The Study on Advanced Placement Exams on Collegiate Academic Performance

Madison Sarmiento, Computer Science
Mentor: Dr. Lan Yang
Kellogg Honors College Capstone Project

Introduction

Throughout high school, students are encouraged to take Advanced Placement (AP) tests in order to receive college credit for certain classes, boost their GPA, and give themselves an edge in college applications[1]. These tests range from humanities to STEM subjects and a good score can indicate that the individual may be successful at their chosen college/university. However, is this assumption true? In this project, I hope to determine the validity of this belief using a machine learning algorithm to predict whether or not a student's college GPA correlates to their AP scores.

Objective

The objective of this project was to build, train, and test a classification algorithm known as the random forest classifier. I used python as my main language with the help of the sci-kit learn[2] package to aid in creating this model. Additionally, data was collected from Cal Poly Pomona students to run this test as it was vital for the algorithm in both training as well as testing.

Methods

The main steps in this project included data collection, understanding the random forest classification algorithm, and determining the feature of importance.

Data Collection: For this project, I sent out a survey to students at Cal Poly Pomona asking for their current GPA, the AP tests they took, their AP test scores, and their major. The survey was open for a month to all majors.

Random Forest Classifier[2]: One of the most important parts in machine learning is the ability to classify data into categories. For the random forest classifier, classification is done by training the model with a part of a data set, then testing it with the remainder - this is known as supervised learning. When starting, the random forest classifier takes in data and creates an individual decision tree for each data entry. Each decision tree then comes up with a conclusion, which are compiled through a regression model. The output from this regression model will be the most common result of the trees.

