HIGH SCHOOL EDITION
Journal of Student Research

Volume 10 Issue 2 (2021)
AP Research August

Feature Correlation with Student Education Performance

Ritvik Gupta[1] and Claire Gueneau[1#]

[1]Carnegie Vanguard High School, Houston, TX, USA
[#]Teacher and Mentor

ABSTRACT

The 21st century has seen the advent of the internet as well as the spread of increasingly powerful computer technologies. One of these new technologies is Artificial Intelligence and Machine Learning. These computer models assist in pattern recognition, task performance as well as prediction. One place where this technology can be used is Educational Data Mining. This study used these ML technologies on the Student Performance Dataset to see what features are correlated with high student academic performance. This study also utilized Feature Engineering to derive features that represent the interactions of different features from the original dataset in order to conduct further analysis. This study found that multiple different features such as parent relationship status, travel time between home and school, among others, had a positive correlation with student academic performance. Features such as past failures and increasing frequency of hanging out with friends after school was correlated with negative student academic performance. However, results with the ML models as well as Feature Engineering were inconclusive due to the results not having a high enough accuracy to merit analysis.

Introduction

The 21st century has seen the advent of the internet as well as the spread of increasingly powerful computer technologies. One result of this phenomenon is the development of sophisticated computer algorithms that can analyze vast amounts of data that wouldn't have been previously possible. This study of creating these computer algorithms that can improve themselves from data is called Machine Learning – a subset of artificial intelligence. These computer models are now present in our everyday lives as different technological services such as voice assistants (Siri, Bixby, Alexa), Google autocomplete, and facial recognition. These technologies have taken off and become more widespread thanks to the creation of open-source libraries that contain prebuilt modules of code that can be used for creating, testing, and application of virtually any model without heavy experience or knowledge in programming or data analytics. And with more powerful computers being made available for cheaper prices, more and more data can be analyzed with increasingly complex models.

One place where this technology can be used is in Educational Data Mining. Education Data Mining is "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in" (International Educational Data Mining Society, 2011). Every student has unique factors that affect their performance in school and so studying these unique factors can bring about some interesting results and applications. Current studies have found that a teen's brain is not fully developed until they turn 25 (Zimmer 2016). As a result, the years that a teen spends in secondary education (Grades 6-12) are vital as the unique factors they experience can shape their future life based on the factors they are subjected to as well as their performance in school.

This paper proposes to use machine learning to perform data analysis on student performance data to identify key features of students that correlate with high-grade performance in secondary education. By identifying what factors are most correlated with high student grade performance, educators can pay attention to these features of their students and proactively take measures to ensure that students who need more help will be identified and helped.

Literature Review

As machine learning is not a relatively new technology, there has been much research already done into analyzing student performance data utilizing machine learning models. However, many of these papers do not provide many results into what features could be the most impactful on student performance. These studies either only study the reliability of machine learning models for predicting student performance, do not study students in secondary education, or do not perform data engineering/manipulation to see if a combination of some features is more influential compared to individual features.

Ofori, Maina, and Gitonga (2020) wrote in their paper about previous studies that have been conducted on this subject. They focus on the different models that previous papers have used and talked about the accuracy and process time these models had. Many papers had conflicting results as shown in how Hussain, Muhsin, Salal, Theodorou, Kurtoğlu, and Hazarika (2019) found that Neural Network models performed the best whereas Belachew and Gobena (2017) found Naïve Bayesian to be the best performing model. Belachew and Gobena's (2017) results were supported by the paper written by Jayaprakash, Balamurugan, and Chandar (2018). However, Obsie and Adem (2018) found that Linear Regression and Support Vector Regression were better than Neural Networks. Also, Acharya and Sinha (2014) found that the decision tree class of algorithms was the best. Looking at these past studies, it is clear that there is no clear consensus as to what models are the best performing. While all these different models have high levels of accuracy, they all have around the same levels of accuracy which means that for Machine Learning analysis, it is not necessary to pick one or two models as no model is shown to be the best one to use.

