Amandaliss Dropik

CSCI 4502

Professor Ravi Starzl

12 December 2023

Final Project Report

For this project, I would like to figure out which gender demographic performs better in university and why. Overall, there are endless kinds of people in university around the world, and each will have their own methods for succeeding in school. I am curious to see if there are any correlations between gender and possibly more successful studying habits that might present in each group. There are also countless factors to account for when looking at how someone might perform in university, like whether someone works or not, time dedicated to hobbies, marital status, parental status and more.

Regardless of gender, we will learn what spells for success in university, and using that information, we can help whichever group learn these habits. University organizations would benefit from this because they could use this information to improve their processes with things like admission, class structures, and more. As a female in university, I am interested in seeing how my peers on average do and what might be done to help better myself (or my friends). There is also something to be said about the social implications that can be found here. For example, if it is found that women are more successful in university, why are they overall less “successful” in the workplace? There are many ways to cross-examine this problem.

There have been many studies on the differences between males and females in academics, with many different conclusions as well. This is because there are many sectors where one gender might do better than the other. However, it seems there is an overall consensus that women tend to be more successful in general in university. We see this from graduation rates, drop out rates, workforce statistics and more. There is a lot of work being done using ML to find out why one demographic is more likely to succeed than another, and how we can use that to implement policies to help the other. The main piece of literature I am referencing is *Conscientiousness as a Predictor of the Gender Gap in Academic Achievement* by Verbree et al (2023). According to Verbree et al., women are found to be more successful in postsecondary education because of their conscientiousness. Conscientiousness "includes being hardworking, reliable, organized, ambitious, self-disciplined, and persevering"(pg. 2), and these are major qualities that impact achievement which therefore account for some of the differences between males and females. They found that this is especially true for female students with non-dominant ethnic backgrounds, meaning "students who were born or with at least one parent born in a non-Western country"(pg. 2). From other resources, many say that there is a gap favoring men when it comes to STEM majors (Vooren, 2022). The fact that performance is something that is impacted by many different things makes it somewhat hard to narrow down. The current state of the field is a bit mixed, with arguments for either gender being favored.

With this project, I intended to find various data showcasing overall performance of students, study habits, etc. to find which demographic was typically more successful and why. I did this by using 3 models: random forest regressor, random forest classifier and linear regression. The random forest regressor was my main and most successful model, albeit not hugely successful. I wanted to use these models to see if I was able to predict grades based upon the features given in the model, specifically sex and another like attendance. My main dataset is *Student Performance* by Joakim Arvidsson (https://www.kaggle.com/datasets/joebeachcapital/students-performance/data), where the data collected includes:

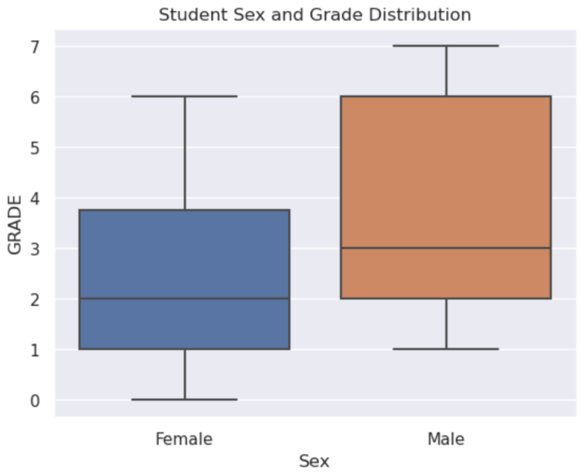
- Age
- Sex
- High-school graduation type (meaning private, public or other)
- Scholarship type
- Employment Status
- Total salary (if available)
- School transportation type
- Accommodation type
- Mother's education
- Father's education
- Scientific reading frequency
- Attendance to seminars/conferences
- Project/activities impact on success
- Attendance to classes
- Who Students Studied With
- When Students Studied for Exams
- Taking notes in class
- Listening in class
- Discussion improves my interest and success in the course
- Flip-classroom
- Cumulative grade point average in the last semester (/4.00)
- Expected Cumulative grade point average in the graduation (/4.00)
- Course ID
- OUTPUT Grade

