# *SG7*

# *5302 Final Project Report*

Student Performance Factors

Austin Steel

Sarah Starkey Lanham

Mahender Bandi

Sai Bhargav Reddy Kanuparthy

**Problem Statement**

Educators are always searching for ways to help their students improve on tests. However, there are so many different factors to consider, often Educators do not even know how to help their students. This is why collecting data about students can be very helpful in directing Educators on what makes students pass or fail tests.

To answer this question, we looked at a dataset titled "Student Performance Factors". This dataset includes a wide variety of information about students, and their performance on a test. The goal of this analysis will be to see if any specific factors or combinations of factors effect students passing or failing the test, and to create a model which we can use to predict if students will pass or fail the test.

**Suggestion**

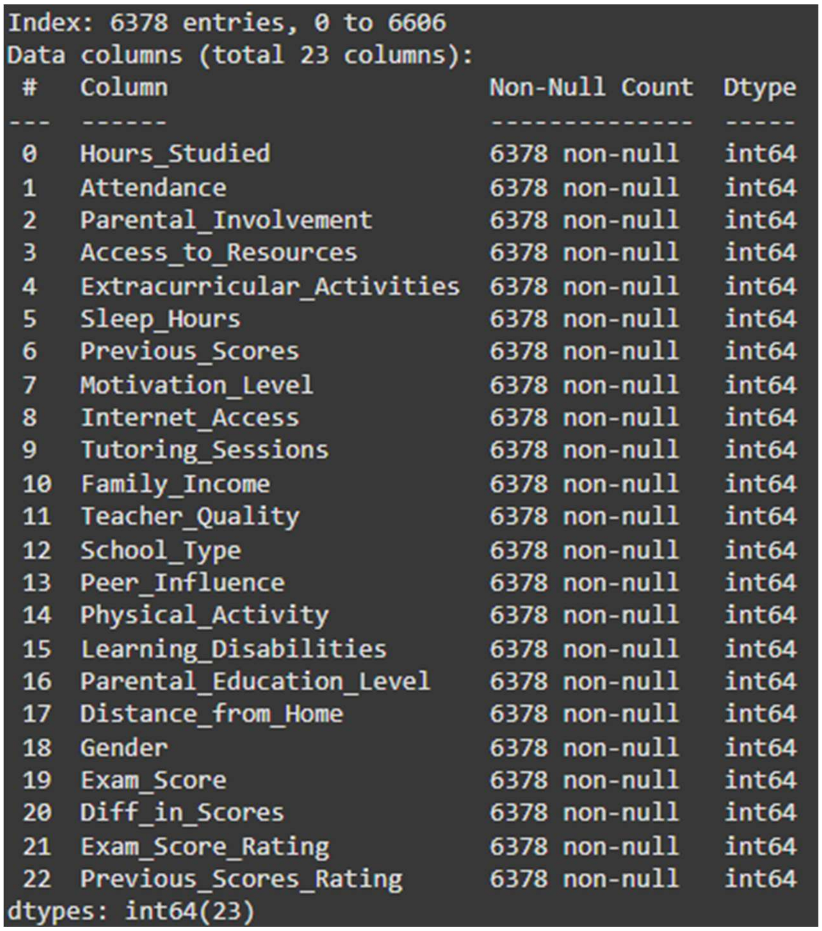
**Data Preprocessing**

Before we can begin our analysis of the dataset, we had to preprocess the data. Initially there were less than 1% of values missing from the data. These missing values were random and did not seem to have any correlations. Therefore, we decided to remove any entries with missing values. Next, we considered the categorical variables with string entries. These variables all had two or three unique entries such as *High, Medium, Low* or *Positive, Neutral, Negative* or *Male, Female.* We replaced all of these with numerical values (1, 2, 3 or 0, 1, etc.). We also added a column *Score_Difference* because we were given *Previous_Exam_Scores* and *Exam_Score*, so we could consider student improvement as well.

For our model we want to be able to predict if a student will pass or fail the test. Therefore, we needed to convert the *Exam_Score* column into a binary option, *Pass* or *Fail*. We used the standard cutoff of 70% as our pass/fail line and set passing as 1 and failing as 0. Since we want to keep the original *Exam_Score* column still, we added this variable and called it *Exam_Score_Rating.* We also added a column called *Previous_Exam_Rating* by applying the same algorithm to the *Previous_Exam_Scores* variable.

