CS 418 FINAL PROJECT

By: Saikrishna Yadavalli, Saivikas Nethi, & Kaushal Mamgain

I. PROBLEM SELECTION

Our goal is to predict a secondary school student's final grade given all of the information on all students from 2 Portuguese secondary schools. In order to achieve this goal, we can use a linear regression model to predict the final grade that a student gets based on various attributes. We will use attributes like the student's age, their parents' employment, address, family size, and more to predict how they affect the final grade of that student. Furthermore, there are 33 different attributes that we can work with in order to find the more precise regression model. We will then determine the coefficient of determination for each model to inspect the performance of these models. In addition, we are using both a **student-por.csv** dataset and a **student-mat.csv** dataset for this project. We can use these datasets to predict the grade for each student on the basis of the collected information. This would, in turn, help us figure out the performance of the schools with respect to the students that are enrolled in them.

II. DATA COLLECTION

The **student-por.csv** and the **student-mat.csv** datasets both have 649 instances with 33 attributes. The attributes include age, sex, address, parents' education, parents' employment, and many more. These are multivariate datasets and have no missing values. The associated tasks that we can use on these datasets are linear regression and classification.

III. DATA PREPARATION

We followed the step-by-step guideline of the data preparation process in order. The first step was to inspect if the datasets were tidy. We didn't have to reshape the datasets to either long or

wide format. However, we noticed that both of our datasets included identical observations. In order to address this issue, we merged both of the datasets together based off of school, sex, age, address, family size, cohabilitation status, parents' education, parents' employment, and many more attributes. After doing so, the merged dataset successfully included all 382 students. Afterward, we moved onto the next stage of the data preparation process, which was detecting and correcting data quality problems. Since there were no missing values in both of our datasets and in the merged dataset, we applied the technique of dimensionality reduction on the merged dataset. In other words, we removed variables that were redundant or irrelevant because they would make our data analysis much more complex, according to the curse of dimensionality. Thus, selecting only a subset of all of the original variables in the merged dataset was more beneficial for our data analysis. In other words, selecting only the more important variables that contribute towards predicting final students grades for secondary school would help yield a more accurate data analysis. The resulting merged dataset that we obtained is shown below. This dataset below includes all of the unjoined attributes being separated into either math or Portuguese. The grades for all 3 periods were determined for both math and Portuguese.

	school	sex	age	address	Medu	Fedu	traveltime_mat	studytime_mat	schoolsup_mat	activities_mat	...	schoolsup_por	activities_por	freetime_por
0	GP	F	18	U	4	4	2	2	yes	no	...	yes	no	3
1	GP	F	17	U	1	1	1	2	no	no	...	no	no	3
2	GP	F	15	U	1	1	1	2	yes	no	...	yes	no	3
3	GP	F	15	U	4	2	1	3	no	yes	...	no	yes	2
4	GP	F	16	U	3	3	1	2	no	no	...	no	no	3

5 rows x 31 columns

Dalc_por	Walc_por	health_por	absences_por	G1_por	G2_por	G3_por
1	1	3	4	0	11	11
1	1	3	2	9	11	11
2	3	3	6	12	13	12
1	1	5	0	14	14	14
1	2	5	0	11	13	13

IV. DATA EXPLORATION

While working with the next step of the data science pipeline, which is data exploration, we explored our merged dataset using a list of summary statistics. We also generated a series of plots to visualize the relationships between final grades of a student and various attributes in the merged data. Within the summary statistics, we grouped the merged data by sex. The summary statistics were generated for both male and female students separately, as shown below.

```
In [10]: data_merged.groupby(['sex']).describe()
```

Out[10]:

	Dalc_mat								Dalc_por		...	traveltime_mat		traveltime_por						
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
sex																					
F	198.0	1.262626	0.605991	1.0	1.0	1.0	1.0	5.0	198.0	1.267677	...	2.0	4.0	198.0	1.409091	0.644633	1.0	1.0	1.0	2.0	4.0
M	184.0	1.701087	1.067554	1.0	1.0	1.0	2.0	5.0	184.0	1.701087	...	2.0	4.0	184.0	1.483696	0.753683	1.0	1.0	1.0	2.0	4.0

2 rows × 184 columns

The mean values for each attribute for both male and female students were shown below as well.

