Paige Ngo
Professor Lozinski
C111
8 December 2023

Student Grade Prediction Based on Different Variables

For this final project, I applied machine learning techniques to investigate how different variables in students' lives impacted their performance in school. Below is my report.

Introduction:

In this final project, I explored machine learning to investigate the relationships between several variables in students' lives and their impacts on academic performance. One of the main objectives of this project was to discover patterns, relationships, and any valuable insights capable of highlighting any factors that affect the success of students. In addition to analyzing different factors, the project extends its purpose to prediction. Using different machine learning techniques, the project attempts to develop models capable of forecasting the grades of students. The intention was to understand how accurately different models can predict the results. Analyzing these educational variables not only contributes to a deeper understanding of academic outcomes but also aims at fostering student success.

To help solve this problem, a dataset called "Student Grade Prediction" from Kaggle was analyzed. Using coding in python, I was able to better understand the data using graphs and a machine learning approach. The dataset uses student grades and demographics from the secondary education of two Portuguese schools. The data was collected through school reports and questionnaires and analyzed grades for both mathematics and Portuguese language classes [1]. To better understand the data, I used supervised machine learning since I had labeled training data. More specifically, I used regression techniques since the purpose of this project is to forecast student grades.

To solve the problem, we used linear regression, Random Forest Regressor models, analyzed feature importances, analyzed REC curves for different models, and used a neural network to help predict final grades and to see which features impacted the final grade the most. We concluded that the final grades were highly related to the grades the students received earlier in the year and found that the machine learning models did not work nearly as well when we took those features out.

Data:

The dataset used in this project was found through Kaggle but originally comes from the UCI Machine Learning Repository. The dataset contains information about student performance in secondary education at two Portuguese schools. The data includes various features, including demographic, social, and school-related features. The dataset focused on the students' performance in the subjects of Mathematics and Portuguese language.

The dataset describes a range of attributes such as school, sex, age, familiy size, parent's cohabitation status, and education levels of both parents. Additionally, it covers details about the student's home address, mother's and father's jobs, reasons for choosing the school, primary guardian, travel time to scool, study time, extracurricular activities, and more. Not all attributes were analyzed in this project, but the preprocessing steps showcase which features were chosen for my models. The dataset also includes three grades related to the course subject: G1 (first period), G2 (second period), and G3 (final grade). The dataset mentions that G3 has a strong correlation with G2 and G1 [1].

Preprocessing Steps:

The first step in my project was loading the dataset into a Google Colab notebook. The pandas library was used to load the file as a dataframe object. I also used the pandas drop method to remove certain columns that were not used in our project such as: mother's job, father's job, nursery, wanting to take higher education, reason for attending school, romantic relationships, extra paid classes within the course subject, which school they attended, parent's cohabitation status, extra educational support, extracurricular activities, travel time to school, family educational support, student's guardian, and student's home address type. Instead, I focused on the categories: sex, age, family size (less than or greater than 3), mother's education, father's education, time spent studying, failures, access to internet, quality of family relationships, amount of freetime, how much students go out, workday alcohol consumption, weekend alcohol consumption, health, absences, G1 (first period), G2 (second period), and G3 (final grade). Additionally, I removed rows where the final grade (G3) were zero to prevent distortion in some of my models since it was creating poor performing models. Furthermore, I cleaned my data by replacing any NaN values with a mean value and changed any categorical variables like "famsize", "sex", and "internet" to numeric representations since strings weren't compatible with some machine learning models. Lastly, I used the train_test_split function from scikit-learn to split the dataset into training and testing sets with G3 being my target variable.

To better understand the data, I also used the pandas.describe() method to create a detailed chart of values that gave me insight for my data.

	sex	age	famsize	Medu	Fedu	studytime	failures	internet	famrel	freetime	goout
count	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000
mean	1.518207	16.655462	1.700280	2.795518	2.546218	2.042017	0.271709	1.162465	3.955182	3.246499	3.098039
std	0.500370	1.268262	0.458778	1.093999	1.084217	0.831895	0.671750	0.369395	0.885721	1.011601	1.090779
min	1.000000	15.000000	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000
25%	1.000000	16.000000	1.000000	2.000000	2.000000	1.000000	0.000000	1.000000	4.000000	3.000000	2.000000
50%	2.000000	17.000000	2.000000	3.000000	3.000000	2.000000	0.000000	1.000000	4.000000	3.000000	3.000000
75%	2.000000	18.000000	2.000000	4.000000	3.000000	2.000000	0.000000	1.000000	5.000000	4.000000	4.000000
max	2.000000	22.000000	2.000000	4.000000	4.000000	4.000000	3.000000	2.000000	5.000000	5.000000	5.000000