**Exploratory Data Analysis (Descriptive Analysis)**

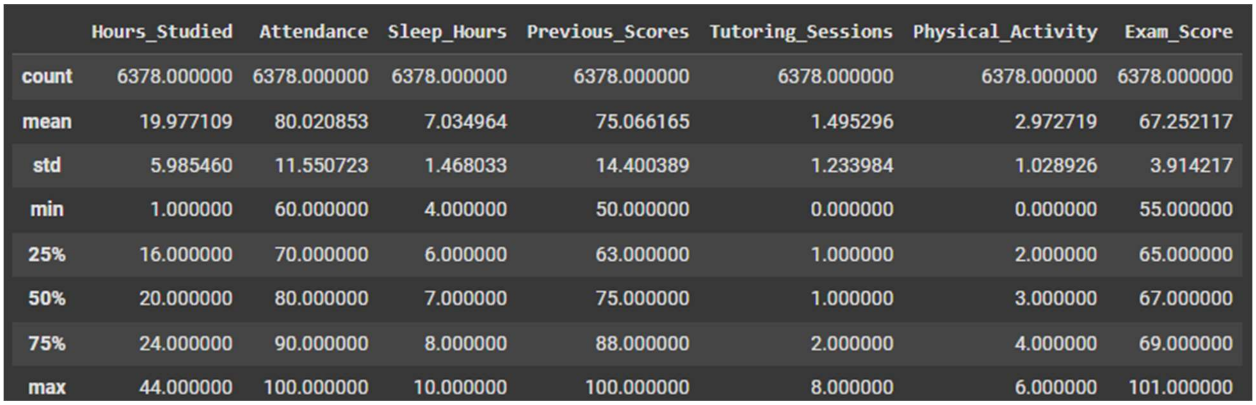
After we preprocessed our data, we were left with 23 columns and 6378 row entries. Each student is a row entry, and each column is a factor about them. These variables give us a wide range of factors.



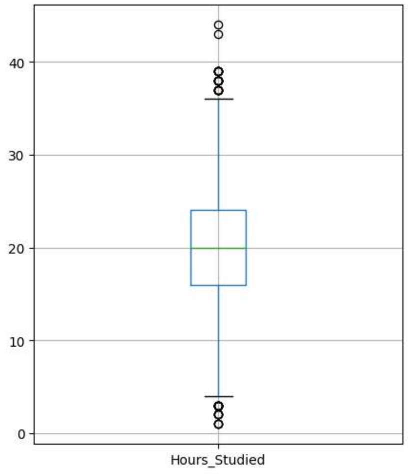
```
Index: 6378 entries, 0 to 6606
Data columns (total 23 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Hours_Studied               6378 non-null   int64
 1   Attendance                  6378 non-null   int64
 2   Parental_Involvement        6378 non-null   int64
 3   Access_to_Resources         6378 non-null   int64
 4   Extracurricular_Activities  6378 non-null   int64
 5   Sleep_Hours                 6378 non-null   int64
 6   Previous_Scores             6378 non-null   int64
 7   Motivation_Level            6378 non-null   int64
 8   Internet_Access             6378 non-null   int64
 9   Tutoring_Sessions           6378 non-null   int64
 10  Family_Income               6378 non-null   int64
 11  Teacher_Quality             6378 non-null   int64
 12  School_Type                 6378 non-null   int64
 13  Peer_Influence              6378 non-null   int64
 14  Physical_Activity           6378 non-null   int64
 15  Learning_Disabilities       6378 non-null   int64
 16  Parental_Education_Level    6378 non-null   int64
 17  Distance_from_Home          6378 non-null   int64
 18  Gender                      6378 non-null   int64
 19  Exam_Score                  6378 non-null   int64
 20  Diff_in_Scores              6378 non-null   int64
 21  Exam_Score_Rating           6378 non-null   int64
 22  Previous_Scores_Rating      6378 non-null   int64
dtypes: int64(23)
```