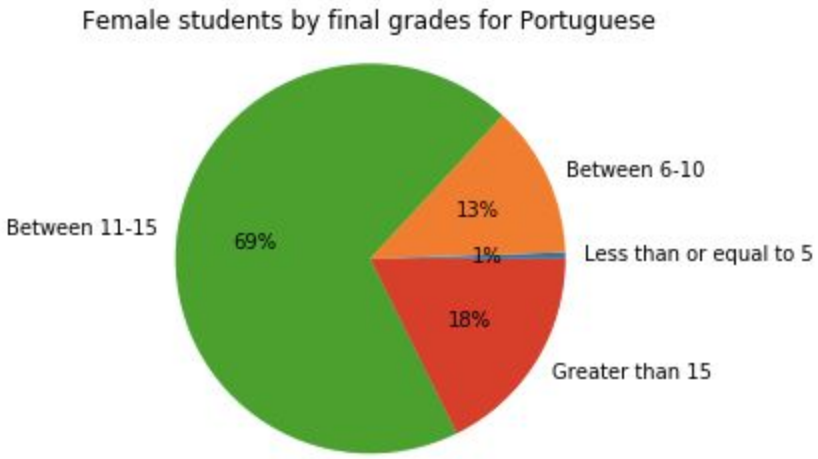
```
In [12]: data_merged[data_merged['sex'] == 'M'].mean()
```

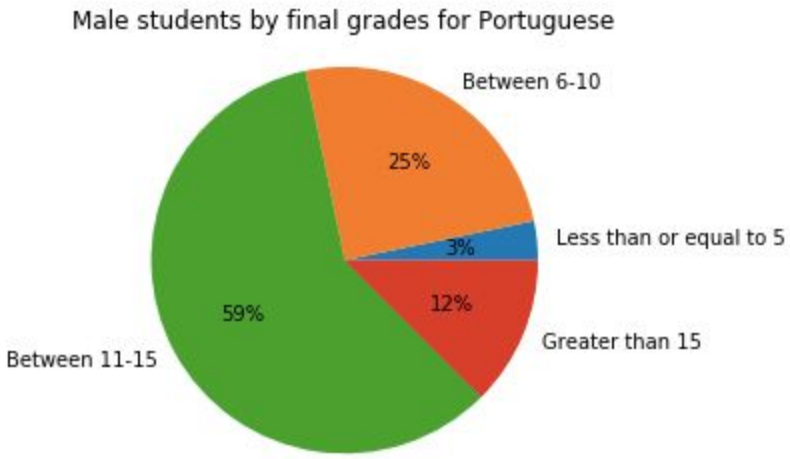
```
Out[12]: age               16.565217
         Medu               2.923913
         Fedu               2.619565
         traveltime_mat     1.483696
         studytime_mat      1.777174
         freetime_mat       3.445652
         Dalc_mat           1.701087
         Walc_mat           2.619565
         health_mat         3.782609
         absences_mat       4.815217
         G1_mat            11.293478
         G2_mat            11.173913
         G3_mat            10.978261
         traveltime_por     1.483696
         studytime_por      1.788043
         freetime_por       3.445652
         Dalc_por           1.701087
         Walc_por           2.625000
         health_por         3.782609
         absences_por       3.625000
         G1_por            11.646739
         G2_por            11.750000
         G3_por            11.902174
         dtype: float64
```