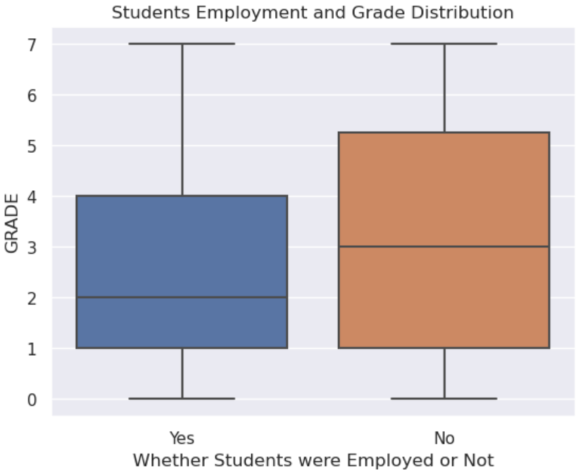
This dataset includes lots of questions about personal information, family, and education habits. It also included grades for every student, which was the output or target feature of this dataset. This dataset is very clean, and at my first impression, I did not have too much preprocessing. At 145 rows, I realized that I would become somewhat limited in what I could perform with this data. I tried to find more datasets to integrate, but I was unable to find some that were as thorough as this one. Many were too simple and lacked the educational habits, which are my main focus here.

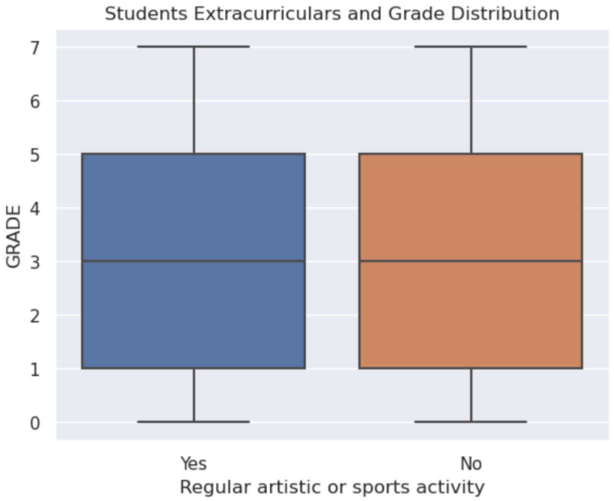
I started with a general exploratory analysis to look at what I was working with. Univariate and bivariate analysis helped me create some initial ideas and impressions. Here is a list of questions I came up with according to my data. The answers to these questions are explored later on in this report:

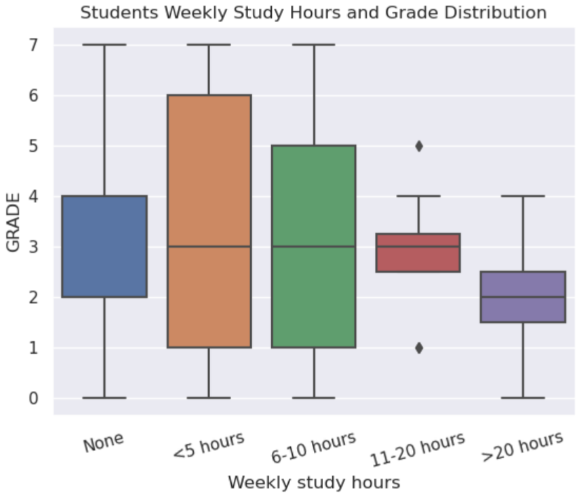
1. Which sex performs better in school?
2. Does having a commitment outside of school negatively impact your grade?
3. How impactful is cramming for exams on your grade?
4. Is it better to study more than 5 hours a week?
5. Does attending seminars/conferences related to your department impact your grade?
6. Is attending class necessary for a good grade
7. Does studying with classmates negatively or positively impact your grade?
8. How much does note-taking during class determine your success?
9. How much does listening during class determine your success?