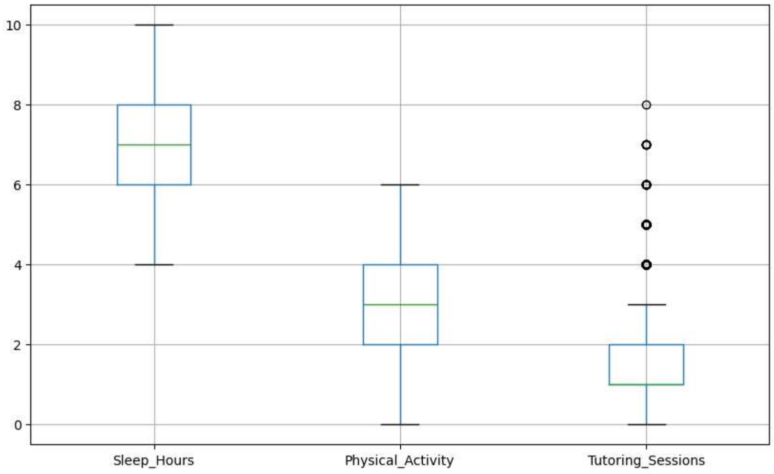
The variables *Hours_Studied, Attendance, Sleep_Hours,* and *Tutoring_Sessions* and *Physical_Activity* are the only variables which were originally numerical, and we did not convert from strings to numerical. For *Hours_Studied* we see a large range from 1 to 44 hours, with an average of about 20 hours. *Attendance* averaged at 80% and ranged from 60% to 100%. *Sleep Hours* varied from 4 to 10 with the mean of about 7 hours.

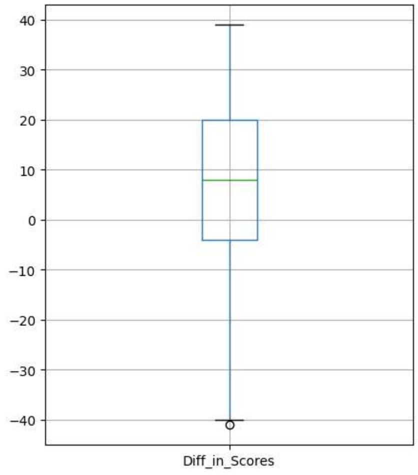
*Tutoring_Sessions* is a count of how many sessions the students attended outside of class and ranged from 0 to 8 with an average of approximately 1.5. *Physical_Activity* is a measure of how many hours the students spend doing physical activities each day and ranged from 0 to 6 with a mean of approximately 3 hours.

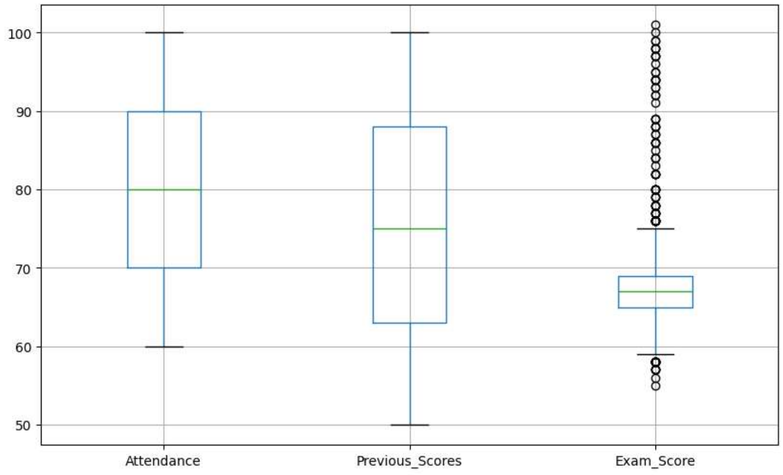


| | Hours_Studied | Attendance | Sleep_Hours | Previous_Scores | Tutoring_Sessions | Physical_Activity | Exam_Score |
|---|---|---|---|---|---|---|---|
| **count** | 6378.000000 | 6378.000000 | 6378.000000 | 6378.000000 | 6378.000000 | 6378.000000 | 6378.000000 |
| **mean** | 19.977109 | 80.020853 | 7.034964 | 75.066165 | 1.495296 | 2.972719 | 67.252117 |
| **std** | 5.985460 | 11.550723 | 1.468033 | 14.400389 | 1.233984 | 1.028926 | 3.914217 |
| **min** | 1.000000 | 60.000000 | 4.000000 | 50.000000 | 0.000000 | 0.000000 | 55.000000 |
| **25%** | 16.000000 | 70.000000 | 6.000000 | 63.000000 | 1.000000 | 2.000000 | 65.000000 |
| **50%** | 20.000000 | 80.000000 | 7.000000 | 75.000000 | 1.000000 | 3.000000 | 67.000000 |
| **75%** | 24.000000 | 90.000000 | 8.000000 | 88.000000 | 2.000000 | 4.000000 | 69.000000 |
| **max** | 44.000000 | 100.000000 | 10.000000 | 100.000000 | 8.000000 | 6.000000 | 101.000000 |

