**Grade Analysis & Predictions**

**Team Undecided**

Anthony Tran, James Nguyen, Martin Nguyen

CS 122

November 30, 2021

**I. Introduction**

As college students, we’ve also wondered what variables play a part of how we academically perform as we’re always striving to improve ourselves. As a result, for this term project, we want to explore what affects a student’s academic performance, and whether we can utilize these features to predict a student's performance.

We will be utilizing the dataset called “Student Alcohol Consumption” by UCI Machine Learning (<https://www.kaggle.com/uciml/student-alcohol-consumption>), which contains data obtained from a survey for students who took math and portuguese courses at a secondary school. The dataset includes different attributes of the student such as their gender or study time and the grades they received in the classes. Our project will perform an analysis of the dataset to draw conclusions about the effect of a student's personal or social life on their academic performance. We will then create three machine learning models to predict the final grades of these students. Finally, we will analyze and decide which machine learning model gives the best prediction results.

**II. Methods**

Our project consists of several distinct complex components which all work together. These components includes the following:

* loading and setting up the dataset, and exploring the dataset
* analyzing and exploring the data regarding a person’s home life and drawing conclusions about the variable’s relationship with academic performance
* analyzing and exploring the data regarding a person’s social life and drawing conclusions about the variable’s relationship with academic performance
* building three different machine learning models to predict the final grades of students and determine which Model gives the best prediction results.

This section will go more in depth about each component of our project and how they all connect with one another.

**A. Loading & Setting Up the Dataset**

To get the project set up, we first downloaded the .csv files from the kaggle dataset, and uploaded them to a github repository. After that, I used the link to the raw file to access the dataset in the notebook (similar to how Professor Lee does it). After that, we used Pandas and Seaborn functions to make analysis more straightforward. For instance, we changed all the integer data types to floats, created an extra column for total grade out of 100 (since grades were split up into 3 categories), and drew out a heatmap that showed correlations between the different factors. Our main objective was to find which factors that had a correlation with grades, which were hard to come by.

**B. Analyzing Personal versus Academic Performance**

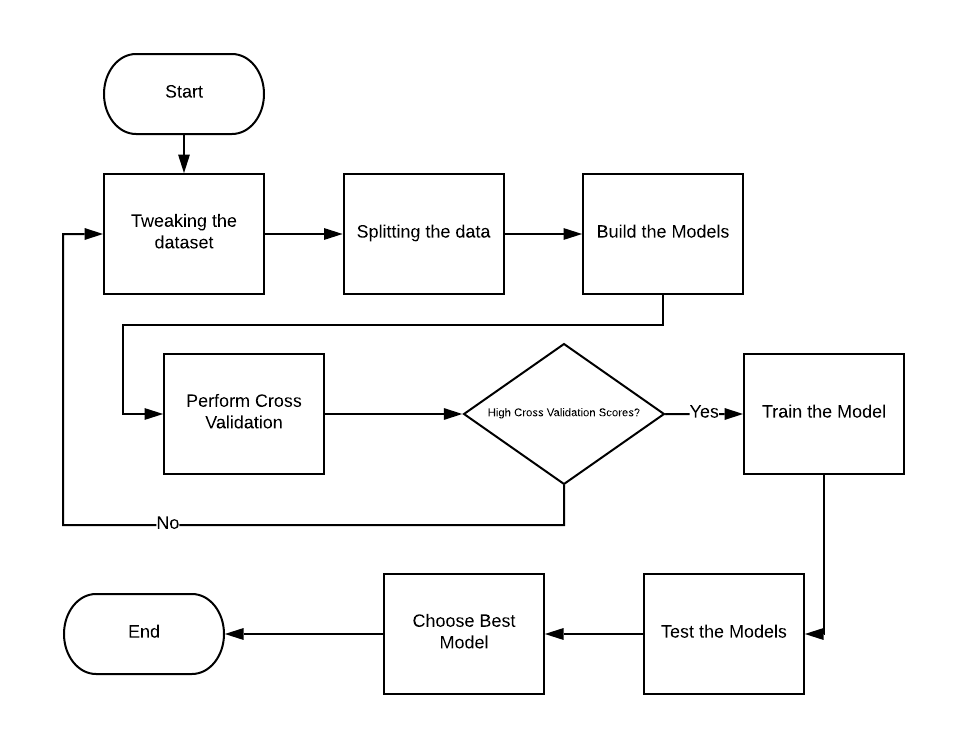
To analyze personal factors on academic performance, we looked at factors such as study time, free time, family size, and education levels of parents and guardians. Because we previously created a heatmap describing the correlations, the objective of this section was to visualize the correlations between certain factors and academic performance. For instance, one would think that studying more would be strongly correlated with higher grades. In this section, we created a violin plot to show that the correlation between study time and grades is actually pretty minor (0.13) using the ‘totalGrade’ column we previously created. Similar to this example, we used bar/violin plots to illustrate a relationship between a factor and the different grades; we would use G1, G2, G3, and/or totalGrade depending on what we were trying to show. For two other plots, we would use a FacetGrid to show the grade distribution across different factors, using the following: col = ‘some factor.’ Using these plots, we could better understand how little of a correlation most of the factors had with grades.