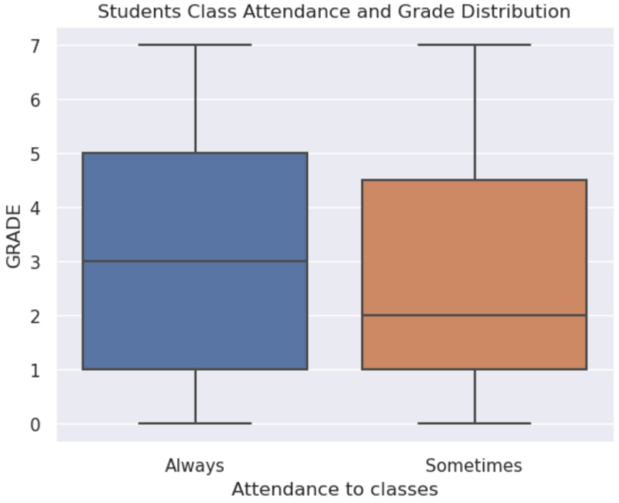
Continuing with the analysis, I found that there were a bit more males than females in the dataset, so that would already impact my goal. I also found that the grade distribution is right skewed, meaning there are more lower grades than higher grades, particularly in the DD category. With box plots, I was able to look at what features I thought would lead to better (or worse) grades. I found that in this dataset, males are doing much better in their classes. Below, I will insert the most significant graphs found