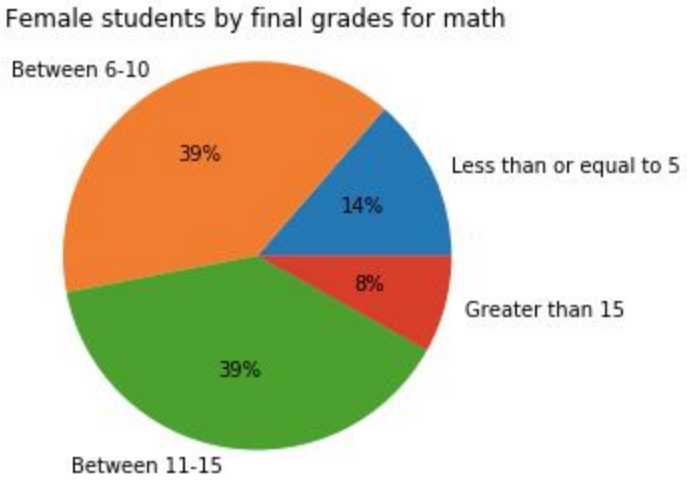
```
In [11]: data_merged[data_merged['sex'] == 'F'].mean()
```

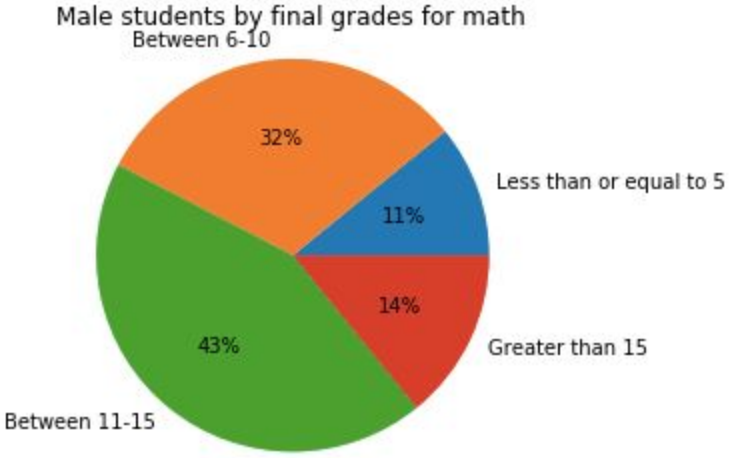
```
Out[11]: age               16.606061
         Medu               2.696970
         Fedu               2.515152
         traveltime_mat     1.404040
         studytime_mat      2.272727
         freetime_mat       3.015152
         Dalc_mat           1.262626
         Walc_mat           1.964646
         health_mat         3.388889
         absences_mat       5.787879
         G1_mat            10.459596
         G2_mat            10.282828
         G3_mat             9.838384
         traveltime_por     1.409091
         studytime_por      2.272727
         freetime_por       3.030303
         Dalc_por           1.267677
         Walc_por           1.979798
         health_por         3.383838
         absences_por       3.717172
         G1_por            12.545455
         G2_por            12.691919
         G3_por            13.085859
         dtype: float64
```

We then started to make assumptions about the final grades that the students got based on multiple different attributes. We used their sex, the amount of time they spend on studying during the week and if they had internet access at home and how that effected the grades.

