Analysing Academic Performance in High School Students

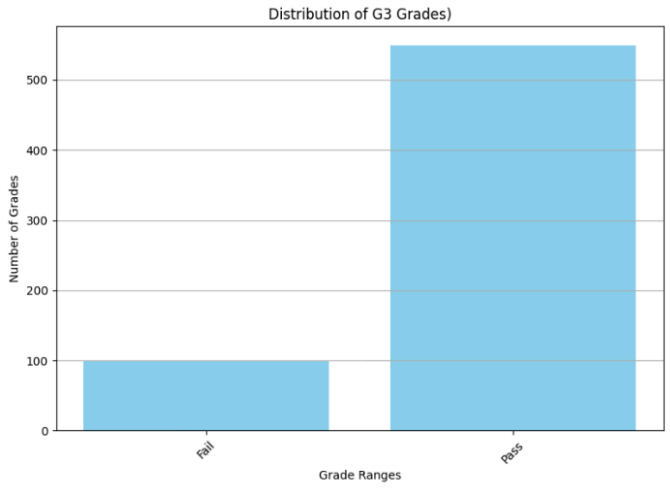
By: Nicholas Hobinca

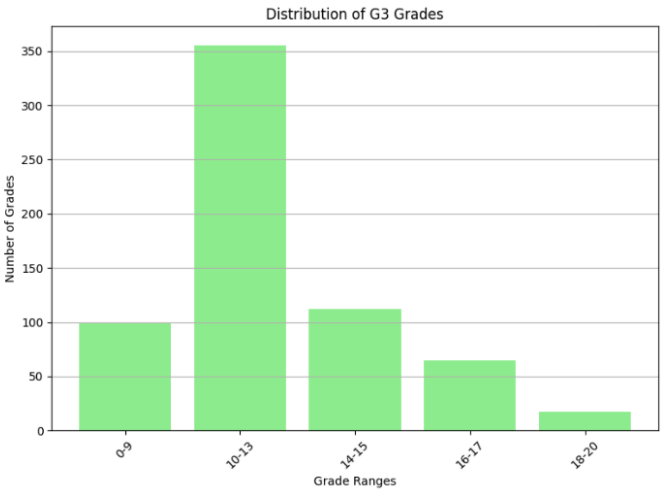
Understanding academic performance of students in school can be difficult to predict, as there is more to it than just measuring study time and focus. Many factors outside of school can potentially impact the ability of a student to do well and can greatly impact their final grade in a course. In a study conducted at two Portuguese secondary schools, 649 students were surveyed to learn what demographic, social and school factors could be at play that could influence academic performance and final grades. Various information was gathered. such as the number of absences from school, time spent studying, distance to school, extra-curricular activities, etc. Their grades for the high school subject Portuguese in semester 1 and semester 2 were gathered for each student, as well as their final year grade.

The first goal in this report is to create a model that can successfully predict whether a student will pass or fail the school subject Portuguese. Logistic regression will be used as this is a binary problem. A confusion matrix will then be used to assess the model's performance and determine if the model did well in predicting whether a student will pass or fail. The second goal in this report is to predict the actual final mark of a student using linear regression. We will use all the demographic, social and school factors in modelling, and then we will take a look using Random Forest to determine what factors have the strongest influence on predicting G3.

To start, the dataset consists of 649 rows, which are the students, and 30 columns, which consist of the demographic, social, and school factors. The dataset itself does not appear to have any missing values, so there is no need to clean the data. The feature variables (X) of the dataset are the demographic, social, and school factors, while the target variables (y) are G1, the semester one grade for the subject Portuguese, G2, the semester two grade for the subject Portuguese, and G3, the final grade of the student for the subject Portuguese. It is worth noting that G1 and G2 are heavily influential in predicting G3. Since the goal in this report is to ultimately predict the final grade G3, we will be using G1 and G2 in our models to help predict G3.

In Portugal secondary schools, the grading scale goes from 0-20, with a final grade of 0-9 resulting in failing the course. We create a bar graph to showcase which students have passed or failed Portuguese, and another bar graph showcasing how many students finished with each grade. It is worth noting that the Portuguese education system is tough, with a failed mark not being uncommon and a grade of 18-20 being almost impossible to achieve. We can visualise this in 2 bar charts; one illustrating students that have passed and failed the course, and another which shows the distribution of the grade ranges.






```
Number of students: 649
Number of students who failed: 100
Number of students who passed: 549
```

We can see that exactly 100 out of 649 students have failed Portuguese, which may seem unusual, but since Portugal's education system is difficult, this is expected.

