Pandemic Classification: A study in New York State Graduation Rates

Jiaqi He

Lane Whitmore

Azucena Faus

Shiley-Marcos Engineering Department, San Diego University

Master of Applied Data Science

# Abstract

The intention of this study was to find a model that could classify the state of New York’s graduation rate data records as either pre-pandemic or pandemic from a set of features, so that further insights and research can develop regarding the shut-down of in-person instruction. The characteristics of importance were the subgroup names, percentage numbers regarding dropout rate, still enrolled rate, local diploma percentage, graduation diploma (which is a sum of the local and regents diploma), GED credential, non-diploma credential, regent diplomas and regent diplomas with Advanced designation. The hypothesis postulated that graduation rates would probably increase for the COVID timeframe, given the ease of working from home. However, there was the doubt that low-income households without access to internet or technology might counter this hypothesis. Data analysis was conducted on information collected by the New York State Department of Education, available on https://data.nysed.gov/downloads.php. It was determined that graduation rates exhibited an overall mean of about 85.3%, with a standard deviation of 8.4. Additionally, strong correlations were found between graduation rates and regent diplomas with Advanced designation. It was also found that count variables had strong associations between each other. The final model chosen revealed the most significant variables that effect graduation rates pre and during pandemic were subgroup name and all the percentage rates for each record. The final model revealed that graduation rates did, in fact, increase slightly for the one-year schools were in the middle of COVID pandemic shutdown, and the key performance indicators for each record were able to accurately classify each record as being Pre-COVID or COVID, with use of a proven ensemble classifier.

[Abstract 2](#_Toc99544442)

[Introduction 5](#_Toc99544443)

[Objective 5](#_Toc99544444)

[General Formula 6](#_Toc99544445)

[Method 6](#_Toc99544446)

[Bivariate Relationships 7](#_Toc99544447)

[Correlation 7](#_Toc99544448)

[PCA Analysis 7](#_Toc99544449)

[Modeling 7](#_Toc99544450)

[CART Classifier 7](#_Toc99544451)

[Logistic Regression Classifier 7](#_Toc99544452)

[Random Forest Tree Classifier 7](#_Toc99544453)

[Naïve Bayes Classifier 7](#_Toc99544454)

[Neural Network Classifier 7](#_Toc99544455)

[Results 8](#_Toc99544456)

[Final Model 8](#_Toc99544457)

[Discussion 8](#_Toc99544458)

[Weaknesses/Strengths 8](#_Toc99544459)

[Conclusion 8](#_Toc99544460)

[Future Research 8](#_Toc99544461)

[References 9](#_Toc99544462)

**Introduction**

Amid international news about a deadly variant of the SARS virus later named COVID-19, on March 7, 2020 as confirmed cases grew in the state of New York, Governor Andrew Cuomo declared a state of emergency. On March 15, 2020, New York, along with the much of the country and world, made the unprecedented move to shut down all public schools for in-person instruction. From there, several families chose to either keep their children in online learning, or pursue home schooling options, as well as private schools that remained open for much of the pandemic. Data has been gathered for graduation rates both before the pandemic shutdown of 2020, and during, as well as the tentative opening of schools where children had the option to come back. This data is used to compare graduation rates and look for any patterns regarding the COVID-19 shutdown. Because online learning requires access to technology and internet and, according to Friedman et al., about 10% of children that participated in online learning were found to not have adequate access to internet, it is possible that this affects the study (2021). Additionally, perhaps online learning helped boost education in some children who perhaps performed better without distractions from classmates.

The data on these characteristics was obtained from the New York State Department of Education website (2020) on graduation rates for New York State schools for the years 2018, 2019, 2020, and 2021, 788,342 records to be exact, along with various other parameters such as Report School Year, Aggregation Index, Aggregation Type, Aggregation Code, County Code, County Name, NYC index, Boces Code, Membership Code, Subgroup Code, Enroll count, Graduation percentage. With this information, perhaps a relationship can be found between these parameters and COVID-19 related graduation rates such that this study can add understandings to the effects of the school shutdown in New York State.

## Objective

It is likely that this study will reveal a relationship between statistical performance of students and the pandemic timeframe of Pre-COVID or COVID. This information can later be used to compare future performance of students post-pandemic.

# Methodology

The study consists of 10,587 county records on graduation rates and other related features from school years 2017-2018, 2018-2019, 2019-2020, and 2020-2021. These records were obtained from the State of New York Department of Education Data website (Data NYSED, 2020). The records were also imported from four files into a single file csv file for Python. The data included Graduation Year, County Codes, County Names, Membership Code, Subgroup Code, and Subgroup Name. Additionally, the data has numerical variables that includes count and percentages based off Total Enrollment that includes Non-Diploma Credential, GED, Dropout, Still Enrolled, Regents Degree, Regents Advanced Degree, and Graduation.