**C. Analyzing Social versus Academic Performance**

Analyzing the social factors on academic performance had us looking at attributes like daily/weekly alcohol consumption, going out time, study time, etc. The main purpose of this was to visualize the correlations that were shown from the heatmap of the dataset and use that as a base to explore other possible correlations that might have been missed. The process with which we visualized these correlations was to first visualize just the correlations against each other. An example of this would be Dalc vs Walc, where we would first plot the two categories against each other in a bar plot. This allowed for seeing and understanding the correlation better. From there we would plot them against the final grades and try to use our previous plots to interpret the correlation between the final grades and that specific social attribute.

**D. Building & Training Three Predictive Models**

With the data analysis completed, we can move on to building three different predictive models which will be able to forecast the final grades for students based on the input variables in the dataframe. In order to create these different predictive models we will have to go through several different stages. These different stages can be explained with the following diagram:



As seen in the diagram, we will begin by tweaking the dataset in order so the models can read the input data. First, we will split the dataset into input and output variables. We will put the final grade as the output variable, but the input variables are more complicated. Based on the previous observations in the other sections, we noticed that there are very low correlations between most of the variables and the final grades except for the first and second period grades. This gives us the dilemma of whether we include our G1 and G2 scores in the input variables or not. Including G1 and G2 scores may cause high model variance. This is when the model will fit the training data well, but does not perform well with test data as the input data may be too specific. On the other hand, not including G1 and G2 may result in low performance from all our models.

As a result, we will use a technique called cross validation on our models using the input data which does not contain the first and second period grades. Then, based on the scores, we will either decide to add the grades into the input variables, or continue and train the model.

Now that we have split the input and output variables, we will need to convert the categorical input variables to numerical input variables. This is due to the fact that since we are predicting the final grades of students which are continuous numerical values, we will be using regression models which can only take numerical inputs.

Once we have the input and output variables set up, we will then split these variables into testing and training data. The training data will be used to train and validate the models, while testing will be used to test the performance of our models once the model has been finalized. We will split our variables into 80% training and 20% testing data.

Once our data is split, we will create three different models that we will run cross validation on. The models are all part of the sklearn library but each behave differently. The three models that we use are:

* **Linear Regression:** Model which finds a linear function that best fits our input and output variables. This function will be used to determine the final grades.
* **Decision Tree Regressor:** Model which creates a continuous variable decision tree based on our training data that will be used to decide the final grade based on the input variables
* **Support Vector Regression:** Model which finds a hyperplane that fits the most number of points in our training data. The goal of SVR is to minimize the amount of error but also maximize the margin of the hyperplane.

With the models created, we will then compute the cross validation scores to give us insight on how well our model is doing and whether we should continue with the input data or not. Cross validation works by breaking our training data into X parts, and then for each part, the model will be trained by all the data in the training data except for that specific part. That one part will then be used to validate the trained model and give a score based on how accurate the model is. We will use a 5-fold cross validation which means we will break our training data into 5 parts.