Now, to perform logistic regression, we will create our binary conditions. The first condition will be if the student passes the course, so they will have a final G3 grade of 10-20. The second condition will be if the student fails the course, so they will have a final G3 score of 0-9. Since some of the feature variables consist of categorical variables, it is necessary to use one-hot encoding to convert them into a numerical format so they can be used for logistic regression. This ensures that each category variable is represented by a binary variable.

I used 80% of the data for training the model and the remaining 20% to test, so we will use the data from 130 students to determine if they pass or fail. As mentioned, all the feature variables, and G1 and G2 grades, will be used to help the model predict. After using one-hot encoding on the trained dataset, we standardise the dataset, and then perform logistic regression. I used the library sklearn for this as it provides many tools for data analysis, including logistic regression. In order to ensure replicable results, I set the random state of the training and testing to be 14 to guarantee the same data is used each time. Then, I use the accuracy command to see how well the logistic regression performed, and then print out the report.

```
Accuracy: 0.9153846153846154
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.76      0.78        25
           1       0.94      0.95      0.95       105

    accuracy                           0.92       130
   macro avg       0.87      0.86      0.86       130
weighted avg       0.91      0.92      0.91       130
```

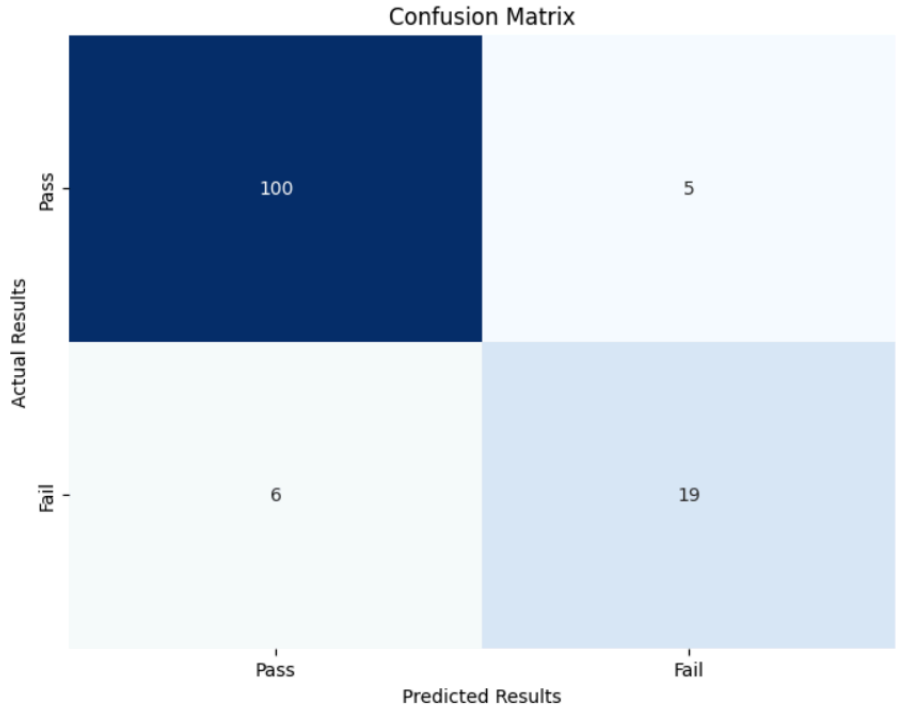
The report suggests that the model is performing well, with an overall accuracy of 91.53%. If we investigate the 'fail' and 'pass' precision of the model. We see the model correctly predicts if a student will pass with 94% accuracy, and claims that 105 people passed the course. However, it seems the model does a bit worse with predicting if a student will fail, with an accuracy of only 79%, and claims that 25 people failed the course.

The reason the model might struggle more with determining if a student will fail might have to do with some strange observation from the dataset. There are some observations where the final grade does not appear to make sense in connection with their G1 and G2 grades.