Null values from unreported parts of the state were forced to be removed from the data as there was no way to replace that data. This took the overall dataset count down to 9,229 records. A Pandemic variable was also constructed by classifying the school years 2017-2020 as Pre-COVID and the final year, 2020-2021, as COVID. In addition to this, a numerical binary version of Pandemic was created. Other cleaning involved included fixing Subgroup Names and County Names since both had some differences between years.

EDA

Descriptive statistics were gathered on the data. A categorical variable that was found to be useful was subgroup name. The mode for this categorical variable was tied between Male and Female.

Numerical characteristics included Graduation Percentage, Non-Diploma Credential Percentage, Sill Enrolled Percentage, Dropout Percentage, GED percentage, Regents Diploma Percentage, Regents Diploma with Advanced Designation, Local Diploma Percentage. The mean value for Graduation Percentage was 85.3%. The standard deviation was 8.4. Also, Dropout Percentage average was 8.1 with a standard deviation of 6.1. Figure 1 shows the differences in graduation rates for each timeframe Pre-COVID and COVID.

**Figure 1**

*Boxplots for Graduation Percentage Rates Pre-COVID and COVID*

Chart, box and whisker chart

Description automatically generated

Figure 1 contains the boxplot for graduation rates after data preprocessing. The difference in medians is apparent and shows COVID had a slightly higher median and mean graduation rate. Also, the variance between the 25th and 75th percentiles (or 1st and 3rd quartiles) is smaller for COVID as well. Any data that lies outside the 3 times the Interquartile range are denoted as single points – outliers, of which there are more in the Pre-COVID timeframe.

**Figure 2**

*Correlation Matrix for New York State Graduation Data (2018-2021)*

Graphical user interface

Description automatically generated with medium confidence

From the correlation heat map, we see that the counts are highly correlated with each other and that reg\_adv\_pct\_d is moderately (0.63) correlated with grad\_pct\_d. The variable Reg\_Adv\_pct\_d is the percentage of graduates that receive regents diploma with advanced designation, so as graduation rates increase, so too does this percentage increase. Additionally, grad\_pct\_d is also related to variables that were derived using graduation rates/counts like ged\_diploma\_cred\_pct. Finally, grad\_pct\_d is also negatively correlated with dropout\_pct\_d.

# Modeling

Six types of classification models are generated and evaluated on the preprocessed data that was randomly partitioned for training and test, on the constructed Pandemic variable.

**CART**

The CART method creates decision trees that are a binary split. Each decision node, then, contains exactly two splits. CART decides the fit based on goodness, which is defined as (Larose, 2019):

(1)

The best split is then defined by the highest Φ(s|t) measure after measuring all possible splits. Figure 4 below gives a visualization of the decision tree built from the dummy subgroup names and percentages predictors listed within the dataset. The decision tree has been limited to a maximum of 25 leaves to be appropriately sized for this format. The variable that shows up most commonly is the Local Percentage variable, which likely means that it is a fairly strong predictor of whether the school year was during the pandemic.

**Figure 4.**

*Decision Tree for CART method.*

Diagram, schematic

Description automatically generated

## Random Forest Tree Ensemble Classifier

The Random Forest Ensemble was run using the gini method. The predictor variables were dummy variables for Subgroup Names in addition to each percentage available within the data. Random Forest Ensembles takes a series of randomly generated CART decision trees (bagging) to then use for the final classification decision, which involves a majority vote decision. The data is randomly sampled with replacement for each decision tree (bootstrap sampling) constructed by the ensemble (Larose, 2019).

(2)

Because the data for Pre-COVID consisted of 73% of the data, a valid classifier needed to perform at least better than 73% (A baseline predictor that predicts all Pre-COVID, would have at least 73% accuracy). This model exceeds this requirement at 86%. Sensitivity is about 66% which means it can predict the smaller class relatively well.

The Area Under the Curve measurement for the ROC curve of the Random Forest classifier further validates the model by having an AUC well above 0.5 at about 0.798. Figure 5 depicts this result.

**Figure 5**

*AUC ROC curve for Random Forest Classifier*

Chart

Description automatically generated

## Naïve Bayes Classifier

The Naïve Bayes model was built using the percentages in the data set in addition to dummy variables representing Subgroup Names. This approach focused on the implementation of the Bayes Theorem where X and Y represent posterior probabilities (Larose, 2019):

(3)

The resulting model showed that separating the statistical difference between Covid and non-Covid years would prove difficult, at least with this method. The model only had 64% accuracy over the test set. The model has flawed sensitivity as only 39% of true positive, Covid classification, predictions were correct. Overall, the model underperformed the baseline model that classified everything as Pre-Covid.