Dalc	Walc	health	absences	G1	G2	G3
357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000
1.495798	2.330532	3.549020	6.316527	11.268908	11.358543	11.523810
0.919886	1.294974	1.402638	8.187623	3.240450	3.147188	3.227797
1.000000	1.000000	1.000000	0.000000	3.000000	5.000000	4.000000
1.000000	1.000000	3.000000	2.000000	9.000000	9.000000	9.000000
1.000000	2.000000	4.000000	4.000000	11.000000	11.000000	11.000000
2.000000	3.000000	5.000000	8.000000	14.000000	14.000000	14.000000
5.000000	5.000000	5.000000	75.000000	19.000000	19.000000	20.000000

Figure 1: Chart from using pandas.describe(). [1]

Figure 1 gives us helpful data to better understand the dataset. For example, the ages of our students ranged from 15 to 22 years old with an average age of 16.66 years old. Additionally, the average G3 (final grade) was about 11.52 (with a range from 4 to 20 since we are excluding the students that received a G3 of 0).

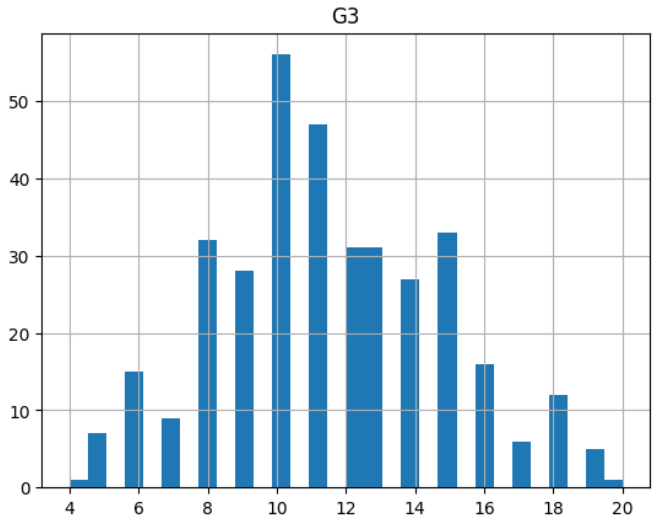

Figure 2: Histogram of the target variable, G3.

Figure 2 is helpful to understand the distribution of scores from the students in the dataset. It appears that the distribution has a bell shaped curve with the highest frequency of grades being around 10 and 11 which corresponds to the mean G3 that we obtained from Figure 1.

Next are a series of bar graphs that relate our features to the target variable. By analyzing these bar graphs, we can determine if there are any patterns between each feature and the mean final grade associated with each category.

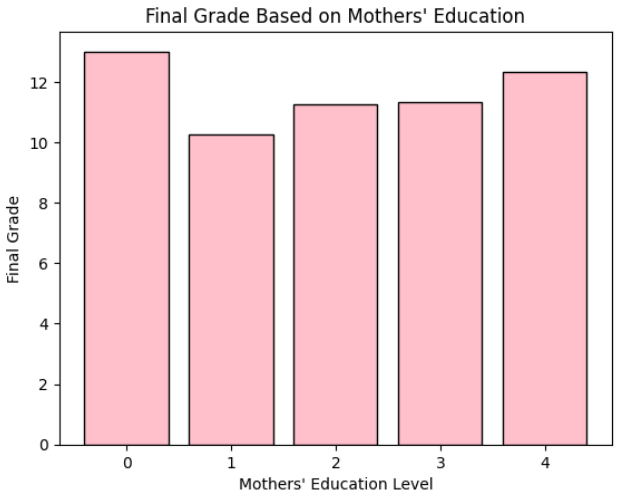

Figure 3: Final Grade Based on Mothers' Education

Surprisingly, the highest grades seem to be related to mothers with an education level of 0 (none). The rest of the data follows trend that the higher the mother's education, the higher the student scored. Perhaps the students whose mother had no education did so well because their mothers were stay-at-home mothers that were able to spend more time helping their children in school.