	G1	G2	G3
0	0	11	11
163	11	9	0
440	7	0	0
519	8	7	0

If we look at observations, it seems strange that observation 163 for example, has a semester 1 grade of 11, semester 2 grade of 9, but then a final grade of 0. It does not seem possible that this student could have finished with a grade of 0. The best guess is that during sampling either some grades were incorrectly taken or some external factors unrelated to the student's performance in the course resulted in a final grade of 0. While there are not too many of these issues observed in the dataset, this could result in accuracy of the model being off by a bit, at least when it comes to predicting if a student will fail.

To visualise the number of students that passed and failed, a confusion matrix is used to showcase if the model correctly predicted if a student passed, failed, or incorrectly guessed that a student passed when they failed or incorrectly guessed if a student failed when they passed. Remember that our test set consists of 130 students:



From the confusion matrix, we can see that 100 students that were predicted to pass the course did pass and 19 students that were predicted to fail did fail. However, it incorrectly predicted that 6 students that were supposed to pass actually failed, while 5 students that were predicted to fail actually passed. Overall, the model has done well in classifying between students who passed and students that failed. To address the issue above where some of the grades of G3 do not appear to make sense, if we remove all rows in the dataset which contain a 0 mark and attempt to perform logistic regression again with the same random state (14), our accuracy ends up a bit better, at 93.7%, and it predicts students that fail at a better rate:

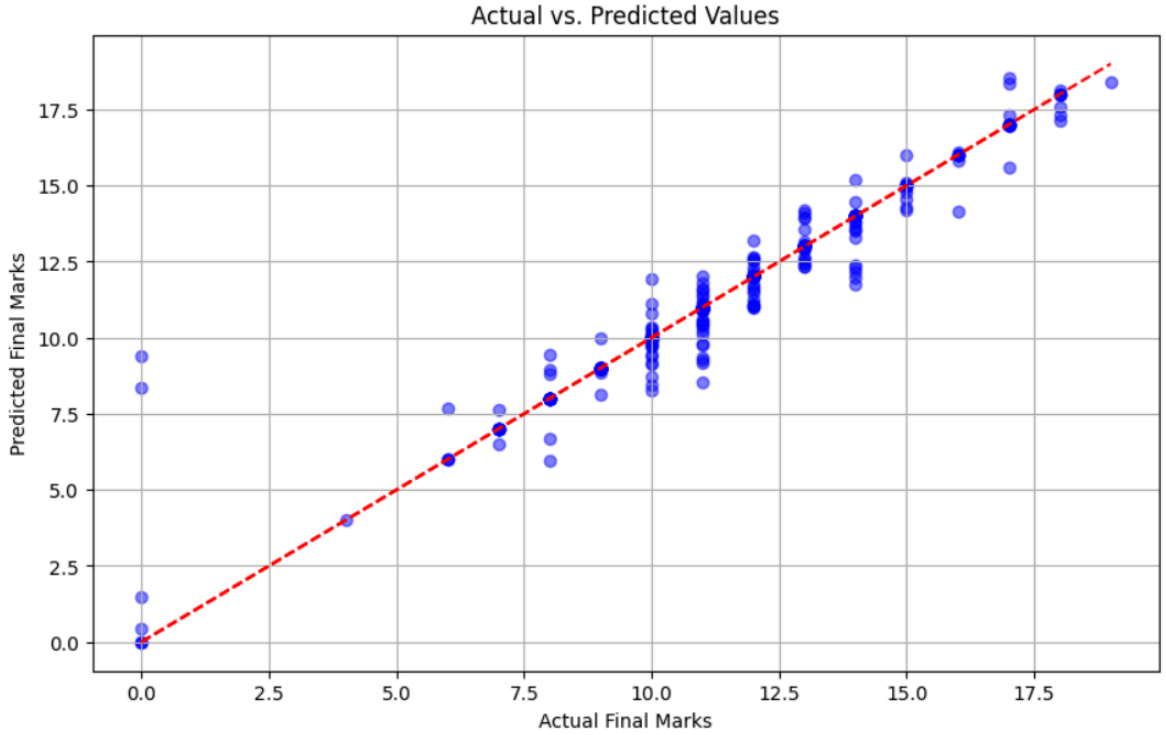
```
Accuracy (no rows have 0s): 0.937007874015748
Classification Report (no rows have 0s):
              precision    recall  f1-score   support

           0       0.83      0.62      0.71        16
           1       0.95      0.98      0.96       111

    accuracy                           0.94       127
   macro avg       0.89      0.80      0.84       127
weighted avg       0.93      0.94      0.93       127
```