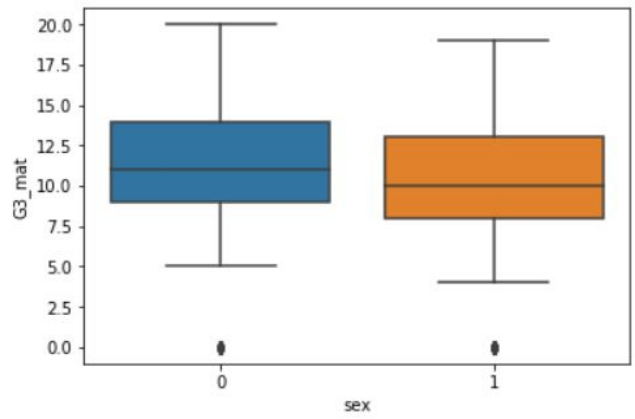




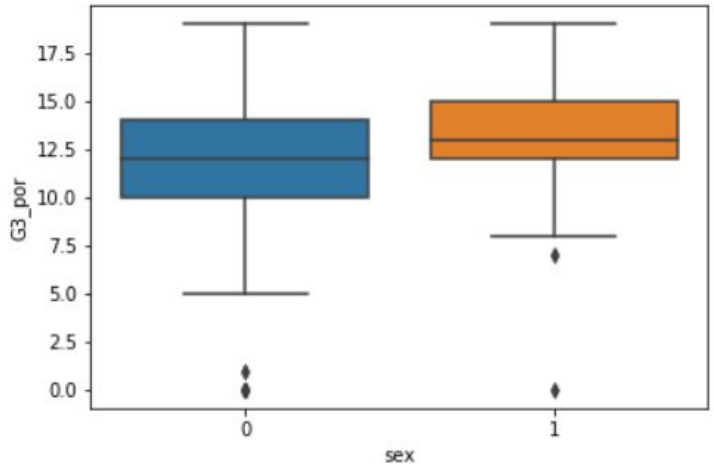




These pie charts were the first way we used to measure the grade distributions in the classes according to the sex. We were able to see how the grades were distributed for both males and females.



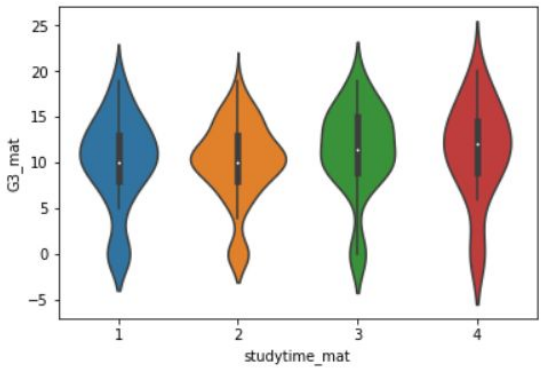
This box plot shows the grade distribution for math grades in both male and females. 0 is male and 1 is female. From this box plot we see that the average is approximately 10.978 for the males and that is a higher average than the females whose average is approximately 9.838.



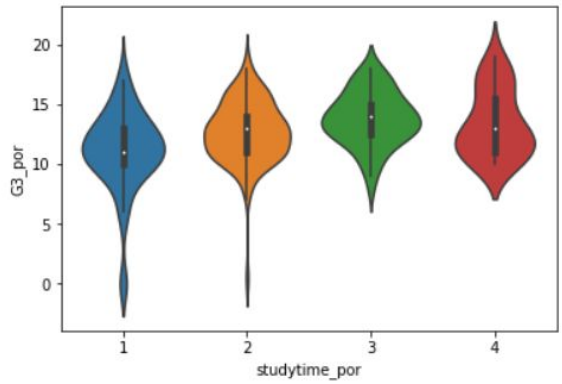
This box plot shows the grade distribution for the porteguese class grades for both male and females. The average for the male students was approximately 11.902 and the females students is approximately 13.086. The female students had a higher average in this class.

With this data, we can conclude that males did better for math but the females did better for the porteguese class.

We then plotted violin plots to see how grades are affected based on the amount of time the students spend on studying. 1 is less than 2 hours a week, 2 is 3 to 5 hours a week , 3 is 5 to 10 hours a week and 4 is 10+ hours a week.

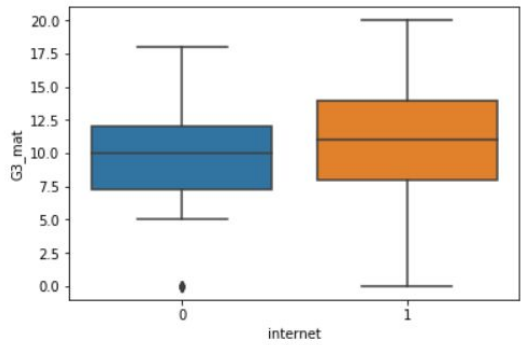


This violin plot shows the amount of time the students in the math class spent studying in the week. The grade distribution were higher for those who studied more than 5 hours a week.