Based on the cross validation scores of the different models, we will then decide to continue or go back to tweaking the data. If the scores are good, we will continue with training and testing the model where we will train the model with the whole training data and evaluate the models with our testing data. We will use the following metrics to evaluate our model’s performance with our testing data:

* **Coefficient of Determination:** This metric shows the percentage of how closed the predicted data fits to the true values.
* **Mean Squared Error:** This metric exemplifies the average of the squares of the differences between the predicted and true values. The metric gives large penalizations to errors since it takes the squares of the errors.
* **Mean Absolute Error:** This metric is similar to mean squared error but instead of the square of the difference, the metric takes the absolute value of the difference. Unlike Mean Squared Error, the metric does not give large penalizations and treats all errors as the same.

With the scores computed, we can then determine which model performs the best in predicting student grades based on these different metrics.

**III. Results**

After completing the project, we were able to make several conclusions surrounding the dataset as well as figure out which predictive model performs the best on our dataset. This section will go over these results that were found through the different project components.

**A. Loading & Setting Up the Dataset**

All columns of int data type were changed to floats. Then I used basic math to make a totalGrade column, which adds G1, G2, and G3 grades scaled out of 100. Then we added statistical descriptions for each numerical column and a correlation heatmap. Lastly, we added plots that illustrate the correlations between G1 and G2, G2 and G3, and G1 and G3.

**B. Analyzing Personal versus Academic Performance**

For analysis of personal factors vs. academic performance, here were some of the important correlations (or lack thereof).

From the violin plots, we can see that time spent studying had a small positive correlation with grades. This was true across G1, G2, and G3 grades, and so we plotted the totalGrade for convenience and readability purposes. **(See ‘Effect of studying on grades’)**

Using a swarm plot, we can see that the majority of students believe that they have "average" free time, and the grades do not correlate with a lack or abundance of free time. We used a hue for the internet factor just to loosely show that it’s much harder to have better grades without an internet connection at home. **(See ‘Free time vs Grades’)**

From these plots, we can see that smaller family sizes performed better on average, but this is most likely due to less data points. **(See ‘Family Size vs G1, G2, G3’)**

A majority of students said their mother was their guardian, which did not affect academic performance. **(See ‘Guardian vs G1, G2, G3’)**

During the study, one of the questions was "Why did you choose this school?" The plots above show the different reasons, and how the grades are correlated. The only small conclusion we can draw is that those who chose reputation as their reason for picking the school tended to perform better on average than those who had other reasons. **(See ‘Reasons for choosing school vs Grades’)**

From the first plot **(‘Fedu vs Medu’)**, we can see that parents tend to have the same level of education. From the next two plots **(‘Fedu, Medu vs G3’)**,we can see the following: while there is a positive correlation between the education of one parent in relation to the other. There is absolutely no correlation between the parent's education and the student's grades.

From the plots **(‘Medu vs G1, G2, G3’)**, we can see that students who come from families where the mother is either highly educated or not educated at all tend to have better grades.

Looking at the boxplots concerning failures **(‘Medu, Fedu vs failures’)** and parental education **(‘Failures vs G3’)**, a parent's education level seems to affect the amount of failures the students have. Also, we can see that failures affect the grades of a student inversely. So we can say that a parent's education level indirectly affects the grades of a student.

**C. Analyzing Social versus Academic Performance**

For the results of analyzing social versus academic performance, we were able to find various correlations between the different social factors.

From the **Dalc vs Walc** plot, we can clearly see that students who drink more on the weekdays end up drinking more on the weekend and vice versa. I believe that this is likely because if they’re drinking during the weekdays, it becomes a habit and that carries over to the weekends.

For **Dalc/Walc vs G3**, it seems like lower alcohol consumption won’t guarantee a good grade, but higher levels of alcohol consumption can definitely prevent achieving the highest grades.

The **goout vs Dalc/Walc** charts shows how students who go out often end up drinking more alcohol. This could potentially be because students who go out often are looking to escape the stress of school and alcohol is probably one of the best ways to do this.

From the **goout vs G3** chart it seems like going out a moderate amount leads to better grades. When a student goes out too often or doesn’t go out at all, grades seem to be lower.

The **freetime vs goout** chart shows a positive correlation between the two meaning that when students have freetime, they often go out and hang out with friends.

On the other hand when we see the **freetime vs studytime** plot we can see that the more freetime that someone has the less studytime they have.

Because of that when we compare **Dalc/Walc vs studytime**, we see the same thing as the previous chart where the more alcohol one consumes the less studying they do.