Now, to move on to the 2nd part of this report, we take a look at using linear regression to predict a student's grades. We can approach analysing the dataset through linear regression to predict the student's grades and compare them to their actual grades. We will use the sklearn library to perform linear regression and once again use one-hot encoding to change the categorical variables to binary. We will again use all 30 feature variables and G1 and G2

target variables to predict G3, the final grade, and split the data 80% for training and 20% for testing. I chose to use a random state of 8 to guarantee reproducible results and showcase the mean squared error and R^2 score.

```
Mean Squared Error: 0.6776991639052931
R^2 Score: 0.936384301639675
```

From the results, we can see that the model does a good job of predicting G3 scores using all the feature variables and G1 and G2. A low mean squared error and a high R^2 score suggests the model predicts very well the actual final grade and captures approximately 93.6% of the variance. To visualise, we can plot the linear regression:

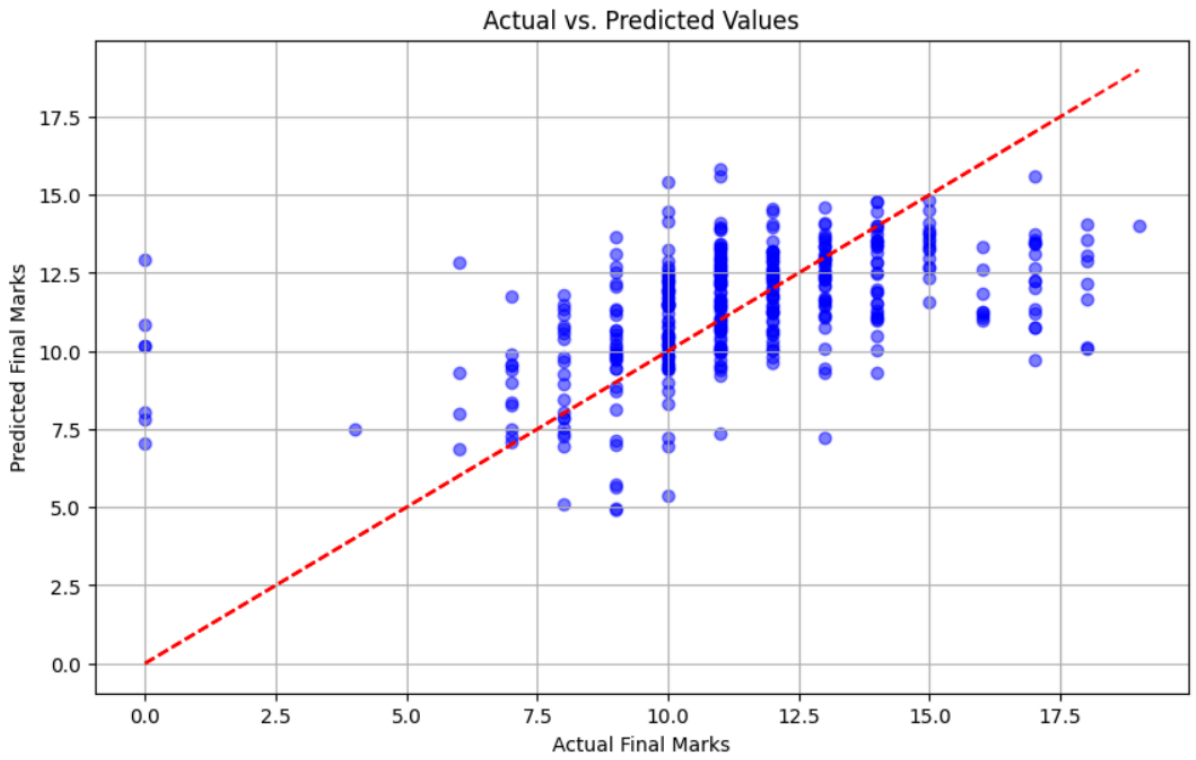


The plot shows the model performing well and predicts G3 nicely, aside from a few dots that showcase students who finished with a grade of 0.

