# Project Report: Student Performance

Student performance on standardized tests is used as a measure of how much the students have learned and how successful their education program was. Each individual’s performance is driven by numerous factors, which often includes socio-economic factors like gender, race, and parental education levels. Factors that the school or program can control like providing nutritious meals that keep the students energized and focused, or ensuring that they complete exam preparation courses can have a significant impact as well. Being able to track these various factors and used them to model and predict student outcomes can help identify areas to improve a school or education program to make sure that any potential issues or biases in the outcomes are identified to be addressed.

Three data modeling algorithms were tested on an example set of data from Kaggle, attempting to predict two factors, whether the average score for a student represents a passing or failing grade, as well as whether as student is likely to significantly underperform expectations and may require additional assistance. An ensemble predictor combining all three models was assessed as well. The results for each model are shown in Table 1.

**Table 1: Model Prediction Results**

|  |  |  |
| --- | --- | --- |
| Model Type | Accuracy | Type I Error |
| Decision Tree | 60% | 42% |
| Naïve Bayes | 66% | 39% |
| Support Vector Machine | 53% | 46% |
| Ensemble | 64% | 38% |

All of the models had similar levels of accuracy of Type 1 error, but the Naïve Bayes classifier performed the best with 66% accuracy and 39% Type 1 error. These are acceptable results for the sake of demonstrating the usefulness of these modeling algorithms, though overall performance of these models is likely driven by the fact that the data selected was generated for educational purposes and is not necessarily representative of real-world data.

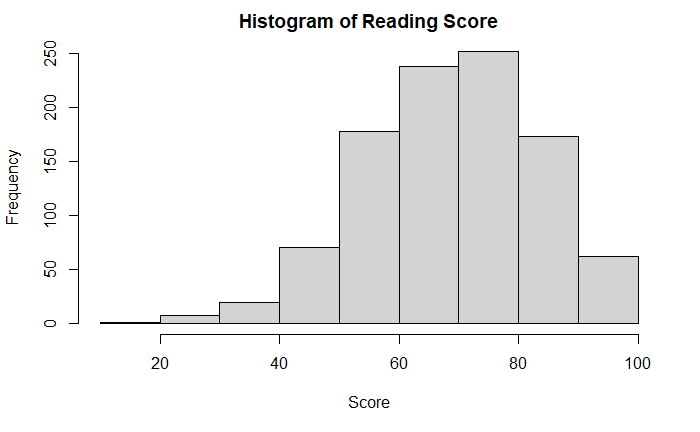
# Data Exploration

The example data for this report comes from the Student Performance on Exams dataset from Kaggle, <https://www.kaggle.com/spscientist/students-performance-in-exams>. This dataset contains 1000 observations of 8 total variables, 3 test scores for math, reading and writing, and 5 socio-economic and education factors, all of which are detailed in Table 2. It should also be noted that this dataset was artificial generated for instructional purposes only and may not be representative of an actual dataset.

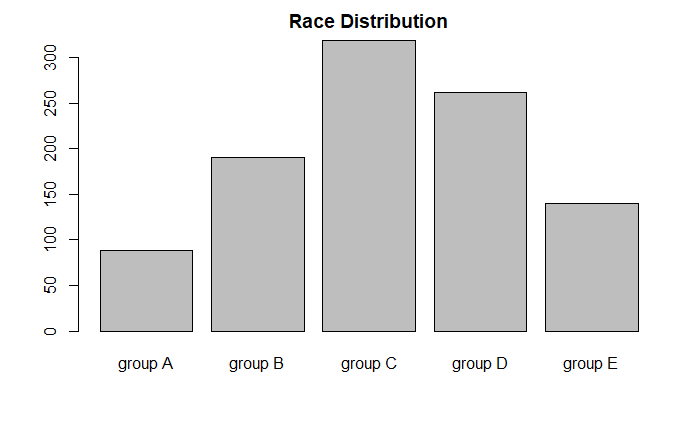
**Table 2: Data Parameters**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Values |
| Gender | Factor | Male/Female |
| Race Ethnicity | Factor | Group A/B/C/D/E |
| Parental Level of Education | Factor | Some High School  High School  Some College  Associates Degree  Bachelor’s Degree  Master’s Degree |
| Lunch | Factor | Free or Reduced  Standard |
| Test Preparation Course | Factor | Completed  None |
| Math Score | Integer | 0 to 100 |
| Reading Score | Integer | 0 to 100 |
| Writing Score | Integer | 0 to 100 |