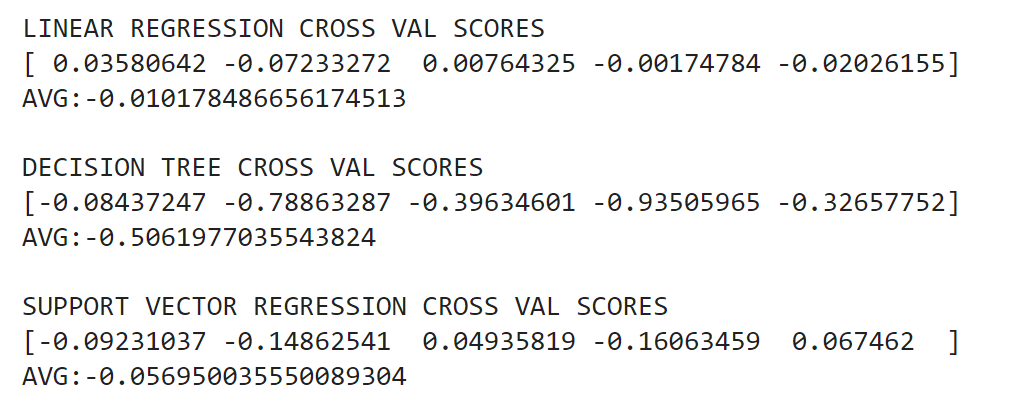
This is shown to be counterproductive as in the **studytime vs G3** plot, the students who have the most studytime end up with the highest possible scores.

The next chart had us looking at **freetime vs absences** which showed that students with the least freetime had the highest amount of absences by far. This is important as when comparing **freetime vs G3**, it explains why students with the 2nd lowest amount of freetime had the highest grades. It was these students who had a low amount of absences and didn’t go out often meaning they had more time to study.

The final attribute we looked at was extracurricular activities. When comparing **activities vs Dalc/Walc**, we noticed that students who participated in extracurricular activities drank less alcohol. This matched up with the **activities vs G3** plot as students who participated in extracurricular activities had the higher final grades.

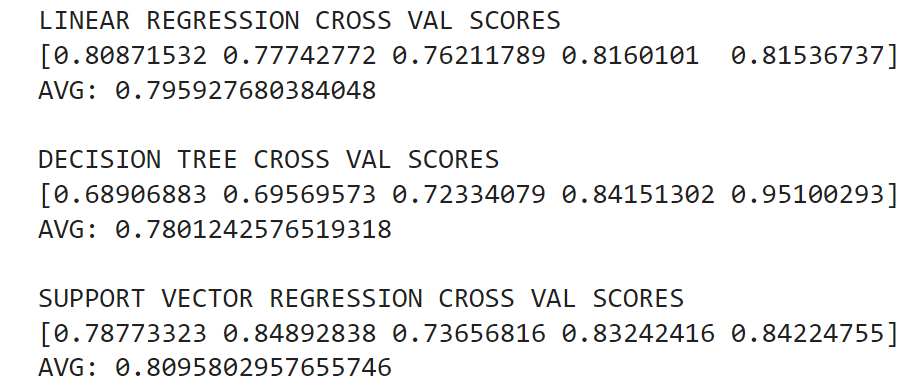
**D. Predictive Modeling**

After completing this component, we found that training the model without including G1 and G2 scores caused extremely low cross validation score as seen below:



This could be due to a bad dataset or the models being incorrectly configured. However, all models seemed to be configured correctly. Therefore, this is due to the low correlations of the variables as expected.