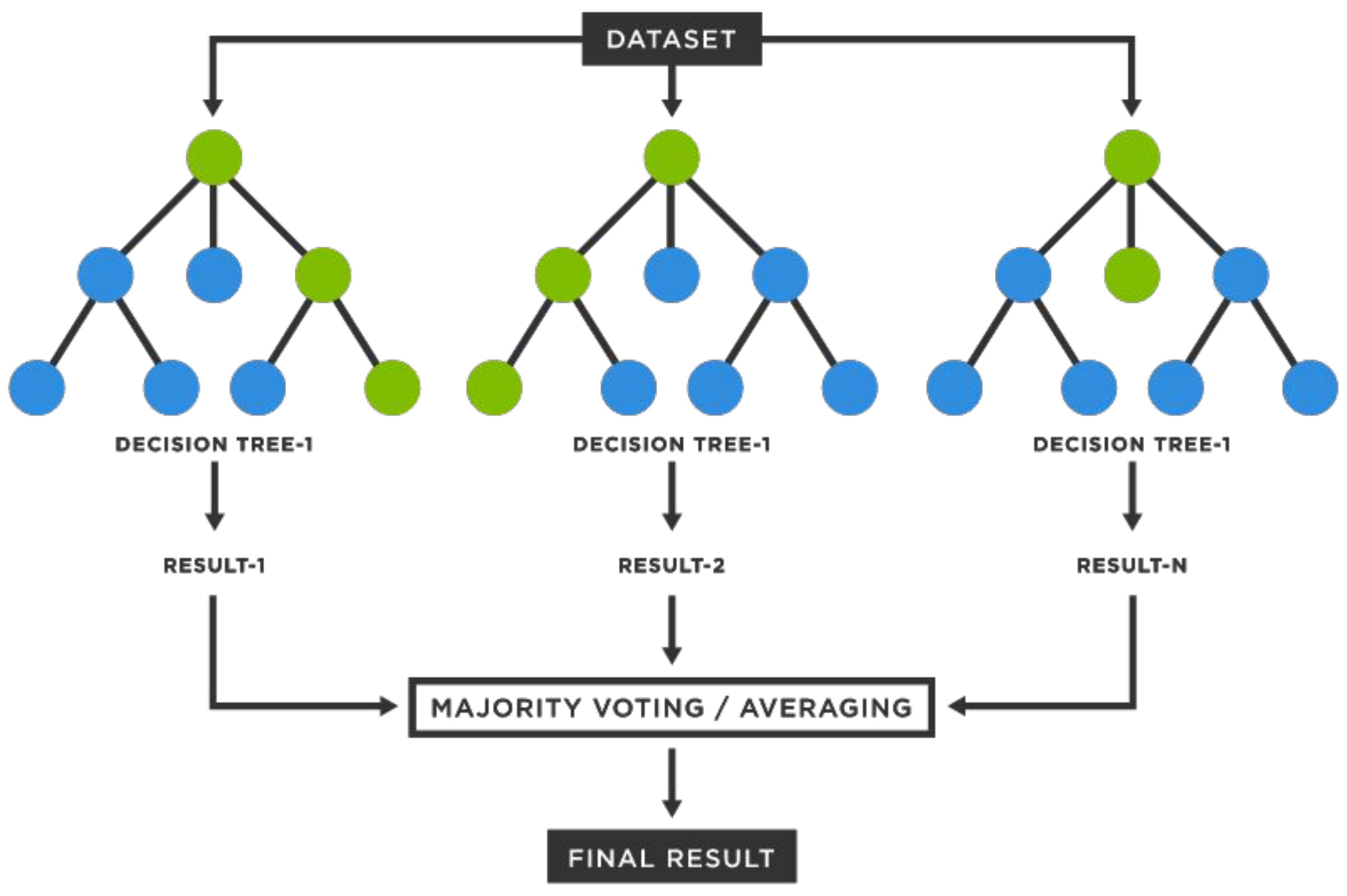


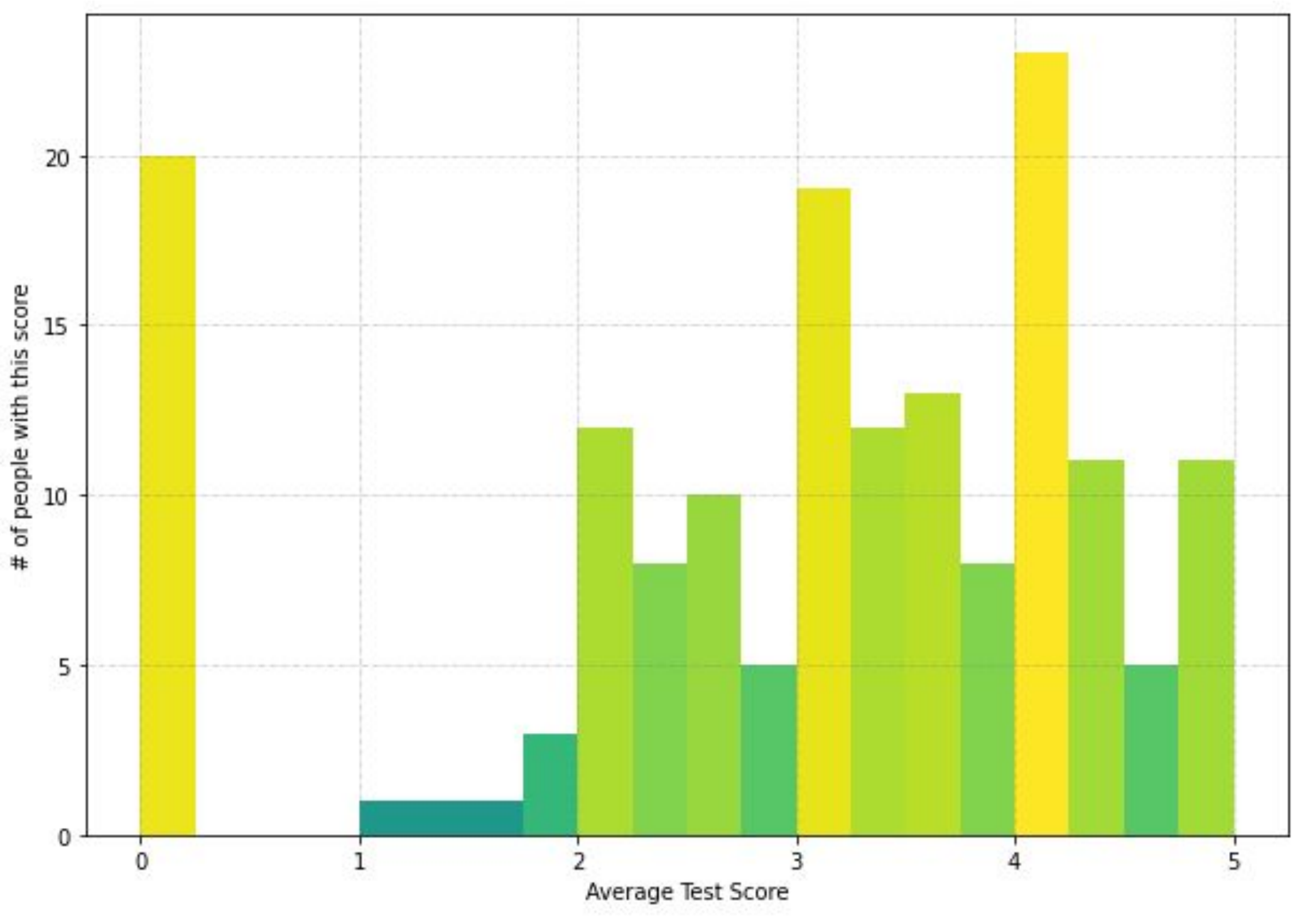
Fig 1. Diagram of how the random forest classifier algorithm works.