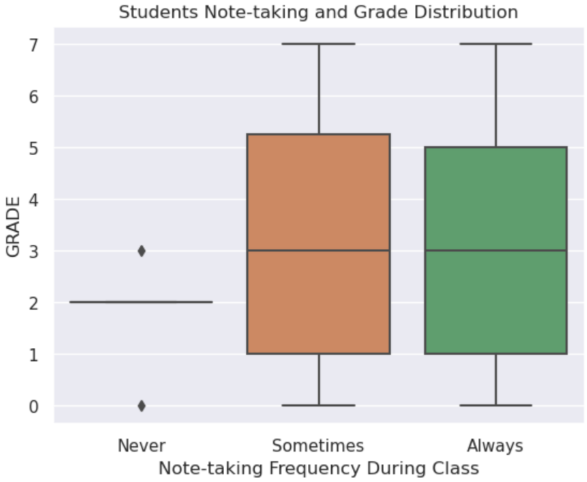

Student Sex and Grade Distribution

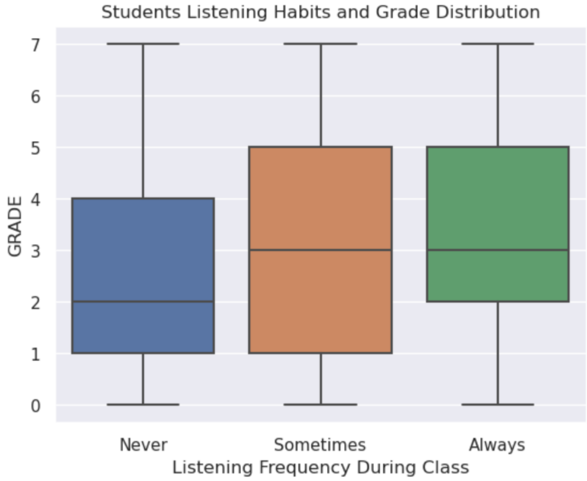

Students Employment and Grade Distribution

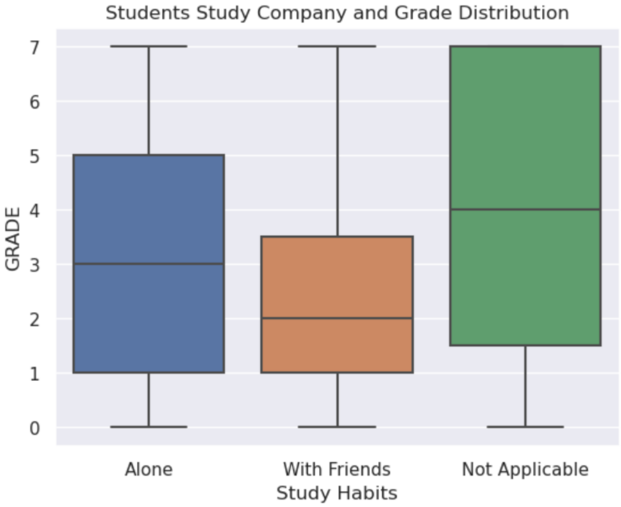

Students Extracurriculars and Grade Distribution

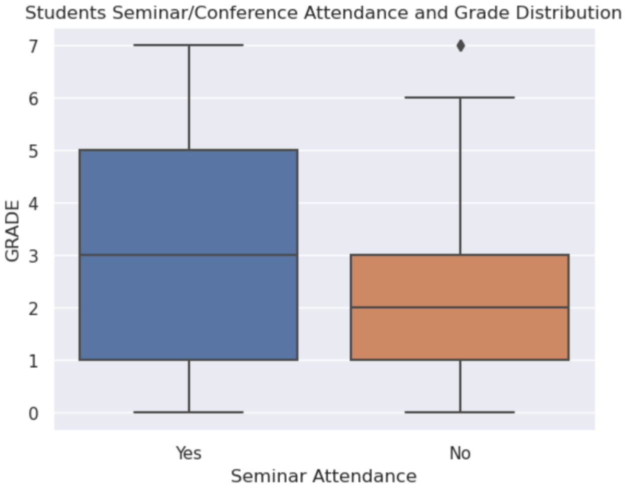

Students Weekly Study Hours and Grade Distribution

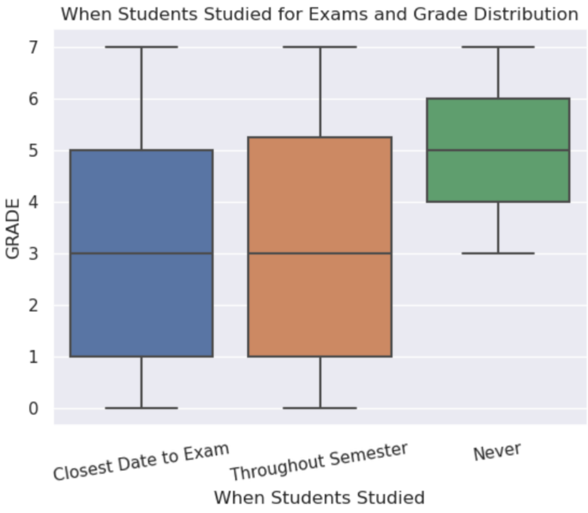

Students Class Attendance and Grade Distribution


Students Note-taking and Grade Distribution



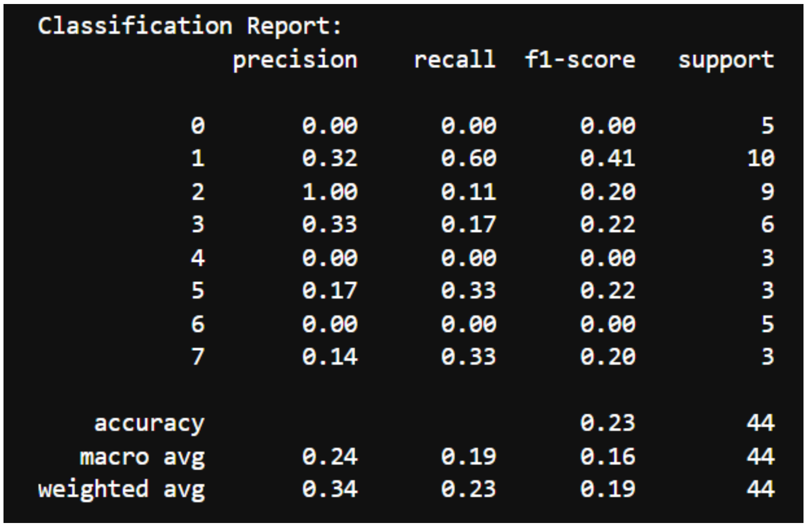






As stated above, I tried 3 different algorithms: Random Forest Regressor, Random Forest Classifier and Linear Regression. I went with the random forest models because they are known to have high accuracy and predictive power. They are also perfect for my goal- to understand the relationship between the features and the target, like what influence they might have. Starting with my random forest regressor, my root mean squared error was about 2.057 (2.057179366323528). This is the best result I was able to achieve after trying things like changing the test size and the number of trees in the forest. This is saying that my model is predicting about 2 grades off. The features with the highest Pearson’s correlation are sex, at 0.3355, cumulative grade point average in the last semester, at 0.3155 and expected cumulative grade point average in the graduation at 0.2486. The features with the highest feature importance score are cumulative grade point average in the last semester at 0.1146, father’s education at 0.0683 and sex at 0.0657. As we can see, although the cumulative GPA has the highest feature importance score, it is still not that strong of a predictor. However, there are moderate correlations between sex and cumulative GPA with grades.

