**Project 2**

For this project I am attempting to answer a question very relevant to myself, and every other student who desires to get good grades, without having to spend huge amounts of unnecessary energy. The factors in this dataset cover a wide range of things, all of which will likely impact a student’s ability to succeed on an exam. Therefore, the goal is to test which factors have the greatest impact on a student’s performance. Using this information a student can then prioritize the aspects of their life which will have the greatest impact on their education, and therefore their future.

The dataset which I am working with takes the exam scores for many students, along with many other factors commonly associated with successful students (https://www.kaggle.com/datasets/lainguyn123/student-performance-factors). The factors included in the dataset are Hours\_Studied, Attendance (as a percentage, ranging from zero to one-hundred), Parental\_Involvement, Access\_to\_Resources, Extracurricular\_Activities (either yes or no), Sleep\_Hours, Previous\_Scores, Motivation\_Level, Internet\_Access (either yes or no), Tutoring\_Sessions, Family\_Income, Teacher\_Quality, School\_Type (either public or private), Peer\_Influence (either positive, negative, or neutral), Physical\_Activity (average hours of physical activity per week), Learning\_Disabilities (either yes or no), Parental\_Education\_Level (either high school, college, or postgraduate), Distance\_from\_Home (either near, moderate, or far), Gender, and Exam\_Score (the target). The columns Parental\_Involvement, Access\_to\_Resources, Motivation\_Level, Family\_Income, and Teacher\_Quality were all either low, medium, or high. The data set has data from 6,607 students, and takes 20 different factors into account.

The first step taken during preprocessing was to check for null values. Fortunately, this dataset was remarkably null-free, having only 3 columns with nulls and a total of only 235 nulls in total. Because nulls represented only a small portion of the dataset, any rows with null values were dropped entirely, using the dropna() function. The second preprocessing step was to take the column ‘Exam\_Score’ and determine how to convert it into a categorical variable, so that it could be used in a classification model as the target. In order to investigate the column more thoroughly the .describe() function was used, which showed that, on average, students scored a 67.2 on the exam. However, the goal of this study was to find not just how to be an average student, but one who excels. Therefore, the value for the 75th percentile was noted, which was 69. A new column was then created, called ‘Top\_Student’. This column was a simple boolean, where any student who scored higher than a 69 on the final exam was marked with the value “Yes”, and any student who got a score equal to or less than 69 was marked with the value “No”. Finally, the original ‘Exam\_Score’ column was dropped from the dataset, to avoid multicollinearity.

Two steps were taken to further understand the dataset. The first was to loop through every column in the dataset, creating a histogram for each column, in order to quickly check the columns for outliers, skews, and general trends. There were a few noteworthy takeaways from this step. Firstly, two columns were extremely skewed toward one of their two possibilities. Specifically, ‘Internet\_Access’ and ‘Learn\_Disability’ were both overwhelmingly skewed toward neurotypical students with adequate internet access. Because of this information I removed both columns from the dataset, as the model would likely struggle to predict any atypical values from those columns. Another internet note was that the graph for ‘Exam\_Score’ showed that the overwhelming majority of students scored between 60 and 75 on their final test. This supports the above decision to define any student who received a 70 or higher on their exam as a top student. For, while a 70 may not typically be considered an overwhelmingly great score on a final exam, this exam seems to be particularly difficult (for example, only 19 of the 6,378 students scored a 90 or higher). Therefore, it is necessary to adjust expectations and accept a score of 70 or higher as being worth pursuing. The second step taken to further understand the dataset was to create a correlation matrix between all the numerical columns and ‘Top\_Students’, which would give an early indication of which factors tend to be most impactful. This matrix showed that Attendance and Hours\_Studied were the most important variables, having correlation scores of 0.49 and 0.37. Conversely, Sleep\_Hours was shown to have a slight negative relationship with ‘Top\_Students’, though it was incredibly minimal at -0.001.

A decision tree classifier model was chosen for this project. The first step toward running this model was to one-hot encode the dataset, so that every categorical variable was divided into multiple boolean columns. This allows the decision tree model to easily understand categorical values, which are otherwise hard for it to interpret. This step resulted in 20 new columns being created, resulting in a new total of 38 columns. At this point the independent and dependent variables were established and the model trained. The trained model then outputted the decision tree. The way a decision tree shows its output is to create “branches” where levels of yes or no questions about the data are asked, with the most important ones at the top. For instance, the top question for this output was whether a student’s attendance was less than or equal to 65.6. The data is then divided by whether or not it is less than or equal to that value. This process is repeated multiple times, until data is sorted into highly correlated bins. For example, the model found that students who attend more than 65.6 percent of classes, have less or equal to 0.9 hours of physical activity per week, and participate in extracurricular activities have a 98% of being a top student.

A few evaluation metrics were used to test this model’s efficacy. Firstly, the model’s accuracy was 0.71, which means that it was correct 71% of the time. The precision for this model was only 0.52, however, the recall was better at 0.76. Interestingly, the model seemed to do better at predicting which students would not do well, with a precision of 0.91 and a recall of 0.76. Precision tells what percentage of the values predicted to be true actually were. This means that the current model does an okay job at predicting which students will be a top student, but there are a fair amount of false positives. However, if this model does not predict a student to be a top student, it is very unlikely that they will be one. Recall shows what percentage of the true values were correctly predicted to be true. The model did a fairly good job, though some improvement could certainly be made. These evaluation metrics were chosen as they work well with classification models, such as a decision tree. Paired with a confusion matrix, which is included in the Jupyter Notebook, this information is easy to understand at a glance and gives an accurate representation of the predicting power of a model.

The final results for this model showed that the most important factor toward predicting a student’s final exam grade was attendance, with a correlation of 0.496. This makes sense, as these students are likely more attuned to what their professor is teaching, as well as more engaged in general. The second most important factor was Hours\_Studied, with a correlation of 0.373. Again, this is logical, as students who put in the work to study the material would likely perform better in their final exams. Other important factors included a student’s previous scores, their access to resources, and their use of tutoring services. Conversely, the factors most likely to correlate with low final exam scores were a student having low access to resources, having low parental involvement, and having parents who’s education stopped after high school. So, in general, if a student wants to succeed on the final exam they should prioritize engaging with the schoolwork, school’s resources, and class. Additionally, it helps if their parents are involved in their education and have had first-hand experience with higher education (though this part is out of a student’s control).

The impact of this study is that students who aspire to succeed in their final exams can prioritize their efforts. It shows that students who put effort into their school work tend to reap the dividends of their work. So, attending classes, completing homework and additional studying, and a prior understanding of the topics tends to correlate highly with student success. A danger with this study is that it does show that a parent's education achievements tend to correlate with their child’s success. While this is true, it could potentially discourage students who did not grow up with their advantages to give up, when there is still a chance for them to beat the odds, if they engage with the work.

**References**

https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

https://chatgpt.com/

Note: ChatGPT was used in this assignment for specific code segments. Specifically, to assist in creating the graphs and modeling.

**Link to Code**

https://github.com/bigbadraj/Project-2-School-Correlation