For this project, the dataset given to the model included the number of tests a student took and their average score. The data was split into two subsets - 67% of the data was used to train the model and 33% was reserved for testing. With the training set, the algorithm took in the number of tests a student took and the test scores of each one and predicted if their GPA was above or below a 3.5 - a common GPA that grants a student honors distinction at graduation. After training, the test data was used to check how accurate the predictions of the model were. The returned percentage was then analyzed to answer our original question.

Feature of Importance: The feature of importance in a model was used to determine which input value influenced the random forest classifiers outcome. There are various ways to calculate the feature of importance but for this project I only used two - the permutation method and the Mean Decrease in Impurity (MDI) method.

- Permutation Method[2]: The model will "shuffle" the different values and run the random forest classifier model again. If the shuffled value increases the model error, it means that the classification relied on that feature, and thus, it was important. This method is fairly simple to implement in any dataset and does not require new data in order to run. The drawbacks of this method are that it is more expensive to run based on the number of inputs and it can favor higher values in the model.
- MDI Method[3]: This model calculates the feature of importance based on the number of "splits" in all trees divided by the number of samples it splits. In other words, it weighs the probability of reaching a decision over the number of trees in the "forest". This method is great for eliminating "noisy" data (data that doesn't fit within the chart), however it can be biased towards the most common data in a data set.