## Neural Network Classifier

The same predictors were used to create a Neural Network Classification, but for the Neural Network model, the numerical data was min-max scaled to help boost the model performance. The Neural Network was constructed using the Tensorflow and Keras packages in Python. Our model has been activated using the sigmoid function:

(4)

While the overall formula utilized can be described by (Larose, 2019):

(5)

By using the sigmoid function, the different behaviors within the data set with act the same in order to input into the model. With dummy variables there are 15 input nodes, 32 nodes in the hidden layer that activates using the ‘relu’ function, and the final output layer, that is activated by the sigmoid function, produces two results either 0, Pre-Covid, or 1, Covid.

**Boosting Ensemble Classifier**

While Boosting Ensembles are more challenging to conceptualize than other classifiers, they are often better at classifying difficult problems. The Boosting Ensemble was created using the same data prepared for the Naïve Bayes Classifier. The Boosting base classifier depends on its error rate, defined by:

(6)

The importance of this classifier is then further defined by the following hyperparameter:

(7)

If error rate is high the hyperparameter value will be negative. If the error rate is low the hyperparameter will be a positive value. The parameter will then be used to update the weight in order to begin the next run through the data (Tan et al., 2019).

**Bagging Ensemble Classifier**

Bagging Ensemble is another method that is more challenging to visualize or conceptualize. The same set of predictors has been implemented here as the previous ensemble. Bagging takes a bootstrap sample of the predictors where each sample has a probability of being implemented into the bootstrap. A split point is then chosen to that most effectively minimizes the entropy. Which creates a decision stump (Tan et al., 2019). The bootstraps are then used to make multiple classifiers that are combined to make a single classifier.

# Results

All the models’ evaluation measures are listed below in Table 2. As far as overall accuracy, the Random Forest Ensembles performed the best at 85.99% accuracy. For comparison, a baseline model that predicts everything to be Pre-Covid would have an overall accuracy of 73.71%. The model that performed below the minimum baseline is the Naïve Bayes model.

In terms of sensitivity, the Bagging ensemble performed best, with a sensitivity of 77%. The other model that stands out is the Decision Tree. The Decision Tree has a sensitivity of only 32%, meaning, the Decision Tree does not predict the minority class at a very high proportion. This imbalance makes the Decision Tree among the weaker models with the Naïve Bayes.

All the models, aside from the Naïve Bayes, performed above 75% specificity, the most specific being the Random Forest. Yet, the lower of sensitivity with the Random Forest detracts from that performance. Of the ensemble methods that returned high sensitivity, the Bagging method returned the highest sensitivity at about 77%. Overall, Random Forest gave the highest accuracy, Specificity and Precision, but Bagging gave relatively close accuracy and good Specificity and Precision, so Bagging performed best when concerned about the smaller class COVID (which we are concerned with), with Random Forest as a close contender for final model.

**Table 1.**

*Evaluation Measurements for Each Model*

## 

## Final Model

Ultimately for the final model, the Bagging Ensemble proved that it showed overall the best performance for the minority classification.

(3)

# Discussion

## Conclusion

Although the data for the COVID classification consisted of only one year, there was enough of a difference in student performance to allow recognizable patterns to be found by some ensemble classifiers that exceeded the minimum accuracy requirement of 75%. This can provide insights into future analysis of student performance with the added use of technology in schools which will help students work from home and make up work easier than in the Pre-COVID era. Furthermore, perhaps predictive analytics can assert future behaviors in the post-pandemic era and find correlations yet to be seen in performance and the events of the COVID timeframe not currently available in the dataset used in this study.

# References

Larose, C., & Larose, D. (2019). *Data Science Using Python and R*. John Wiley & Sons, Inc.

Tan, P.N., Steinback, M., Karpatne, A., Kumar, V. (2019). Introduction to Data Mining, 2nd Edition. Pearson Education.

New York State Education Department. (2018-2021). *Graduation Rates*. Retrieved March 11, 2022. <https://data.nysed.gov/downloads.php>

Friedman, J., York, H., Mokdad, A. H. (February 17, 2021). *U.S. Children “Learning Online” during COVID-19 without the Internet or a Computer: Visualizing the Gradient by Race/Ethnicity and Prental Educational Attainment*. SAGE journals. [https://doi.org/10.1177/2378023121992607](https://doi.org/10.1177%2F2378023121992607)