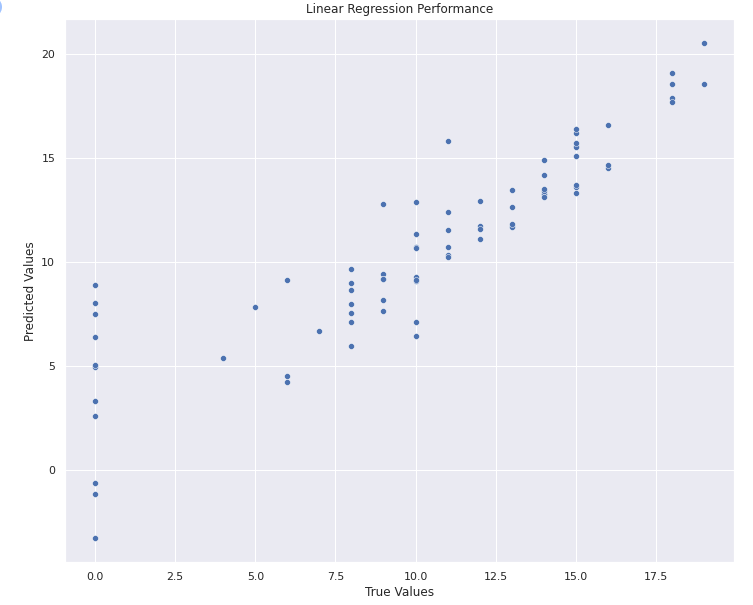
Due to these low correlation scores, we decided to go back to tweaking the dataset where we included the first and second period scores in the input dataset this time. This gave us increasingly better scores than before.

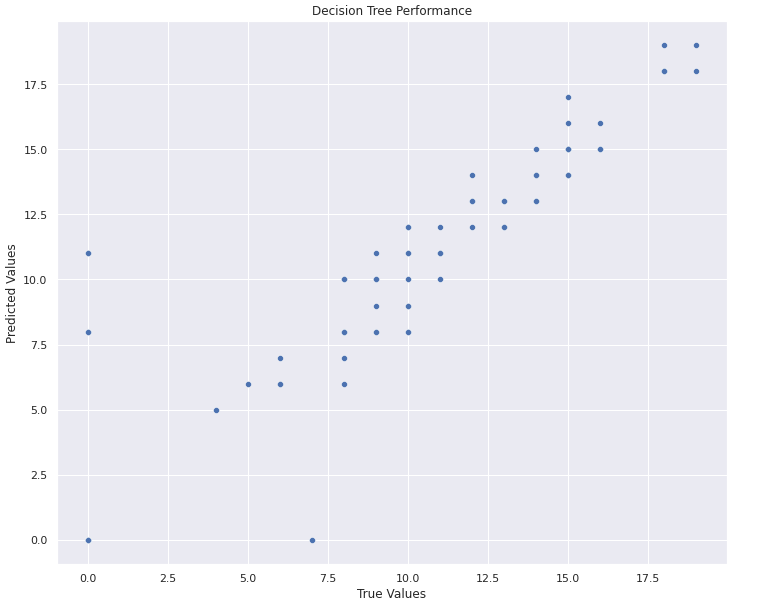


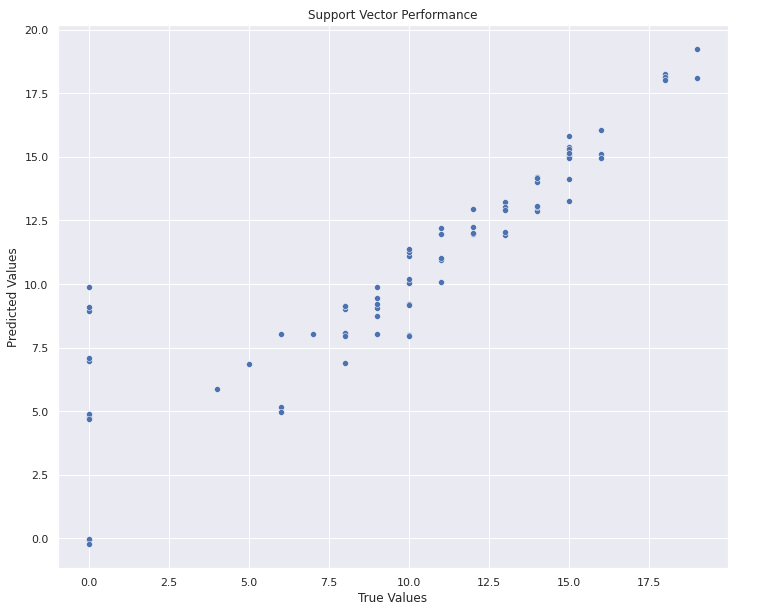
With the high cross validation scores, we were then able to train the models based on the new input variables. We then inputted the testing data into our models and received the following performance results for each model:

| **Model** | **Coefficient of Determination** | **Mean Squared Error** | **Mean Absolute Error** |
| --- | --- | --- | --- |
| **Linear Regression** | 0.7881 | 5.845 | 1.610 |
| **Decision Tree Regressor** | 0.776 | 6.19 | 1.203 |
| **Support Vector Regression** | 0.779 | 6.087 | 1.203 |