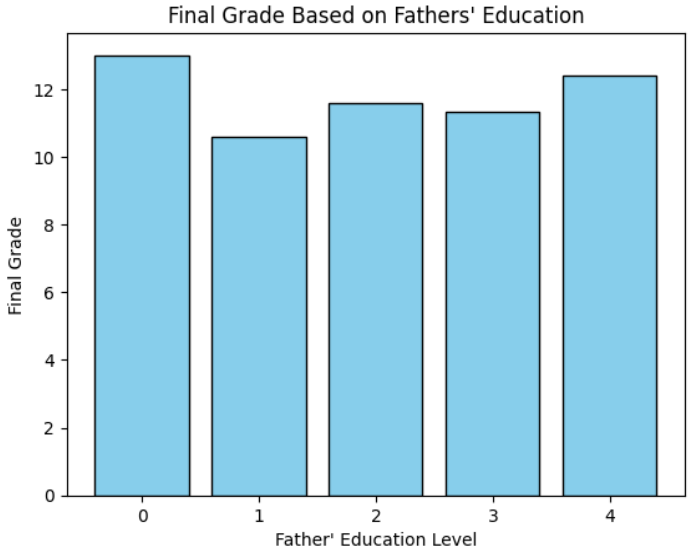

Figure 4: Final Grade Based on Fathers' Education

Once again, students whose fathers had no education seemed to have the highest final grade. This could be due to a similar reason where the fathers with no education might have been stay-at-home fathers that were able to spend more time helping their children with school. The rest of the data seems to follow the trend that the higher the fathers' education, the higher the students' grades.

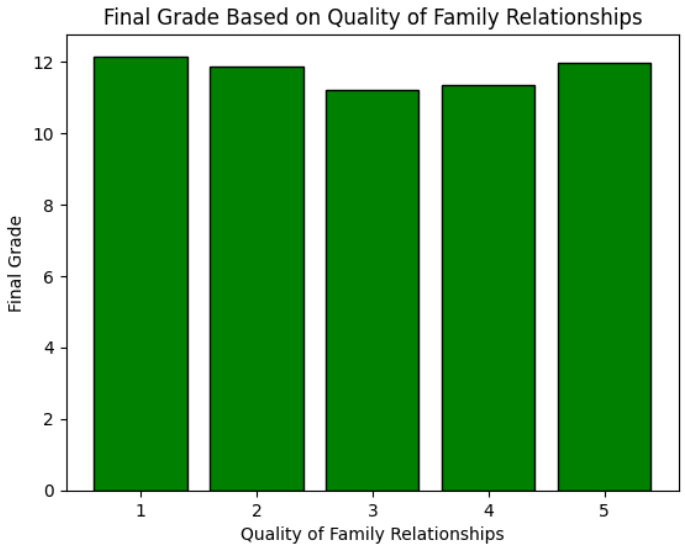

Figure 5: Final Grade Based on Quality of Family Relationships

The quality of family relationships doesn't seem to have much of an effect on how students performed in school. The average final grades are all quite close despite having a low quality of family relationships or a high quality of family relationships.

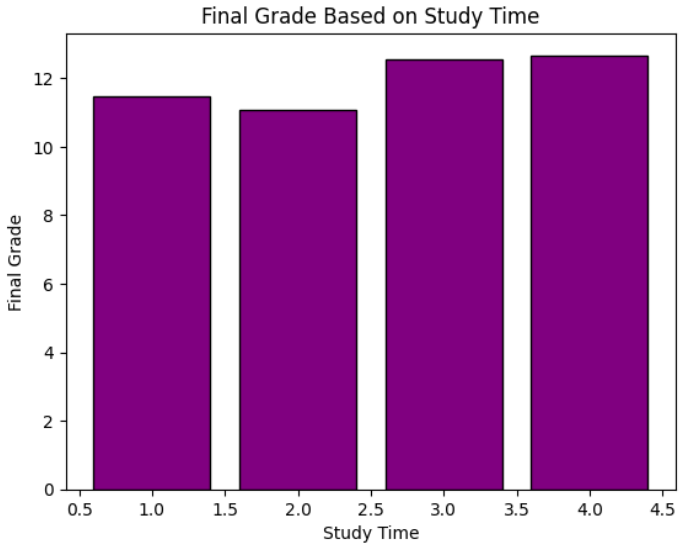

Figure 6: Final Grade Based on Study Time

Figure 6 shows that students who studied longer received better final grades compared to students who studied for less time. Students that rated their study time a 4 (more than 10 hours) received the highest final grade average. Students that rated their study time a 1 indicated that they studied for less than 2 hours, a rating of 2 indicated studying for 2 to 5 hours, a rating of 3 indicated studying 5 to 10 hours, and a rating of 4 indicated studying more than 10 hours. For some reason the final grade average for students that studied less than 2 hours is higher than the average of those that studied 2 to 5 hours but perhaps they were more efficient and focused during their shorter study period.