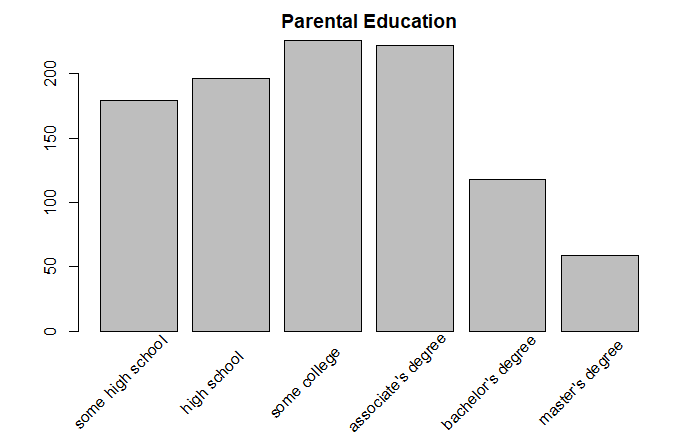
An initial exploration of the data shows that all three test scores have pretty similar distributions, with mean scores in the mid to upper 60’s, and left-hand skewed towards the lower scores. Figure 1 shows the distribution of reading scores as an example. There are roughly equal numbers of male and female students, a varied array of ethnicities (Figure 2), and parental levels education (Figure 3) vary as one might expect with most having completed high school or some level of college, and fewer with higher level degrees like a bachelor’s or a masters. Two variables of note are the lunch and test preparation course, which shows roughly 1/3 of the students receive some sort of free or reduced-price lunch while the rest receive the standard lunch, as well as only 1/3 of the students having completed the test preparation course while the rest have not.



**Figure 1: Reading Scores**



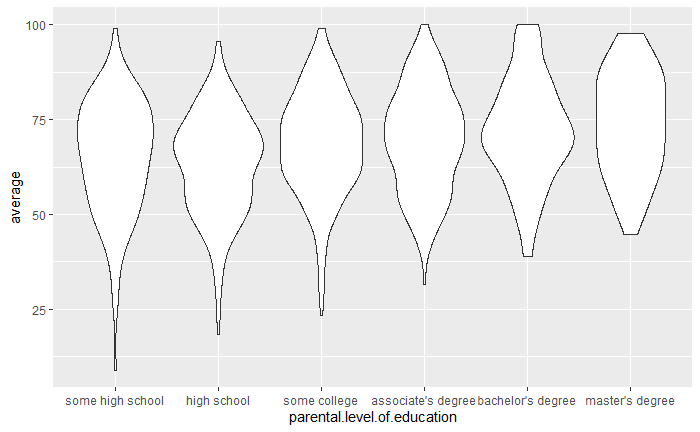
**Figure 2: Race Distribution**



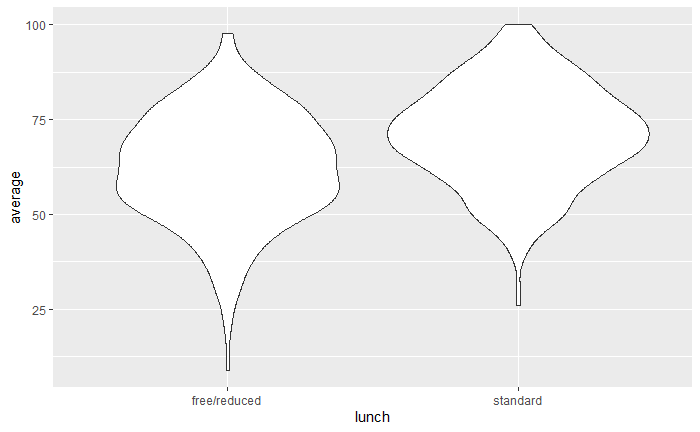
**Figure 3: Parental Education**

There are some very peculiar shapes in the distributions of some of the data. If we look at the distribution of average scores across levels of parental education, we see that that there each level of education has a slightly different shape, some that could be a normal distribution, while others are possibly multi-modal or even somewhat uniform. These distributions are plotted in Figure 4. There is also a trend in that the minimum grades tend to increase as the education level increases, but the average seems to remain around the same level which will make it quite difficult to identify any statistical differences between each group.