The results can also be visualized using the following graphs where the x-axis represents the true values while the y-axis shows the predicted values:







Based on these performances we can conclude that although all models perform very well, the Linear Regression Model seems to be performing a little better than the other models. The model performances on our testing data showed that the Linear Regression model had the highest R squared score which means that it was the most accurate. Furthermore, even though the Linear Regression Model did have the highest absolute error, the model’s errors were not as severe as the other model’s errors which can be seen in the models low mean squared error.

**IV. Discussion**

Throughout our project, we faced several different technical challenges and there were many aspects that we could’ve improved on. This section will go over the different challenges and improvements.

**A. Technical Challenges**

One of the challenges of the project was figuring out how to visualize the data. Most of the data was either binary (M/F, yes/no, etc) or based on a likert scale (1-5) so it really limited the options we had for visualizing the data. This shows in our project as many of our plots were either bar plots, violin plots, or box plots.

Another technical challenge that we faced was that a majority of us didn’t know much about Machine Learning. There was a lot of research, self learning, and knowledge transfer that needed to happen in order for us to fully understand and grasp the different concepts of Machine Learning such as building and training a model. It was definitely a steep learning curve as we not only had to learn about machine learning techniques but also we had to brush up on our statistics.

Another challenge we faced in this project was finding anything that actually had an effect on student grades. A lot of the factors in the dataset were interesting, but just didn’t have a strong correlation with student grade which made it extremely difficult for us to create interesting plots. This also impacted the models as we were having trouble getting high cross validation scores for any of our models since there was no strong correlation.

**B. Improvements**

An improvement that we would do is to choose a different dataset all together. While the dataset contains many relevant data points, the sample size of them is too small. Along with this another major problem with the dataset is that it was solely collected from students in Portugal. The geographic location could be a huge factor that drastically changes the relationship of the other data points. An example of this would be the daily and weekly alcohol consumption. The drinking age in Portugal is only 16 so it’s probably quite common for highschool students to drink, but if this were taken from America where the drinking age is 18, alcoholic consumption would likely be much lower.

Now if we were to improve this project while keeping the current dataset, we would probably add in the Portuguese class dataset along with the current math class dataset. Not only would this increase the sample size, but it would also increase the possible relations between them. For example, maybe the grades of students in the Portuguese class goes up the more they go out compared to the math class. This could be because students who go out often end up communicating with others in portuguese which helps them understand the language better and do better in class. Or maybe it’s because student’s in math need to spend more time understanding the concepts and going out more takes time away that they could be using on practicing and comprehending the material.

**V. Appendix**

The following section will go over the instructions to run through the program, our input data sources, and any additional data.

**A. Instructions to Run Program**

To run the program, follow the below instructions. You will not need to worry about any of the dependencies missing as Google Collab, the tool we will be using, has pandas and sklearn already installed.

1. Download the file, term-project.zip.

2. Unzip the folder. It should have the following file structure

* term-project folder
  + README.pdf
  + Term Project.ipynb

The README.pdf will contain information about how to run the program, and the Term Project is the program file that will be needed for the next few steps.

4. Navigate to <https://research.google.com/colaboratory/> on either Chrome, Firefox, or Safari as these are the browsers currently supported.

5. Sign into your Google Account. If you don’t have one, create an account.

6. On the website, go to the tab called “Upload”.

7. Click on “Choose File”. This will open a file selection window.

8. Click on the file Term Project.ipynb and press open. You should be navigated to the notebook site.

9. Now you can scroll through the notebook and see the results of a previous run.

10. To run the program again, you can click each cell in order one by one, or you can go to the Runtime tab, and then press “Run all” which will run all the cells for you.

**B. Input Data Source**

UCI Machine Learning. (2016, October). Student Alcohol Consumption, Version 2. Retrieved November 27, 2021 from <https://www.kaggle.com/uciml/student-alcohol-consumption>