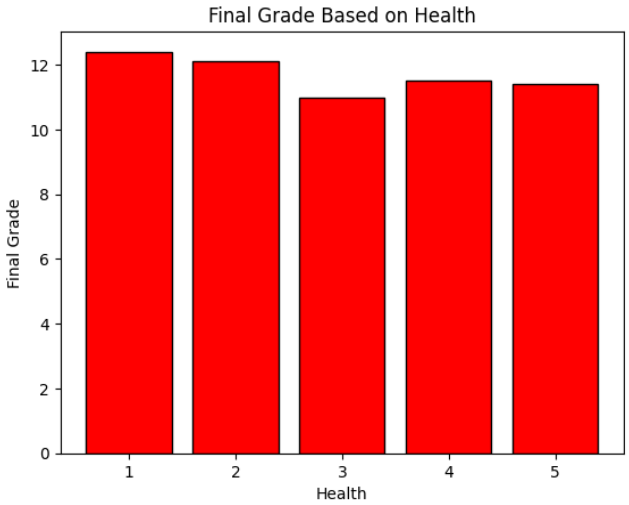

Figure 7: Final Grade Based on Health

Students were asked to rate their health from 1 (very bad) to 5 (very good) and it appears that students with lower health ratings had slightly better final grade averages than those with higher health.

Figure 8: Final Grade Based on Failures

The clear positive skew of this graph shows that students with less failures had higher final grade averages than those with more failures. This makes sense as students who wanted to perform well in school most likely did not fail as often as their counterparts.

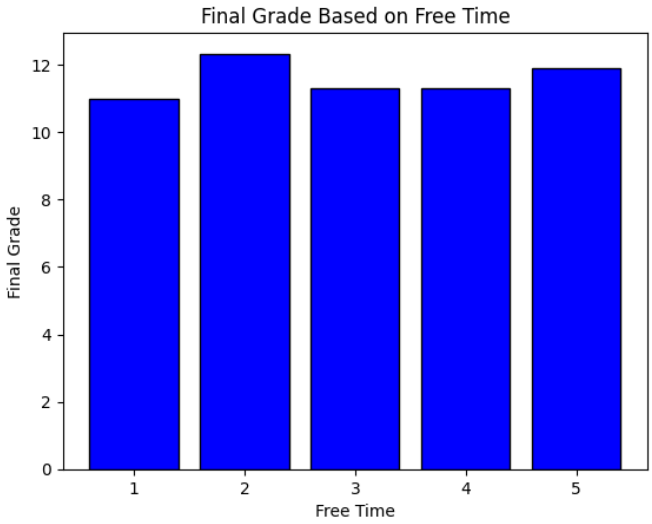

Figure 9: Final Grade Based on Free Time

The amount of free time that students had did not seem to have much affect on their final grades. The students were asked to rate their free time from 1 (very low) to 5 (very high). The students with the highest final grades seemed to have a free time rating of 2.

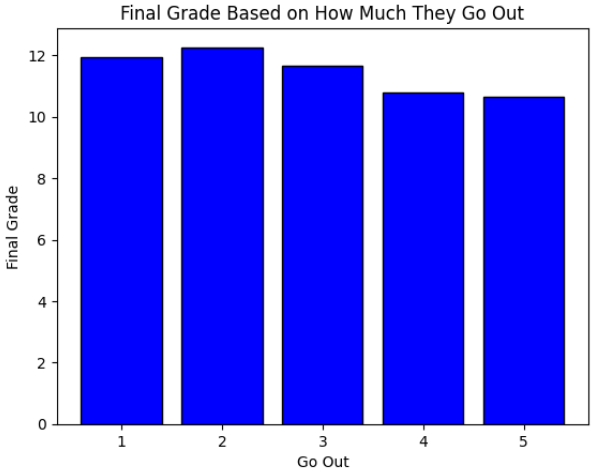

Figure 10: Final Grade Based on How Much They Go Out

Students were asked to rate how much they go out with friends from 1 (very low) to 5 (very high). Students seemed to have a lower final grade the higher they rated their amount of going out.

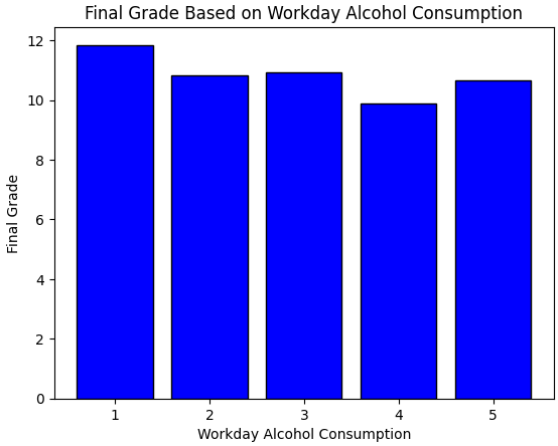

Figure 11: Final Grade Based on Workday Alcohol Consumption

Students were asked to rate their workday alcohol consumption from 1 (very low) to 5 (very high). There seems to be a pattern that the less alcohol students consumed on workdays the higher their final grade was. This is likely due to the negative effects associated with consuming alcohol that inhibit a student from performing well.