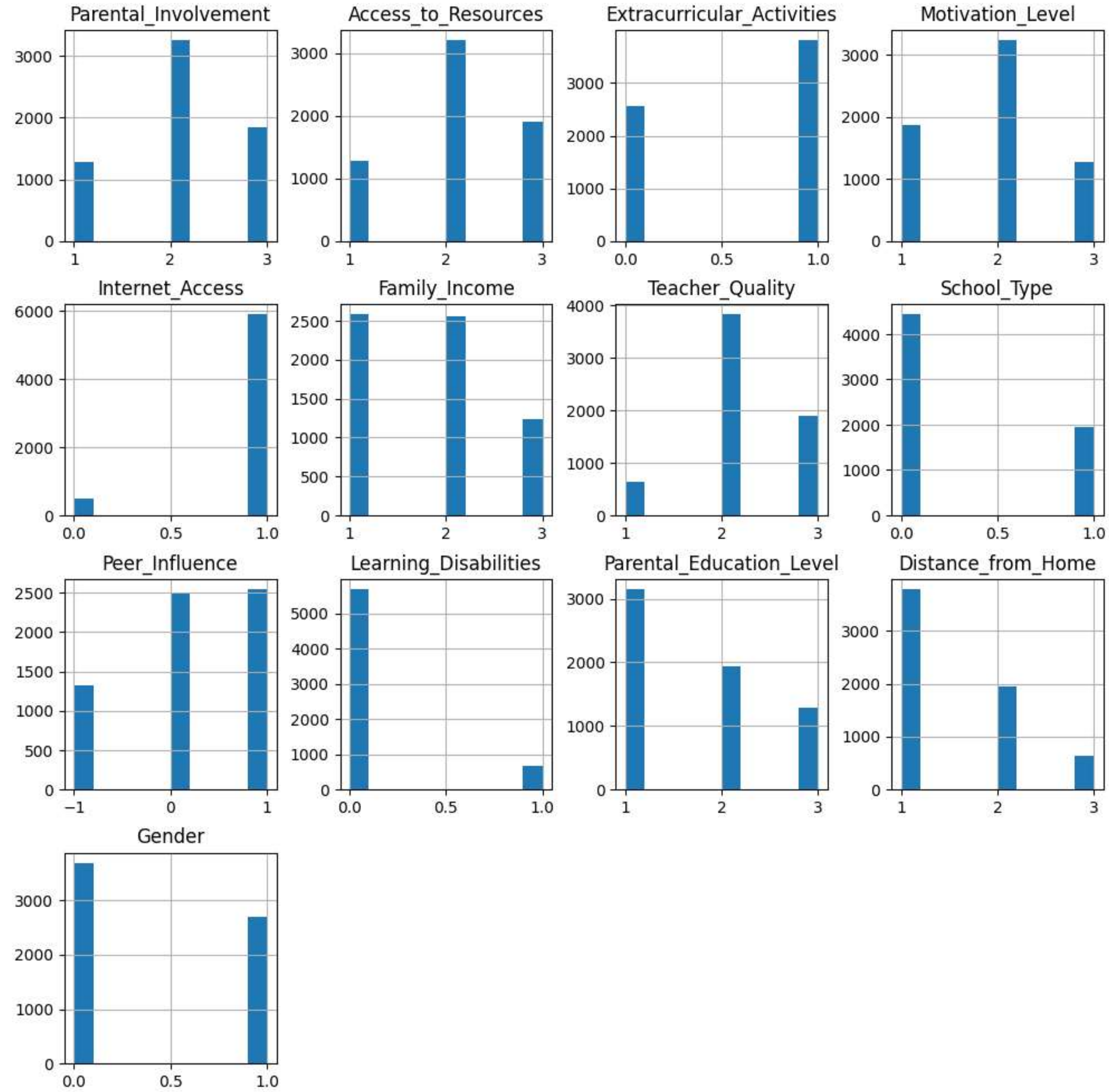




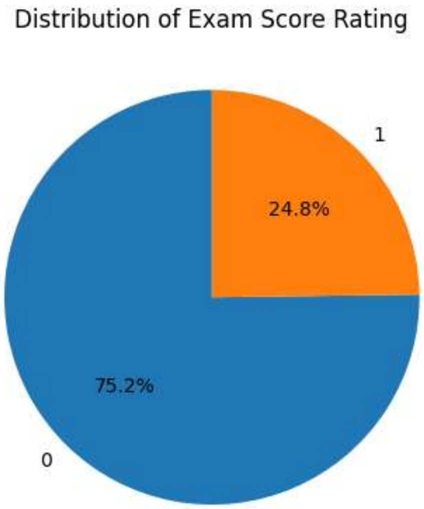




These boxplots show us the distribution of these variables we did not convert. For the variables we converted to numerical entries the distributions are shown in histograms below.