It is no surprise that there is no clear consensus as to which model is the best as each study used unique data and models that are specialized to the data each study was working with. The main thing to note is that while no model was clearly found to be the best, the models in dispute all showed high levels of accuracy – with some showing around 98% accuracy. This fact means that these models can all be used as they have been established to be very accurate.

In a study conducted by Agrawal and Mavani (2015), they focused on using Neural Networks to predict student performance in an academic organization. They wrote "Present studies shows that academic performances of the students are primarily dependent on their past performances. Our investigation confirms that past performances have indeed got a significant influence over students' performance" which further shows that past student performance, such as secondary education performance, greatly influences future student performance. They also found that the performance of a neural network increases with dataset size which makes sense as neural networks are much more complicated and require larger sets of data to train on when compared to other models such as Linear Regression and Support Vector Machines. Another thing to note about this study is that they studied students that are in college, by which time the student's performance/work ethic in school is already established. These students are also studying advanced computer application courses, which means that the students in this study are more likely than not high performers and not a good mix of underperforming to high achieving students.

One key thing to mention about all these past studies is that a few to none of them have used Feature Engineering to create new features from the original data. This fact is important to note as it may not be the raw features that are correlated with student performance data, but a combination or mix of features that, when engineered from the original raw data, have a high correlation with student performance data.

Research Goals

The paper will attempt to cover these gaps left by previous studies:

1. Data is from students in secondary education as this is the time proactive measures are the most effective.

2. Feature Engineering is performed to create features that are respective of the relationship between different features.
3. Analyze which feature were most correlated with high student performance rather than which models are most accurate.

This study will utilize these steps to address the aforementioned gaps:

1. Use data that is recorded from students in secondary education.
2. Use Feature Engineering to create data that is derived from the original raw data.
3. Use a variety of Machine Learning models to see which features are most important from the most accurate models on this study's data.

Research Question

What features of students in the Student Performance Dataset are most correlated with high academic performance?

Methodology

This study will be done using the Python programming language. The reason Python was chosen over other popular programming languages, such as Java and C++, is because Python emphasizes code readability to enable clear, logical code for projects of any size. Another reason is that there are numerous free, open-source libraries - modules of prebuilt code - available for use in Python to assist in performing Machine Learning tasks. The PyCharm IDE, Integrated Development Environment, is where the code will be compiled and run. PyCharm is an open-source IDE that is one of the most powerful Development Environments for Python and Web Development (JetBrains).

For performing Machine Learning applications in Python there are two packages that are widely used – Pandas and Sklearn. Pandas is an open-source library that allows for fast and flexible data manipulation and analysis. It also contains a data management type called Data Frames which is the data format the dataset that will be imported into the environment will be stored as. Sklearn is another open-source library that contains numerous amounts of easy to implement Machine Learning models as well as data analysis and data processing tools. Another library that will be used is the Matplotlib library. This library will allow for the data to be plotted and displayed using graphics.

These are the libraries that this study will use to perform the data analysis. These libraries contain models that already are shown to perform at a high level of accuracy as explained and found by Ofori, Maina, and Gitonga (2020).

The dataset that is being studied is the Student Performance Data, which is publicly available for free from the University of California, Irvine. This dataset was donated by Associate Professor Paulo Cortez of the University of Minho, Portugal. This is the same dataset that has been used in previous studies such as Jayaprakash, Balamurugan, and Chandar (2018) which found that different Machine Learning models had a high accuracy on the data set. It contains 33 different features for 649 students.

The individual feature datum is stored as either an integer or text character depending on the feature type. The entire dataset is stored as a Comma Separated Value, CSV, file format.

The dataset is recorded from students in Portugal who are in secondary education from two different schools. The target data that is collected is their grades from their performance in two different courses, Math and the Portuguese language, over 3 different grading periods. For this study, the data of the students' grades in the Portuguese language was discarded as Portuguese is not a universal course taught globally whereas Math is. Feature Engineering will also be performed on the original data in order to create new data that is the interaction of different features between the original features.