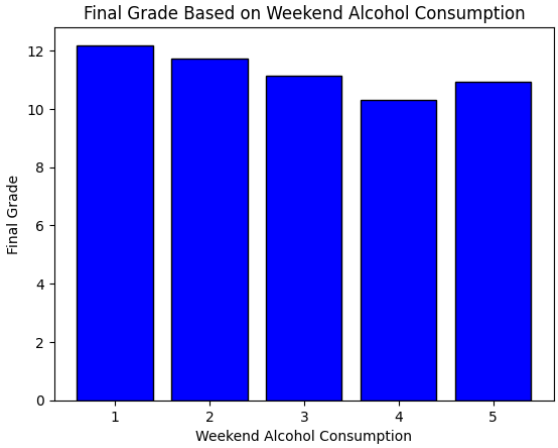

Figure 12: Final Grade Based on Weekend Alcohol Consumption

Similarly, students tended to do better in school when they consumed less alcohol on the weekends as well.

Modeling:

The first machine learning approach I used was linear regression. I used a linear regression model since I was dealing with supervised learning and more specifically, I was working with a regression problem. I used linear regression to model the relationship between the target variable and the independent variables by fitting a linear equation between predicted and actual values.

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt

# Create a linear regression model
model = LinearRegression()

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)
```

```
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
#####
print('R^2 score on testing data =', model.score(X_test, y_test))
```

In my code, I created a model, trained the model based off of the training set, and then made predictions on our test set. I then evaluated the model using the mean squared error and the R^2 score on the testing data. This is how the linear regression model was created.

Another machine learning technique I applied to the dataset was a Random Forest Regressor. The purpose of incorporating this was to be able to calculate the importance of each feature.

```
#calculate the importance of each feature:
importances = regr_model.feature_importances_