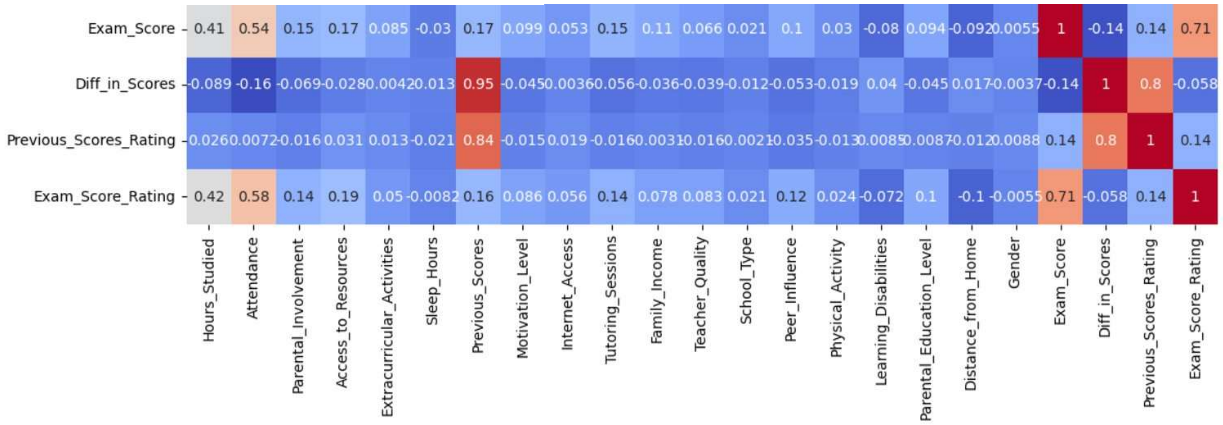


Our target variable is *Exam_Score* and the column we added *Exam_Score_Rating.* For our *Exam_Score* we range from 55 to 101 with an average of 67.23. The average is below the cutoff for passing. When we look specifically at the *Exam_Score_Rating* variable we added, we notice most of the students did not pass the test. Approximately 75.2% of all students failed the test, and only 24.8% of students passed the test.



This is a problem for the model we want to build because the data is imbalanced. To resolve this problem, we used the SMOTE method to oversample the data to balance the dataset. After we did this, our new dataset had 9,594 entries and the distribution of Exam_Score_Rating is now 50:50.

We also normalized the data and then created a correlation heatmap to see what factors were most correlated to our target variables.

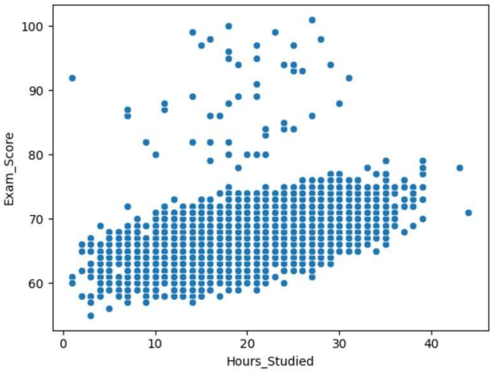


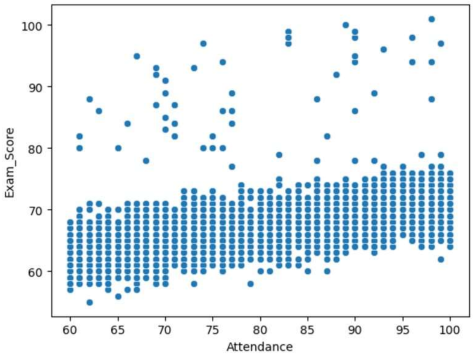
The two factors that had the biggest positive impact on our target variable (*Exam_Score_Rating)* were *Hours_Studied* and *Attendance* with correlations of 0.42 and 0.58 respectively. This makes sense since if students do not study the material or go to class, they typically don't pass tests. These two are much higher than any other factors with the next highest being *Access_to_Resources* at 0.19, followed by *Previous_Exam_Scores* (0.16), and then *Parental_Involvement* and *Tutoring_Sessions* at 0.14 both. The largest negative impacts come from *Learning_Disabilities* and *Distance_from_Home,* although both of these have very low correlations close to 0.