If we look at another factor like lunch type, as shown in Figure 5, we see that the distribution of scores for those that receive the standard lunch appears normal, but the distribution for those that receive a free or reduced lunch is flat at the peak and has an extended tail down to very low scores. This may be indicative a socio-economic factor that is truly impacting student performance, and is likely worth further attention while we build predictive models.



**Figure 4: Average Scores vs. Education Level**



**Figure 5: Average Scores vs. Lunch Type**

# Model Development and Feature Engineering

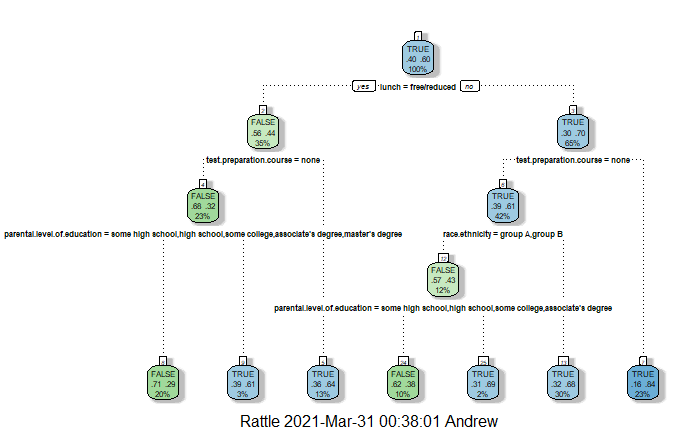
Instead of building a model for each individual test score, we can take an average of the math, reading, and writing scores for each student, and then create a pass/fail grade to assign to each student. For the development of these models, the average grade will be considered passing if it is greater than the mean of the average student scores, which is 67.8. For each model training data is selected randomly, totally 2/3 of the total available data. The remaining 1/3 is used for testing and validating the models. The results of the Decision Tree and the Naïve Bayes classifiers produce a percent-likelihood for whether the average score is passing or failing, which will be discretized into a prediction of a passing grade if the likelihood is greater than 50%.

## Decision Tree

The decision tree model demonstrated 69% accuracy with the training data, and 60% accuracy with the test data in predicting whether a student would have a passing grade on average. The most significant factors in determining the average score were the type of lunch that the students received and whether they had completed the test preparation course. Other contributing factors were the race/ethnicity of the student and the parental level of education, while gender could be completely ignored without affecting the result.

**Table 3: Confusion Matrix, Decision Tree**

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | **Fail** | **Pass** |
| Fail | **94** | **64** |
| Pass | **68** | **108** |



**Figure 6: Decision Tree Model**

One item of concern with this model is the high level of Type I error at almost 42%, meaning these students are projected to have an above average grade but will actually have a below average grade.

## Naïve Bayes Classifier

The Naïve Bayes classifier achieves slightly better performance than the Decision Tree, with a 69% accuracy on the training data and a 66% accuracy on the test data. However, even with the slight improvement in overall accuracy the Type I error is still roughly 40%.

**Table 4: Confusion Matrix, Naïve Bayes**

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | **Fail** | **Pass** |
| Fail | **99** | **50** |
| Pass | **63** | **122** |

## Support Vector Machine

The Support Vector Machine was the lowest performing model of the three evaluated, with an accuracy of only 53% and a Type I error rate of 46%. This was rather disappointing as SVM models are typically quite accurate.

**Table 5: Confusion Matrix, Support Vector Machine**

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | **Fail** | **Pass** |
| Fail | **88** | **83** |
| Pass | **74** | **89** |

## Ensemble Model

In cases where individual models have relatively low accuracy score like those shown here, we can sometimes achieve better results by building a model that takes the predictions of many models in an ensemble approach, and uses the most common prediction across all of the models for each student.

An ensemble model of the three models shown here did not produce significantly better results than any of the individual models, with a 64% accuracy and 38% Type I error rate.

**Table 6: Confusion Matrix, Ensemble Model**

|  |  |  |
| --- | --- | --- |
|  | Reference | |
| Prediction | **Fail** | **Pass** |
| Fail | **100** | **57** |
| Pass | **62** | **115** |

## Conclusion

For the highest accuracy with the most simplicity the Naïve Bayes classifier was the model for predicting whether an individual student would achieve a passing grade or not. All of the models were relative “low learners” with a maximum accuracy of 66%, however this may be due to the data being synthesized for educational purposes. The takeaway is that these types of algorithms can still be useful for educators to help identify students that are likely to succeed on their own, and those that will likely need additional assistance to be successful.