Table 1. The preprocessed student related variables

FEATURE	DESCRIPTION
SEX	student's sex (binary: female or male)
AGE	student's age (numeric: from 15 to 22)
SCHOOL	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
ADDRESS	student's home address type (binary: urban or rural)
PSTATUS	parent's cohabitation status (binary: living together or apart)
MEDU	mother's education (numeric: from 0 to 4[a])
MJOB	mother's job (nominal[b])
FEDU	father's education (numeric: from 0 to 4[a])
FJOB	father's job (nominal[b])
GUARDIAN	student's guardian (nominal: mother, father or other)
FAMSIZE	family size (binary: ≤ 3 or > 3)
FAMREL	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
REASON	reason to choose this school (nominal: close to home, school reputation, course preference or other)
TRAVELTIME	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
STUDYTIME	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
FAILURES	number of past class failures (numeric: n if $1 \le n < 3$, else 4)
SCHOOLSUP	extra educational school support (binary: yes or no)
FAMSUP	family educational support (binary: yes or no)
ACTIVITIES	extra-curricular activities (binary: yes or no)
PAIDCLASS	extra paid classes (binary: yes or no)
INTERNET	Internet access at home (binary: yes or no)
NURSERY	attended nursery school (binary: yes or no)
HIGHER	wants to take higher education (binary: yes or no)
ROMANTIC	with a romantic relationship (binary: yes or no)
FREETIME	free time after school (numeric: from 1 – very low to 5 – very high)
GOOUT	going out with friends (numeric: from 1 – very low to 5 – very high)
WALC	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
DALC	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
HEALTH	current health status (numeric: from 1 – very bad to 5 – very good)
ABSENCES	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

[a] 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

[b] teacher, health care related, civil services (e.g. administrative or police), at home or other.

Terminology

The Correlation Matrix is calculated using the Pearson correlation coefficient method. This method will return the measure of linear correlation between two different sets of data. The Machine Learning models that are being used are Linear Regression, Decision Tree Regressor, Multi-layer Perceptron Regressor, and Support Vector Regressor. These models differ in their architecture and process, but they will all be scored using the same method. All the models are scored using the coefficient of determination or R-Squared.

Linear Regression works by assigning different weights and biases to each feature. From there, it will multiply the numeric value of each feature against the weight and add the bias. After doing this for each feature, it will compute the sum of all these values. Next, it compares the value to the actual result that is needed. It will iterate over the training data changing the values of the weight and biases assigned to each feature by computing the Least Squares Loss value until it finds the set of values with the highest accuracy. The best visualization of this is with the linear equation y=mx+b where “x” is the feature value, “m” is the weight assigned to the feature, and “b” is the bias that is a constant added to the end.

Decision Trees Regression works by creating nodes that branch off into other nodes like a tree branch. The model will train to figure out what the case for each node should be and then what the final value should be by iterating over the training data adjusting the values of each node and adjusting its architecture until it arrives at a Decision Tree with the best accuracy. It predicts values by looking at the testing value inputs and running them through the Decision Tree until it gets a predicted value.