There are some interesting observations about these correlations. The first two, Hours Studied, and Attendance, are much higher than all the rest. The correlation between Previous_Scores_Rating and Exam_Score_Rating is only 0.14, which is very low, and we would have expected it to be higher. This means many of the students who passed the test were not the ones who passed the previous test, and vice versa. Also, we would have expected several of the other variables to have higher correlations than we observe. For example, *Sleep_Hours, Motivation_Level, Teacher_Quality,* and

*School_Type*, are normally factors believed to be very important in student performance. However, the data tells us they all seem to have very little effect.

We created scatterplots of the two highest correlated factors, *Attendance* and *Hours_Studied.* From these scatterplots, we can confirm the positive relationship between these variables and *Exam_Score.*

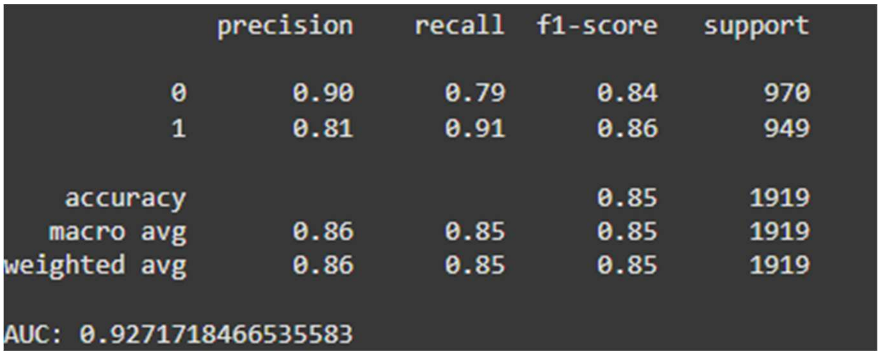




## Model

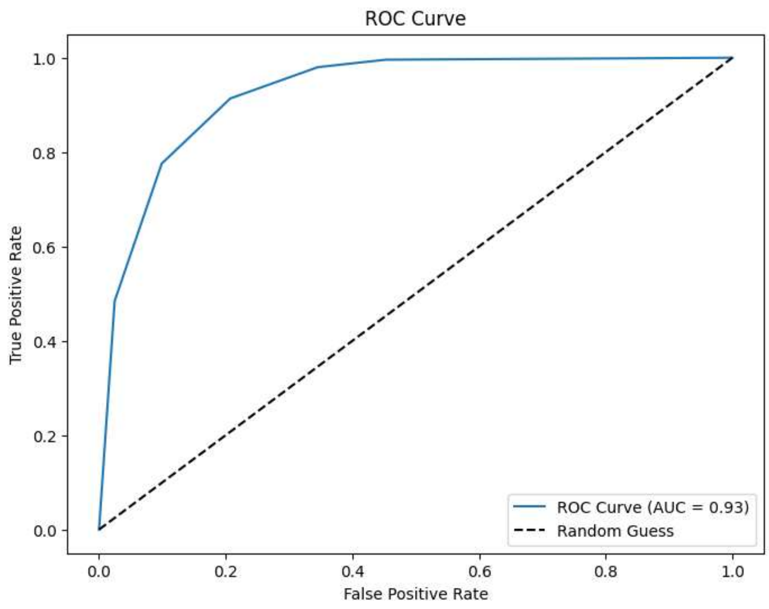
After our exploratory data analysis, we discovered the most important factors in student performance on tests. Most of the variables had very low correlations, but they did have some impact on test scores. Therefore, we still wanted to build a model to be able to predict whether a student will pass or fail based off the data collected about the student. To build a model for our data, we first split the data into training and testing sets at a 80/20 ratio. We used the KNN classifier to fit the data and then make predictions.

The model produced a Training Accuracy of 90.2% and a Testing Accuracy of 85.3%.