std = np.std([tree.feature_importances_ for tree in regr_model.estimators_], axis=0)
indices = np.argsort(importances)[::-1]
```

Calculating the importance of each feature would tell me how much each feature contributed to the model's predictions.

I also analyzed the REC curves for ridge regression, support vector regression, and a regression tree of the data to see how well each model would predict the data. I did this by scaling the X and y training sets using the "StandardScaler" from scikit-learn in order to standardize the features and target variables. Then I created a Ridge Regression model to fit the scaled training data and graphed the REC curve along with finding the root mean square error to measure the model's performance.

For the Linear Regression, Random Forest Regressor, and REC curve models I created two sets of models. The first set of models included G1 and G2 but the second set excluded G1 and G2 to analyze how the importance of other features affected the target variable.

The last machine learning approach I used was a neural network. I used TensorFlow to create a neural network with an input layer, two hidden layers, and an output later. I then evaluated the model on the scaled test set and printed the test loss and mean squared error.

Results:

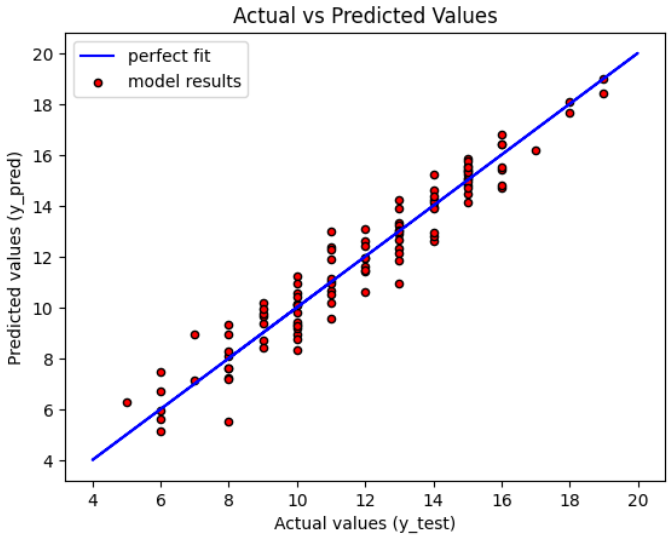

Figure 13: Actual vs. Predicted Values (Linear Regression including G1 and G2)

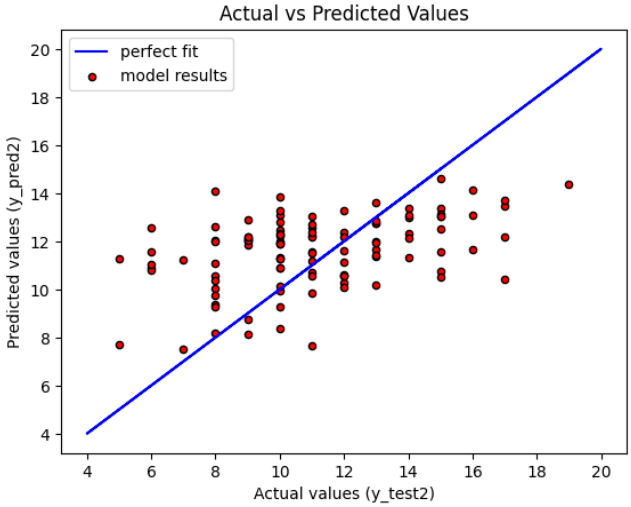

Figure 14: Actual vs. Predicted Values (Linear Regression excluding G1 and G2)

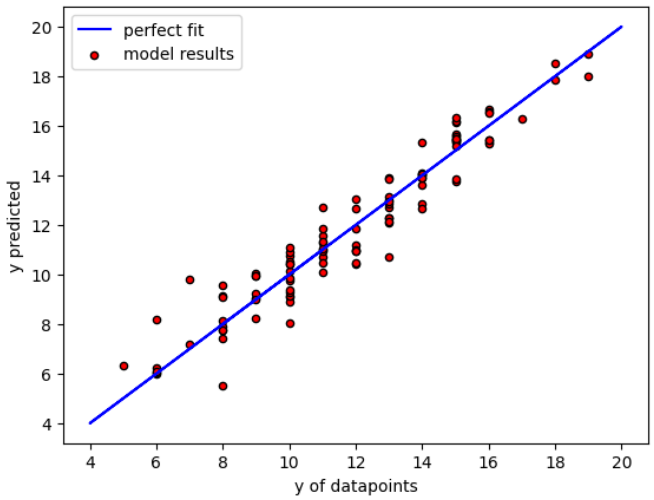

Figure 15: Random Forest Regressor Model (Including G1 and G2)

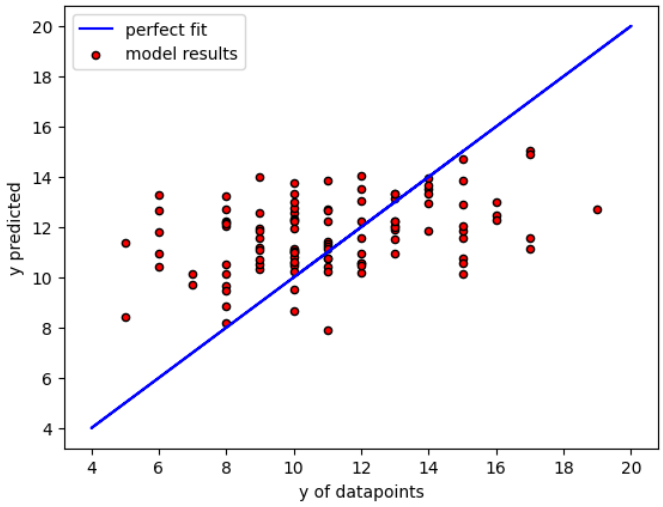

Figure 16: Random Forest Regressor Model (Excluding G1 and G2)

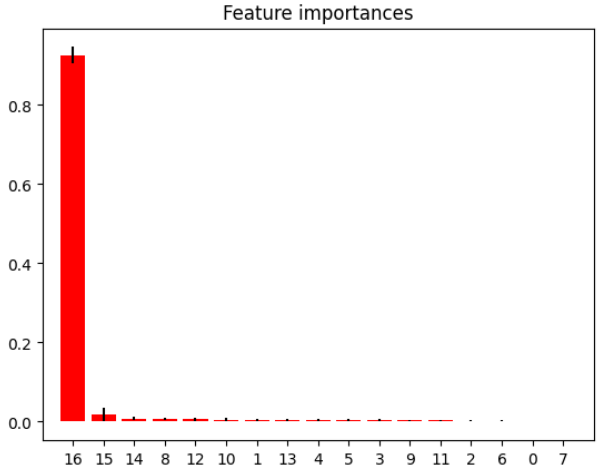

Figure 17: Feature Importances (Including G1 and G2)

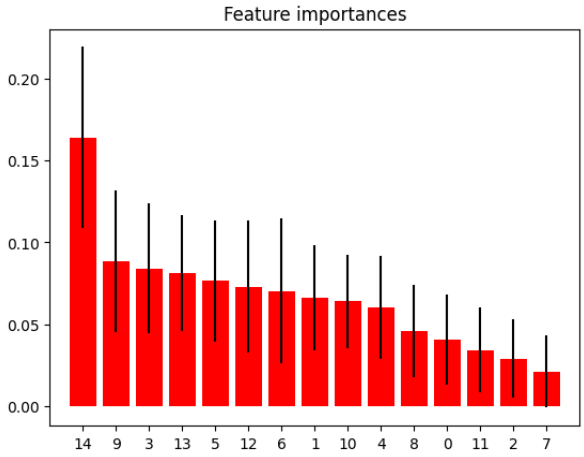

Figure 18: Feature Importances (Excluding G1 and G2)

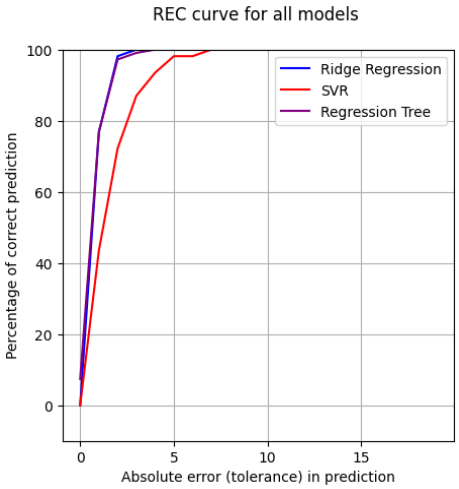

Figure 19: REC Curves for different models (inlcuding G1 and G2)

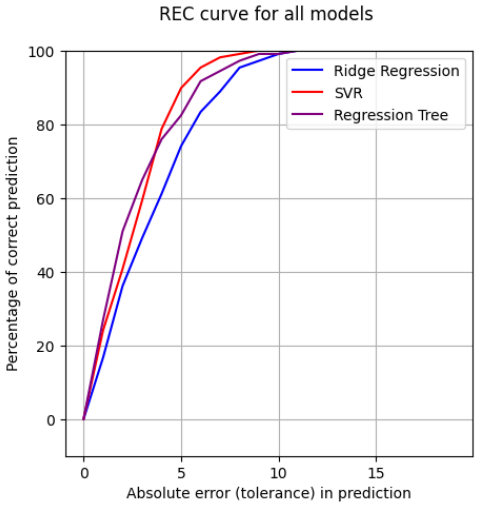

Figure 20: REC Curves for different models (excluding G1 and G2)

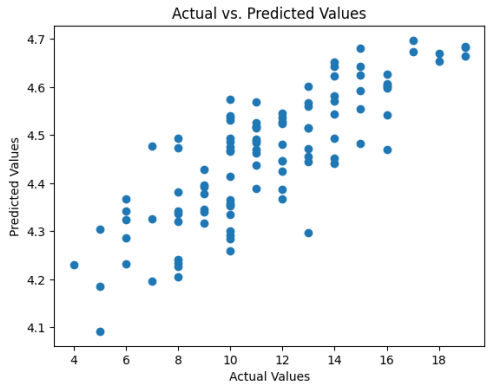

Figure 21: Actual vs. Predicted Values for Neural Network (Including G1 and G2)

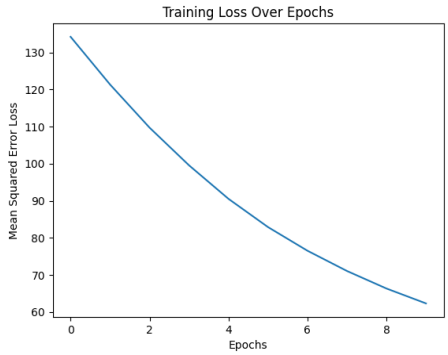

Figure 22: Training Loss Over Epochs for Neural Network (Excluding G1 and G2)

Discussion:

From Figure 13, one can see that the linear regression model's actual vs. predicted values were very closely related to the perfect fit line and had a mean squared error of 0.691 with an R^2 score of 0.929 on testing data. This means that the model performed well when creating predictions for the students' final grades. However, when looking at Figure 14, we see that the model did not perform as well when we excluded G1 and G2. This model produced a mean squared error of 7.345 and had an R^2 score of 0.145 on the testing data. This tells us that the linear regression model heavily relies on the features of G1 and G2 to accurately predict