Results



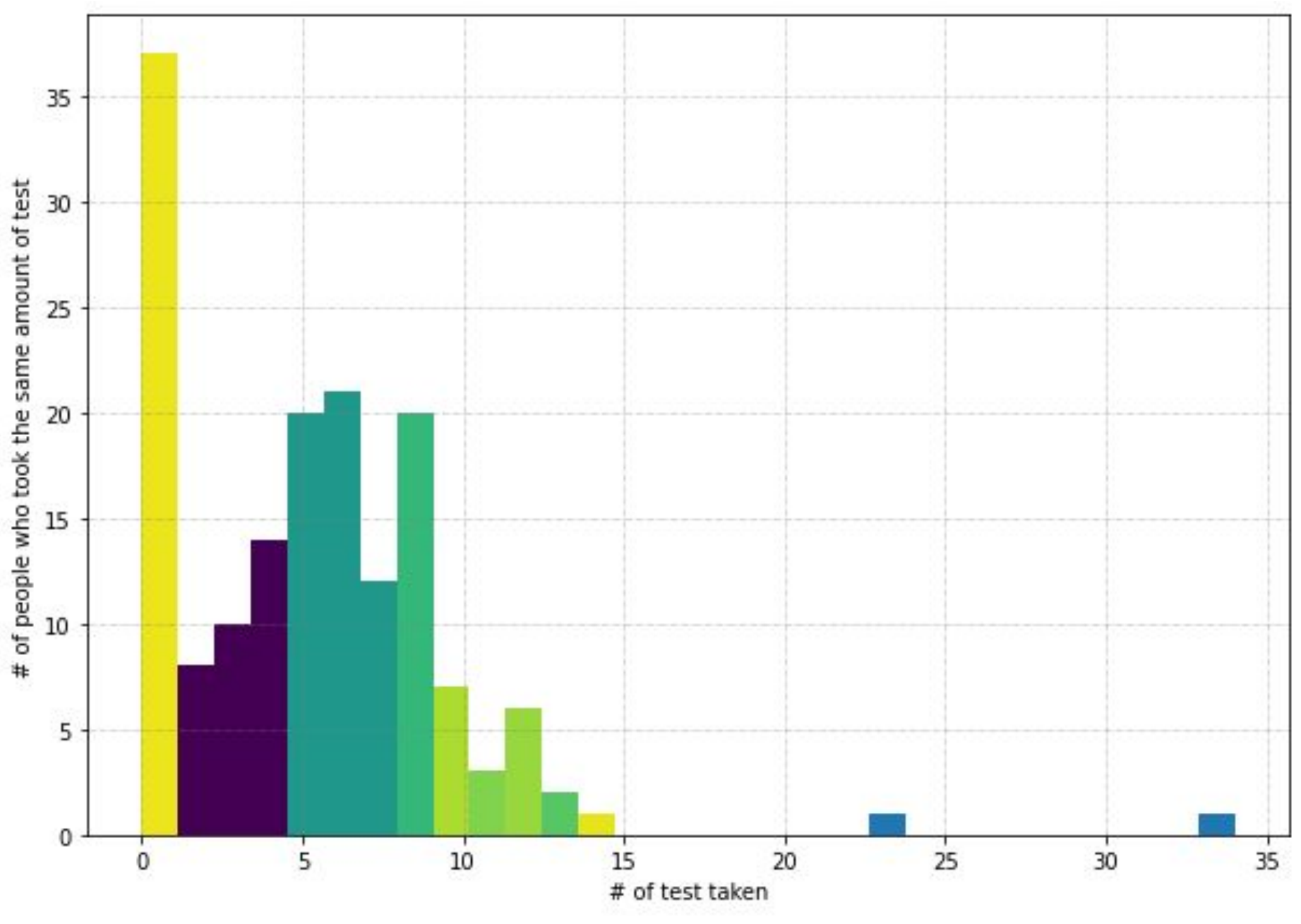
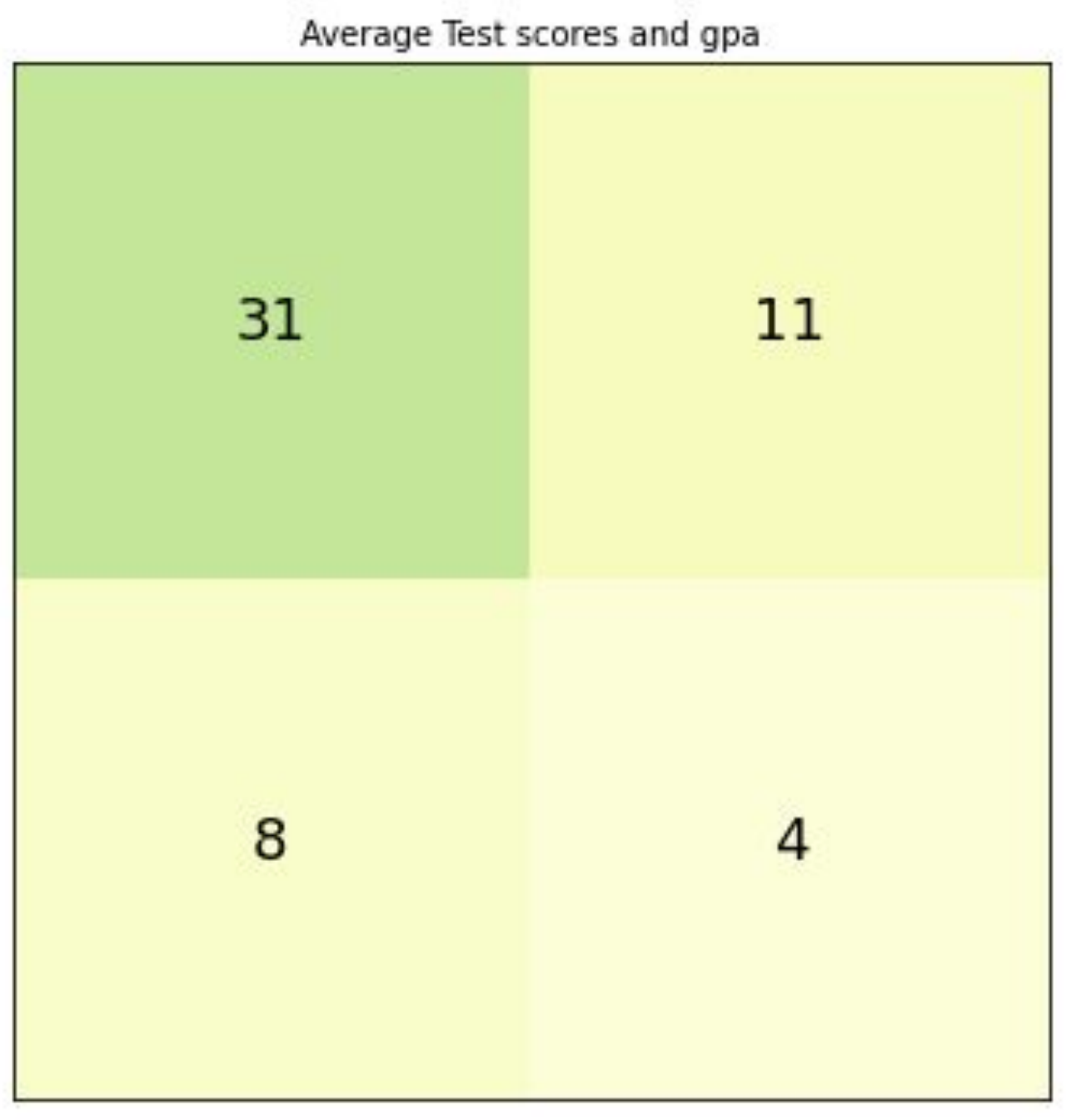