Up to this point, we have been using G1 and G2 to help predict G3. Let's take a look at the same linear regression model with the same random state (8) to see what would happen if we did not train our data using G1 or G2.

```
Mean Squared Error: 6.980092438963116
R^2 Score: 0.2045348751638961
```

Clearly, the model has gotten worse, the mean squared error is much error and R^2 is lower, showing that the model fails to capture most of the variance. Once again we can plot the linear regression:

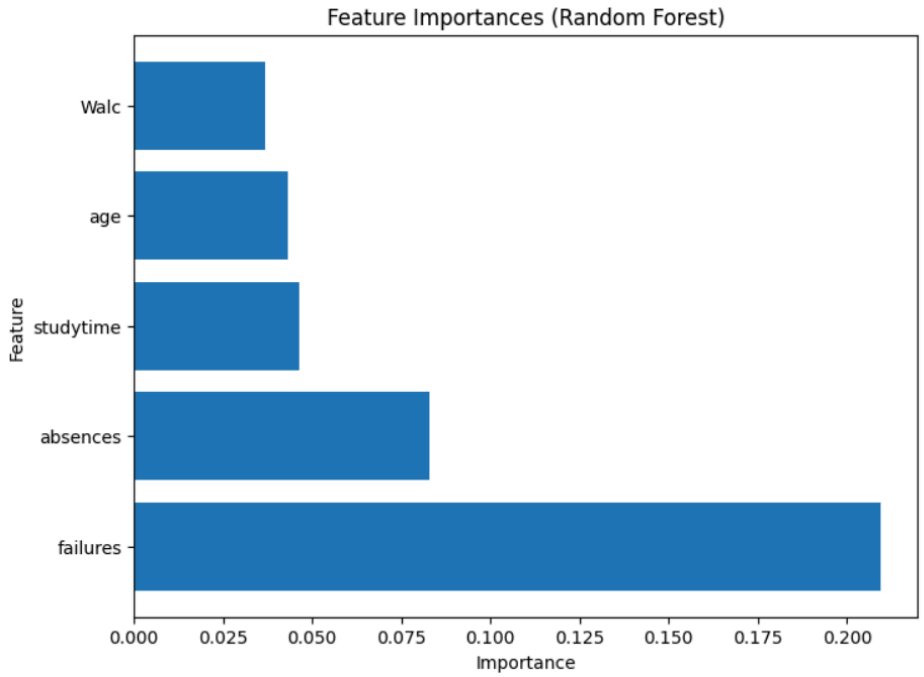


This confirms that G1 and G2 are crucial for testing to ensure that the model is accurate.

Now, we would like to learn which factors are the most important in helping to predict G3. We can use Random Forest regression from sklearn to learn which variables have the strongest influence. Since we already know that G1 and G2 have a great influence on predicting G3, it would be more beneficial to just look at the feature variables to determine which ones have the biggest influence. In order to ensure reproducibility, we choose to plot with the random state of 12 for both testing, training, and the random forest regressor:

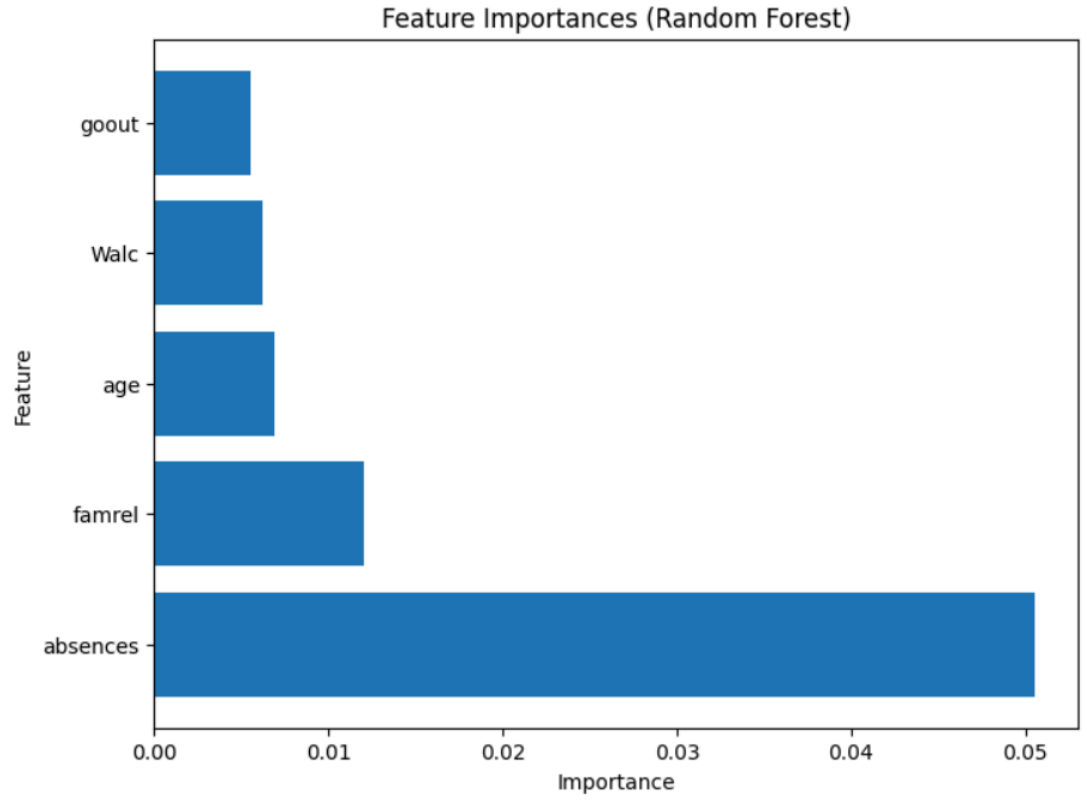
Originally, we only had 30 feature variables, but since we need to use one-hot encoding to transform the categorical vectors into binary values, we end up with 56 feature variables. For example, the ‘guardian’ category is split into 3 categories, guardian - Mother, guardian - Father, and guardian - other. This helps as now there are more categories that could potentially be important features. After performing Random Forest regression, we plot the top 5 most influential feature variables:

	Feature	Importance
48	failures	0.209315
55	absences	0.082892
47	studytime	0.046464
43	age	0.043083
53	Walc	0.036772



Our top 5 most important features in predicting G3 are previous failures in the class, followed by class absences, study time, age, and weekend alcohol consumption.

We can now look at the feature importances with G1 and G2 taken into account, and plot the feature importances, to see which features come out on time. We once again choose to plot with the same random state (12) for both testing, training, and the random forest regressor:



```
Total sum of feature importances: 0.1514180066334973
Number of variables: 56
```

We see that absences from class is the most influential feature variable in predicting G3, followed by the family relationship, health of the student, age and freetime. In total, it seems that the feature variables only help to explain about 15.6% in predicting G3, illustrating the importance of G1 and G2, which is the remaining percentage, in helping to predict G3.

Therefore from the 2 plots, we see the 3 variables that overlap in both cases are absences, age, and weekend alcohol consumption. Therefore, we can see knowing these 3 variables tells us the most about the feature variables to help predict G3.

In conclusion, we were able to analyse the dataset and determine what factors influence a student's academic performance. We first used logistic regression to determine if students would pass or fail the Portuguese course and we were able to create a good model. We then illustrated the results using a confusion matrix. We then used linear regression to plot predicted grades vs actual grades, and learned that without G1 or G2, it becomes very difficult to predict G3 grades and the model performs more poorly. Lastly, we used random forest to determine which factors have the highest influence and learned that approximately only 15% of the feature variables actually help to predict final grades. From the feature variables, we did learn that failures and absences from class showed up top in the most influential feature variables from the 2 plots. This suggests that if we know if a student has already failed the class in the past, then it should be easier to predict the final grade. Since they did poorly in the class before, it logically makes sense that they might not be able to perform well and continue to struggle in class. The other feature, absences, suggests that if a student repeatedly skips class, then they might end up struggling in class more often as they miss content.

Citations

[pdf] using data mining to predict secondary school student performance | semantic scholar. (n.d.). https://www.semanticscholar.org/paper/Using-data-mining-to-predict-secondary-school-Cortez-Silva/61d468d5254730bbecf822c6b60d7d6595d9889c

Grading system. Iscte. (n.d.). https://ibs.iscte-iul.pt/contents/the-experience/international-experience/incoming-mobility-exchange-students-faculty/1588/grading-system#:~:text=Under%20the%20Portuguese%20system%2C%20grades,minimum%20passing%20grade%20being%2010.&text=Exchange%20students%20take%20the%20same,a%20variety%20of%20different%20ways

Sklearn.ensemble.randomforestregressor. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.

Sklearn.linear_model.logisticregression. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Note: All code is in ipynb file