the students' final grades. We have a similar result when applying the Random Forest Regressor model to our data. When we include G1 and G2, we see that the predicted values are very close to the actual y values in Figure 15 and produce an R^2 score of 0.920. When we exclude G1 and G2, we see that our predictions of G3 become very random in Figure 16 and produce a poor R^2 score of 0.094. Therefore, both the linear regression and the Random Forest Regressor models heavily rely on G1 and G2 to accurately predict G3.

When analyzing Figure 17 and Figure 18, we see how important each feature was when calculating the predicted values for our Random Forest Regressor models. Note that we get a better visual when we exclude G1 and G2 from our data since we see which other features have an impact on our predictions. In Figure 17, we see that the model heavily relied on feature 16 (G2) and slightly on feature 15 (G1). When we excluded G1 and G2, we see that in Figure 18 the model relied mostly on feature 14 (absences), feature 9 (freetime), feature 3 (mothers'

education), and feature 13 (health). It makes sense that absences would have a large importance on determining the students' final grades since missing a lot of school would result in a lack of knowledge or understanding or course material.

Next, we can analyze the REC curves in Figure 18 and Figure 19. These are the REC curves for Ridge Regression, Support Vector Regression (SVR), and a Regression Tree. These graphs plot the absolute error in prediction against the percentage of correct prediction. In both models we see that the predictions become increasingly accurate when we allow slightly more error. Figure 18 shows that when an absolute error of about 5 is allowed, the percentage of correct prediction has already reached 100 for all three models. For Figure 19, it appears that we need at least an absolute error of about 10 in order for our percentage of correct prediction to reach 100. Figure 18 is slightly more accurate in the sense that a greater margin of error is tolerated for predictions in Figure 19. In Figure 18, the RMSE for Ridge Regression, SVR, and a Regression Tree are about 0.830, 1.984, 0.930 respectively. In Figure 19, the same results were 4.271, 3.190, and 3.439. This tell us that the models including G1 and G2 were still more accurate in predicting students' final grades because of their high correlation with those two features.

The last models to analyze are the graphs from our neural network model. I found that using 2 hidden layers produced the best output compares to using 1 or 3 hidden layers. Our test loss was 58.175 and our test mean squared error was 58.175 which indicate how the model performed on the testing data. Figure 21 shows the graph of actual vs. predicted values created a graph similar to our linear regression model which tells us both models achieve comparable results when trying to predict students' final grades. Figure 22 represents our graph of training loss over epochs and has a negative sloping downward curve. This provides some insight on the learning process of our model. A decreasing training loss indicates that the model is minimizing errors during training which increases its ability to find patterns in the data.

