omnibus:	2.388	Durbin-Watson:	1.874
Prob(Omnibus):	0.303	Jarque-Bera (JB):	2.474
Skew:	0.016	Prob(JB):	0.290
Kurtosis:	3.164	Cond. No.	692.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

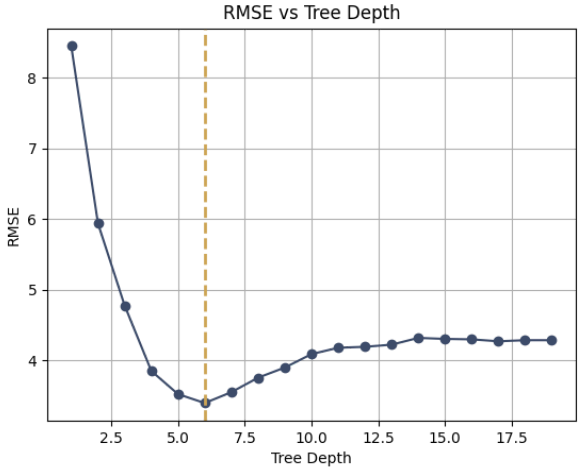
Appendix 2: RMSE Of KNN Models

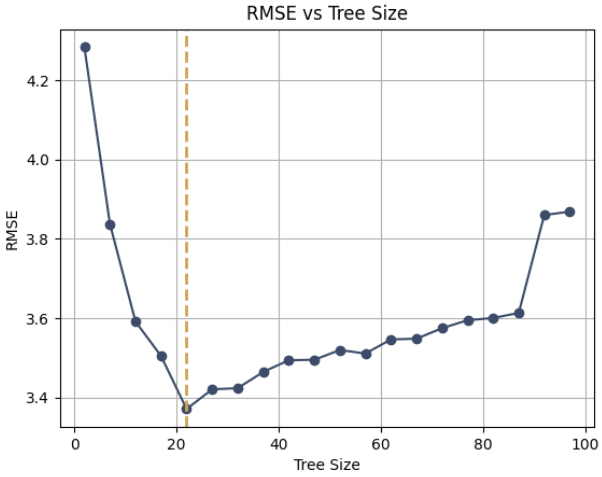


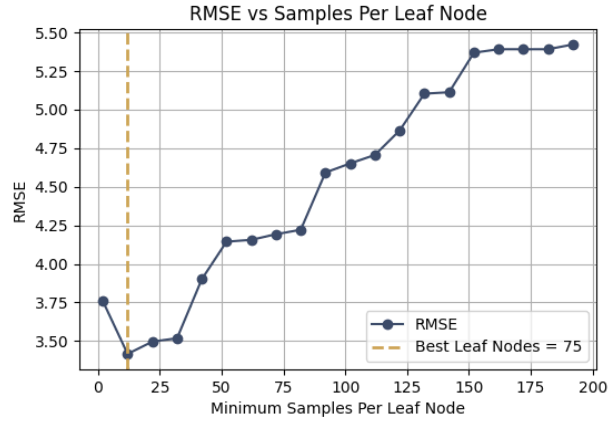
Appendix 3: Full Decision Tree Visualization



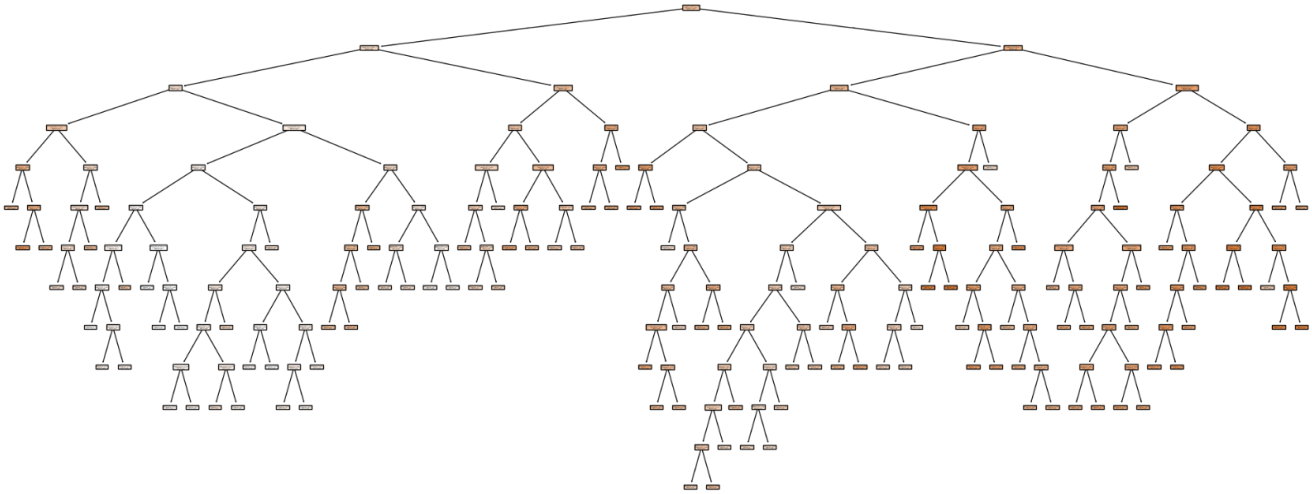
Appendix 4: Optimal Individual Decision Tree Parameters



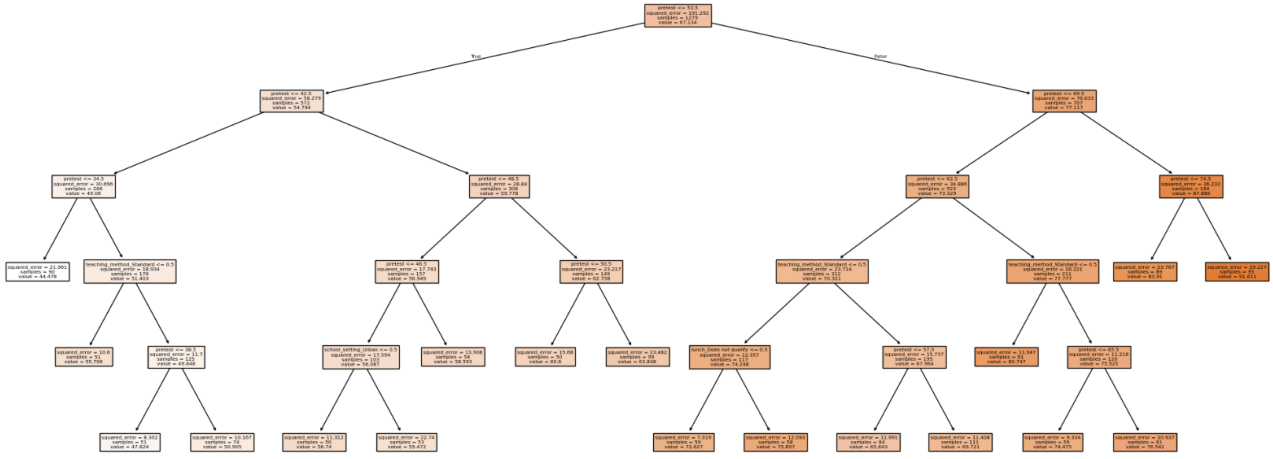




Appendix 5: Tree Visualization With A Minimum Of 22 Leaf Nodes



Appendix 6: Tree Visualization For The Best Combination Of Parameters



Appendix 7: Tree Visualization For XGBoost