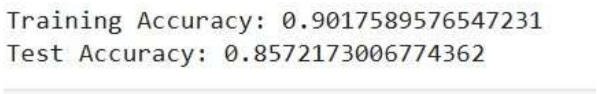


```
              precision    recall  f1-score   support

           0       0.90      0.79      0.84       970
           1       0.81      0.91      0.86       949

    accuracy                           0.85      1919
   macro avg       0.86      0.85      0.85      1919
weighted avg       0.86      0.85      0.85      1919

AUC: 0.9271718466535583
```

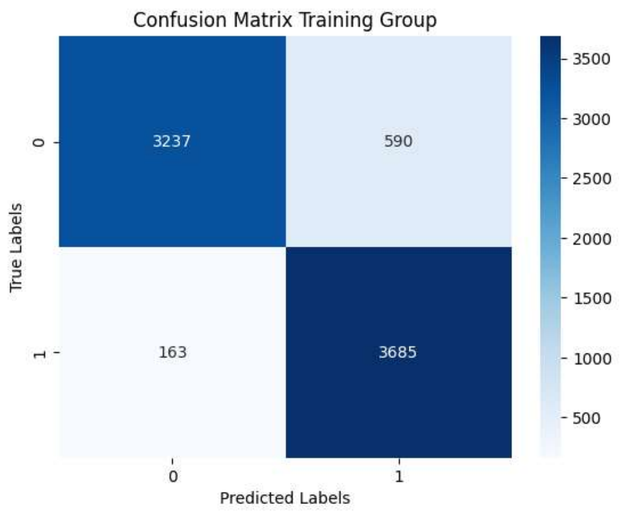
The AUC score of the model is 0.93. The ROC curve is ploted below and compares the True Positive Rate to the False Positive Rate. We see this curve is pretty far above the Random Guess line and shows us that our model is pretty good.

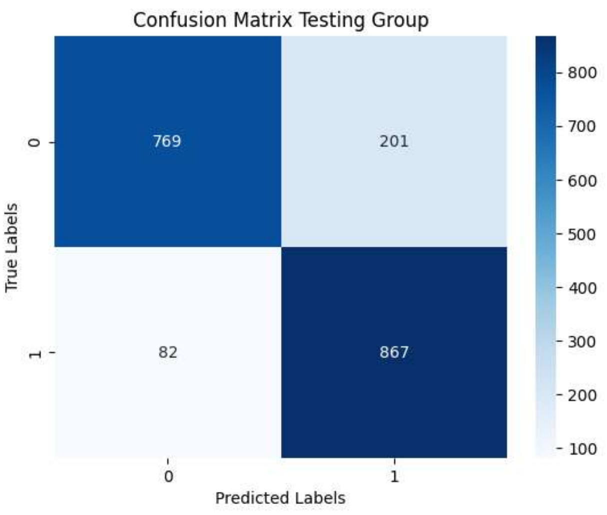


The Confusion Matrix of both the Training and Testing group also provides us with another great visualization of this model. We can see the accuracies are relatively similar.



```
Training Accuracy: 0.9017589576547231
Test Accuracy: 0.8572173006774362
```





Overall, our model had high accuracy and precision rates. The training set is slightly more accurate but not by enough to be considered overfitted.

### Conclusion and Insights

We began our project by asking what makes students pass or fail on tests, and if we can build a model to accurately predict whether or not students will pass or fail.

Throughout our exploratory analysis, we saw the most important factors in student success on tests are *Attendance* and *Hours Studied*. These two factors historically are considered the most important in the field of Education and were to be expected. It was unexpected how much higher the correlations of these two variables with passing the exam were, however. We saw the next most important factors were *Access to Resources, Parental Involvement,* and *Tutoring Sessions.* It was also observed that *Learning Disabilities,* and *Distance from Home* had the strongest negative correlations with our target variable.

These insights are valuable to educators and give educators something else they can focus on to try to help student performance. Parental Involvement, Access to Resources, and Tutoring Sessions are typically outside of a Teacher's control and only something a Parent can fix. But when Teachers have data backing the effectiveness of these factors on student performance, it is much easier to get parents on board with getting their students what they need to pass the test.

After we concluded our exploratory analysis, we built a model splitting our data into training and testing sets and using the KNN classifier. This model gave us a Training accuracy of 90.2% and a Testing accuracy of 85.3%. We felt our model was very accurate and did a good job of predicting whether students will pass or fail the test. Having a model like this would be helpful for Educators because when they get new students they could survey the student, input the data into the model, and get a very accurate idea of whether or not this student will pass the test.

References

Dataset: Student Performance Factors