Conclusion:

In summary, the purpose of this project was to explore the relationships between various factors and see how each one impacted academic performance while also incorporating machine learning techniques to predict outcomes. The analysis covered several factors ranging from demographics, study habits, and family dynamics. From this work, the following conclusions were obtained.

The linear regression model demonstrated a strong predictive performance that closely related actual and predicted values which produced a low mean squared error with a high R^2 score when including the features of G1 and G2. When these scores were taken out, the model did not perform as well and this emphasized the importance of G1 and G2 when trying to forecast the final grades of students. The Random Forest Regressor reinforced this idea since the REC curves that included G1 and G2 performed better than the model that excluded them. This highlighted the importance of using the earlier grades to predict the final grades.

On the other hand, the neural network model produced a similar "Actual vs. Predicted Values" graph to the linear regression model which shows that the predicted values were quite accurate. Although it indicated a different learning process that had a decent amount of test loss

and mean squared error, the neural network seemed to predict the final grades well. The graph that displayed training loss over epochs showed a decreasing trend, which proved that the model minimized errors during training.

The limitations to my project included the fact that the models relied so heavily on factors G1 and G2 to accurately predict G3. The models would have been more helpful if they could rely on other features to predict student grades since the purpose of this project was to focus on which factors mainly affected the success of students.

To further develop this work in future projects, there are several ideas to take into consideration. Firstly, it might be helpful to expand the dataset and gather more factors that would provide a better prediction for the students' final grades without relying on G1 and G2. Possibly exploring factors related to students' financial conditions, academic environments, and gathering data from more schools might produce better predictions for certain models. Additionally, it might be helpful to explore different models that would be less reliant on G1 and G2 so that student performance could be more accurately predicted based off of their personal factors instead of their earlier performance in school. Future projects could also work with hyperparameter tuning to find more optimal combinations of features that would provide the most accurate performance of these machine learning models.

References:

[1] Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.