Next, I tried the random forest classifier model since I have multiple classifiers, those being the scale of grades. This model also has a very low accuracy rating of 23% and with the classification report, seems to perform best with class 1, which is the DD grade. This is probably because this class has the most data compared to the rest.

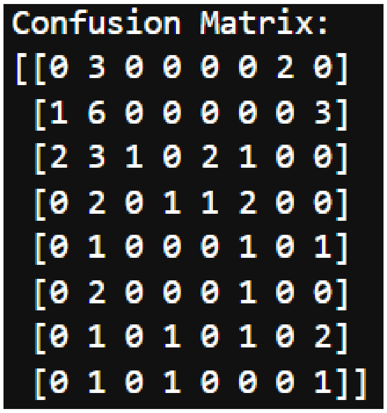


```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         5
           1       0.32      0.60      0.41        10
           2       1.00      0.11      0.20         9
           3       0.33      0.17      0.22         6
           4       0.00      0.00      0.00         3
           5       0.17      0.33      0.22         3
           6       0.00      0.00      0.00         5
           7       0.14      0.33      0.20         3

    accuracy                           0.23        44
   macro avg       0.24      0.19      0.16        44
weighted avg       0.34      0.23      0.19        44
```

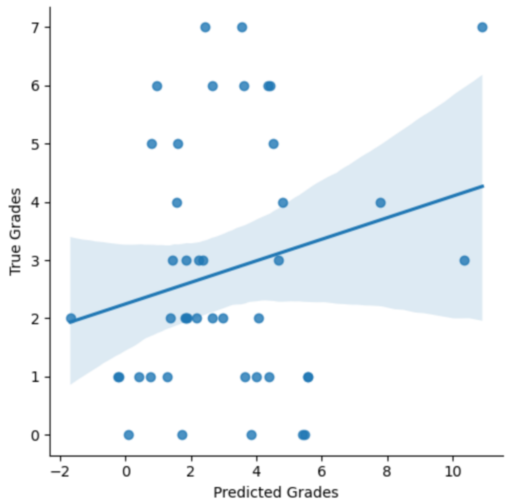
For class 1, the precision is 0.32, indicating that out of all instances predicted as class 1, 32% were actually class 1. Its recall is 0.60, meaning that the model captured 60% of all actual class 1 instances. The F1-score is 0.41, reflecting a trade-off between precision and recall. For the feature importance scores, after looking at every feature, I selected the ones that would seem to help the model the most. That included sex, cumulative grade point average in the last semester, expected cumulative grade point average in the graduation, reading frequency, attendance to classes and weekly study hours. The values ranged between 0.0862 and 0.2642. Although that was done, it only helped a little. In the end, the data will need some feature engineering to be done.