Fig 2. Visualization of the data collected in the survey for average test scores. There is a right skew meaning that the students who took the survey were more likely reporting a score of 3 or higher.

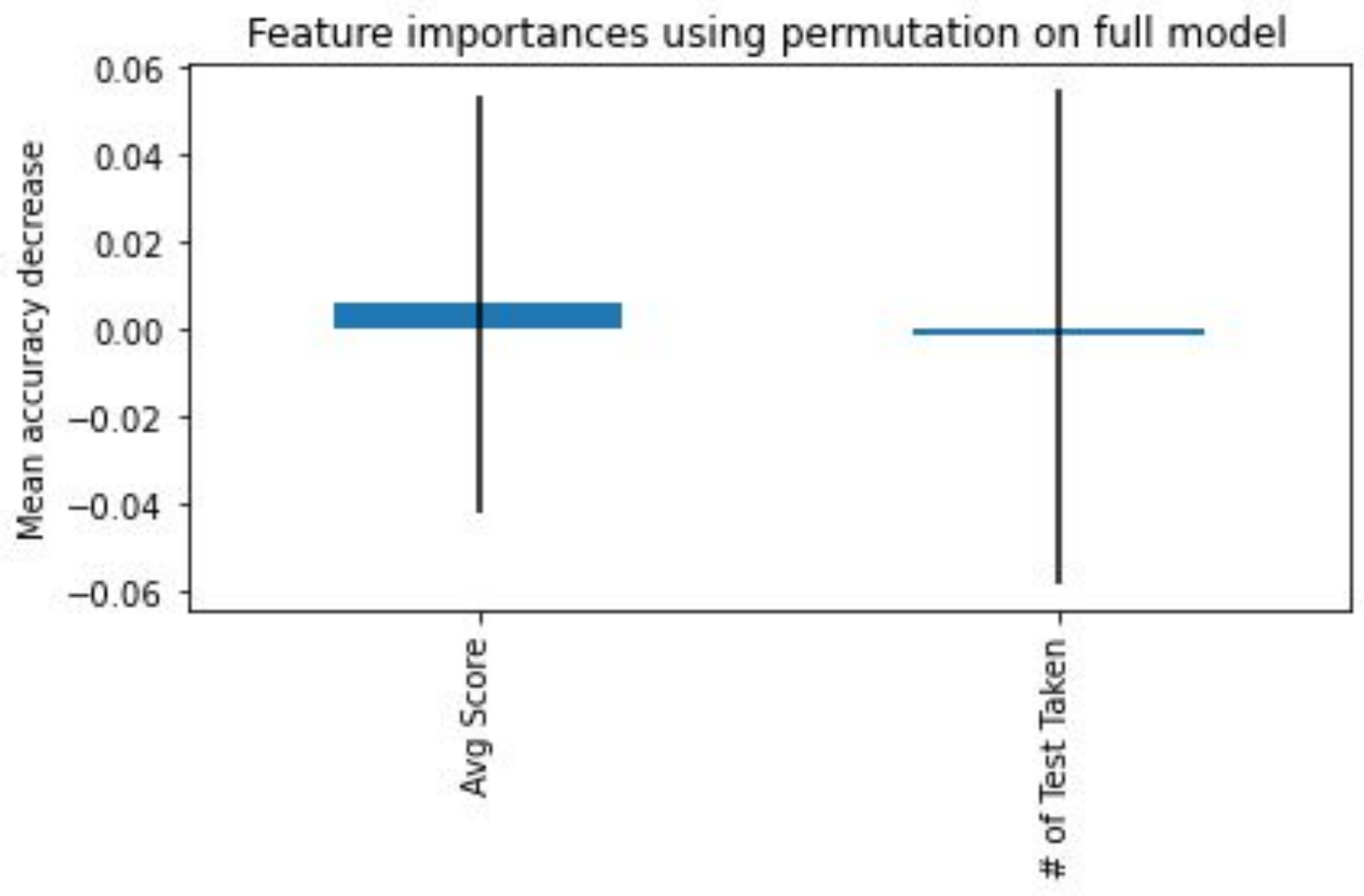
Fig 3. Visualization of the data collected in the survey for number of tests taken. There is a left skew in the graph indicating that of the students who took the survey, they were more likely to take less than 15 AP tests.



Explained

True Positive **Predicted GPA > 3.5** **Actual GPA > 3.5** **31 predicted : 54 test**	**False Positive** **Predicted GPA >3.5** **Actual GPA < 3.5** **11 predicted : 54 test**
False Negative **Predicted GPA < 3.5** **Actual GPA > 3.5** **8 predicted : 54 test**	**True Negative** **Predicted GPA < 3.5,** **Actual GPA < 3.5** **4 predicted : 54 test**

Fig 4. A confusion matrix showing the predictions of a student's GPA in college when given their average AP test scores and the number of tests they took. How to read the confusion matrix is displayed on the right.