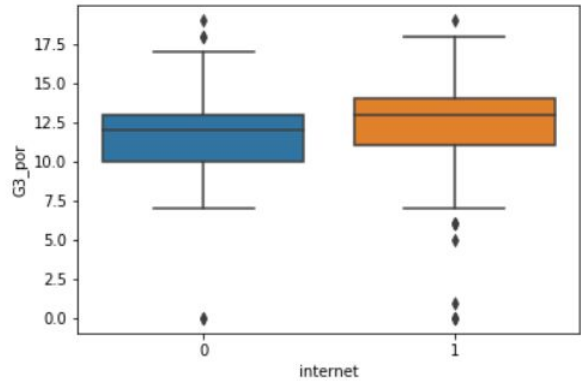


This violin plot shows the amount of time the students in the portuguese class spent studying in the week. The grade distribution were higher for those who studied more than 5 hours a week. From these violin plots, we can assume that the more time students spend studying the higher their grades are.

For the last attribute, we used box plots to analyze whether having internet access at home helps students get higher grades. 0 is no internet at home and 1 is there is internet access at home.



This box plot shows whether the students in the math class having internet access at home helped them get higher grades. This shows the average grade was higher for the students that had internet access at home than those that did not.



This box plot shows whether the students in the portuguese class having internet access at home helped them get higher grades. This shows the average grade was higher for the students that had internet access at home than those that did not.

After analyzing these box plots, we can assume that students with internet access at home have a higher grade on average than those who do not.

V. DATA MODELING

We have performed two operations here:

1) Regression:

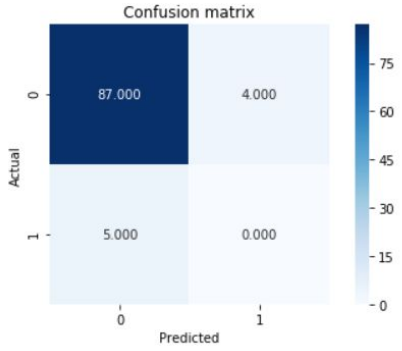
   To perform this we have used a linear regression model. The main purpose of regression analysis is to determine such a set of predictor variables that most affect the predicted variable. Using regression we can interpolate and extrapolate the values for predicted values. Here we have used attributes like Mother's education, Father's education, Failures, Age, Health, Absence to build a linear regression model and we were able to get a score of 86.673 for our model. We have also calculated the value for the test set.

2) Classification:

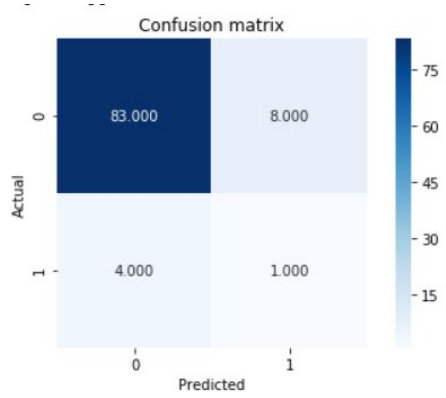
   Our main motive in this task is to predict the school of a student based on his/her Age, Health, Activities, Internet and various other attributes. We have used four different kinds of algorithms to perform this task, which are, K-Nearest Neighbours, Naive Bayes, Support Vector Machine and Kernel Support Vector Machine. For each algorithm we have been able to compute the performance metrics like Accuracy, Recall, Precision, Error and F-1 Score. We are not getting very good accuracy because of class imbalance in the data.

   Below are attached the confusion matrix we got for each algorithm and the performance metric for them.

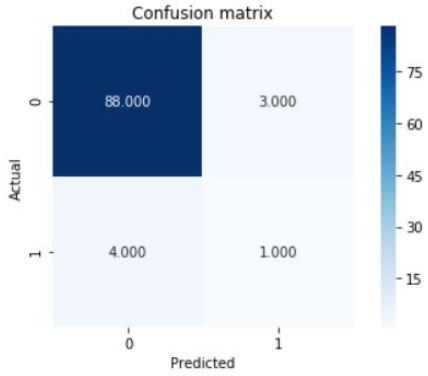
K-Nearest Neighbours



Naive Bayes



Support Vector Machine



Kernel Support Vector Machine