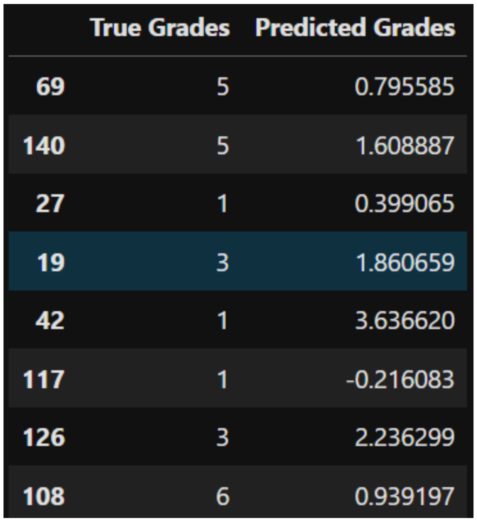


```
Confusion Matrix:
[[0 3 0 0 0 0 2 0]
 [1 6 0 0 0 0 0 3]
 [2 3 1 0 2 1 0 0]
 [0 2 0 1 1 2 0 0]
 [0 1 0 0 0 1 0 1]
 [0 2 0 0 0 1 0 0]
 [0 1 0 1 0 1 0 2]
 [0 1 0 1 0 0 0 1]]
```

The confusion matrix is telling us that every class has 1 or no true positives and class 1 has the most true negatives. Again, the matrix is showing us that the model is struggling to perform with all of the classes.

Finally, I tried a linear regression model just to see what I would get, because at this point I had figured that my Random Forest Regressor was going to be my best model. With the linear regression, my RMSE score indicated that the model is predicting about 3 grades off. The negative r-squared is telling us that the model is a poor fit for the data, which is what I was expecting. I’ve included some visual representations of the model and a sample of its predictions. As we can see, it is not doing well.





	True Grades	Predicted Grades
69	5	0.795585
140	5	1.608887
27	1	0.399065
19	3	1.860659
42	1	3.636620
117	1	-0.216083
126	3	2.236299
108	6	0.939197

To summarize my results, I will start with the initial exploratory analysis. My findings at the beginning had led me to believe that there were a handful of features that would impact the target. I thought I would have many good features to pick. The features that stuck out across all the models was sex and cumulative GPA, and that was really it. From the graphs, I came up with these answers to the questions formed in the early exploratory analysis:

1. This data set suggests males do better
2. There is only a slight impact if you are employed, and none with extracurriculars
3. Seems to have no significant impact on grade
4. Studying for longer than 5 hours appears to lower grade
5. Attending seminars does positively impact grade
6. This data set suggest there is no significance towards grade
7. Studying with others has a moderate significance
8. Seems to positively impact grade
9. Might slightly impact grade

The random forest regressor had the most success, but not by much. Although many reports about this topic indicate that women have better success in university, this dataset suggested the opposite. It also seemed to suggest that many study habits that are known to improve academic performance have no impact. Adding to my initial impression, the number of features I thought would make for a good model was wrong. Many of them are not substantial enough to even need in the model, nor are they really as important in the general survey as well.

In the end, at a surface view, all that was needed was removing some unnecessary columns when it came time to train the model. However, after working with the models, many of the features are not substantial enough to even use in the model, nor are they really as important in the general survey as well. Having fewer features and more rows of data would have helped greatly.

Sources

Verbree, A. R., Hornstra, L., Maas, L., & Wijngaards-de Meij, L. (2023). Conscientiousness as a Predictor of the Gender Gap in Academic Achievement. *Research in higher education*, *64*(3), 451–472. https://doi.org/10.1007/s11162-022-09716-5

Vooren, M., Haelermans, C., Groot, W. *et al.* Comparing success of female students to their male counterparts in the STEM fields: an empirical analysis from enrollment until graduation using longitudinal register data. *IJ STEM Ed* 9, 1 (2022). https://doi.org/10.1186/s40594-021-00318-8