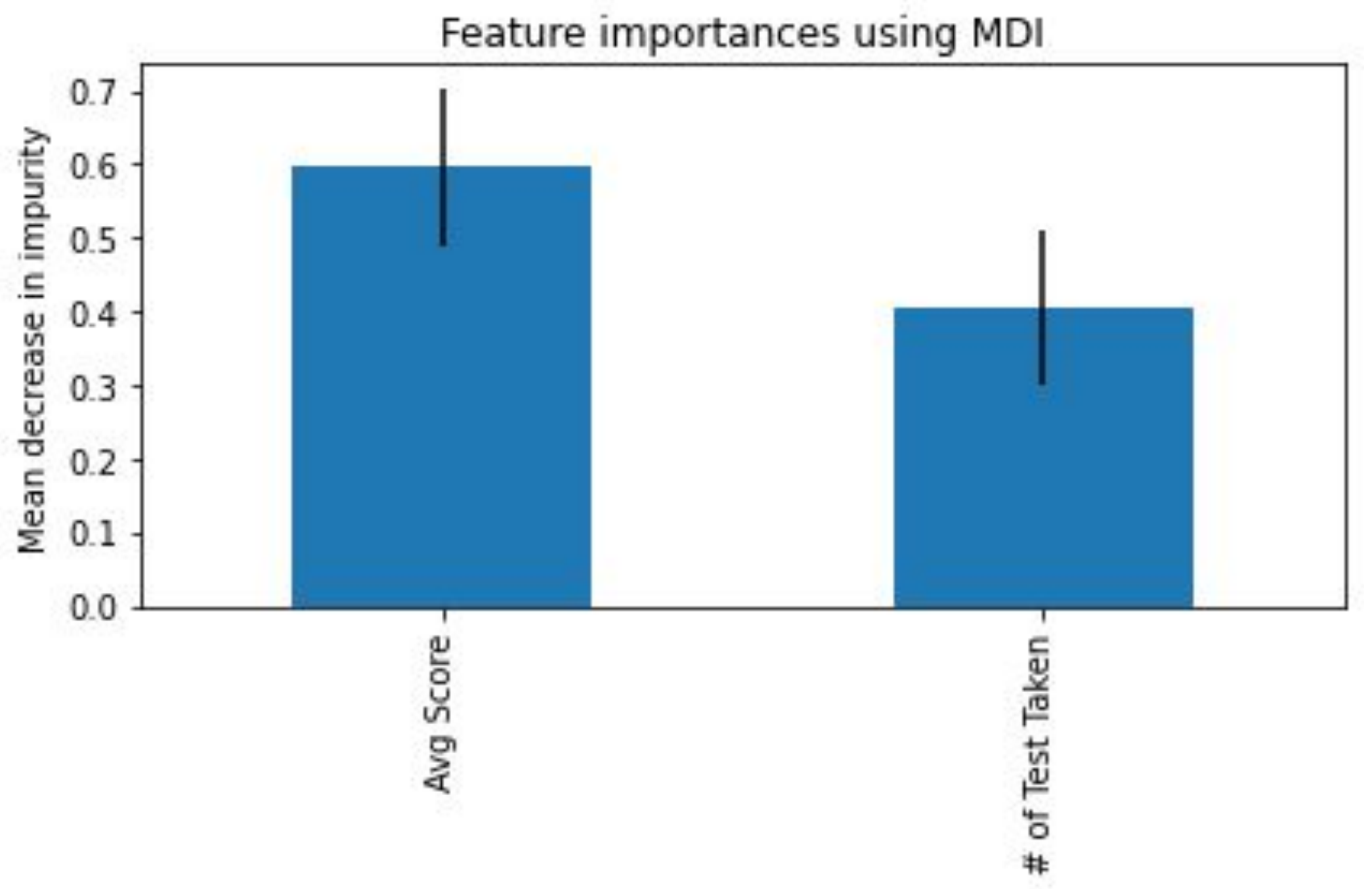


Fig 5. Graph showing feature of importance using the permutation method. In this graph, there is more error when the average score is shuffled. Thus, the average score is the "important feature".

Fig 6. Graph show the feature of importance using the MDI method. In this graph, there is a higher chance of reaching an accurate decision with the average score over the number of tests taken. So, the average score is the "important feature".

Conclusion and Future Work

The above tests yielded a 64.8% confidence rate, and running more tests yielded confidence rates between 55-70%. So, it can be concluded that AP test scores are a *moderate* indicator for a student's performance in college. However, since there are skews in the data, further tests would need to be run to get a holistic answer to the original problem. If this project were to be recreated, more data should be recorded from a broader audience.

References

[1] CollegeBoard (2023, January 1). *AP Program*. Retrieved February 8, 2023, from https://ap.collegeboard.org/

[2] Pedegrosa, F., Varoquaux, G., & E. A. (n.d.). *Scikit-learn*. Scikit-Learn. Retrieved February 8, 2023, from https://scikit-learn.org/stable/about.html#citing-scikit-learn

[3] Menze, B.H., Kelm, B.M., Masuch, R. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10, 213 (2009). https://doi.org/10.1186/1471-2105-10-213

Acknowledgements

I would like to thank Dr. Lan Yang for her guidance throughout this project. I appreciate her insight into this field as well as her support through each step to complete this project. I would also like to thank Dr. Sander Eller, Dr. Rebecca Franko, and Jack Beckley for supervising my survey during my data collection phase. Finally, I would like to thank Dr. Manuchehr Aminian and Dr. Lily Silverstein for introducing me to the world of research and the math behind the algorithms I create.