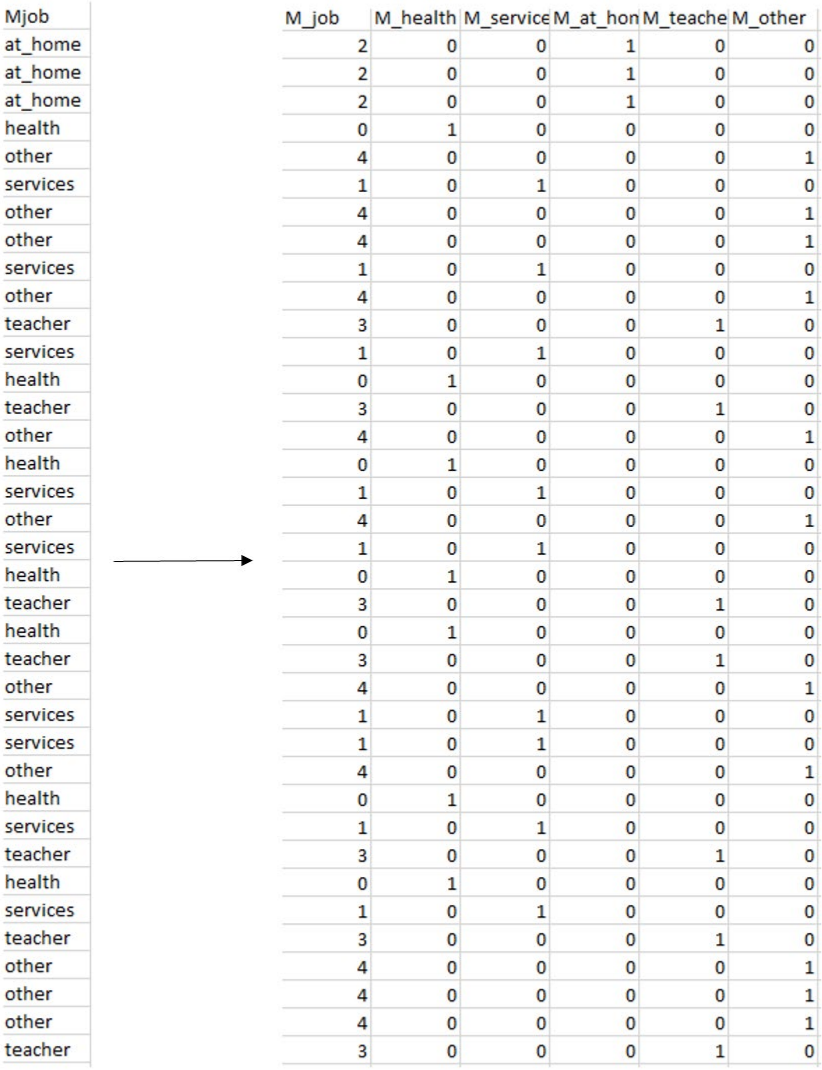
Support Vector Regressors work by first plotting the training data onto an n-dimensional graph (where n is the number of features in the training dataset). This data is then plotted against the target value, which for this study is the student average grade performance. The model will then come up with an equation for a line that best fits the data points by adjusting itself through iterations through the data until it gets the highest accuracy. The type of line that the model will come up with depends on the kernel that it is initialized as. This study will use the “poly” kernel which means that the model will try to create a line in a polynomial expression to best fit the data. The model will then predict values by plotting input data onto the line and seeing the value that is computed. The difference between Support Vector models and Linear Regression models is that Linear Regression aims to find the line of best fit using the Least Squares Loss whereas Support Vectors an epsilon insensitive loss function (Sikder 2019).

It is important to note that the accuracy of these Machine Learning models may not be at the same level as those in previous papers. This is because each study had its own unique approach to data processing and model creation. As Ofori, Maina, and Gitonga (2020) have found that there are multiple models that seem to do well in many studies, this study will use those same model architecture types.

Methods

In order to feed the Student Performance Dataset data into the Machine Learning algorithms, the data must be formatted in order to be fit and analyzed by the Sklearn Machine Learning models. To do this, the data must first be edited by a spreadsheet editing software like Microsoft Excel.

Once downloaded, the data must be formatted and edited so that all text data is converted to numeric values as the Sklearn Machine Learning models cannot function on text data. For some of the input data, it can be turned into 1s or 0s due to it being in a binary data format, such as sex, and stored in the original column. For categorical data with multiple classes such as Residence Type or Mother Profession, the data must instead be converted using the one-hot-encoding method. This method will take the input column and look at all the different values that appear in the column and make a new column for each unique value. From there, the method will iterate through the value in each row of the original column and assign a 1 to the column with the matching name, and a 0 to the rest of the columns in the same row. Using this method, the original categorical data information will be preserved and not lost since the data is simply being represented in a way that the Machine Learning model can understand. To perform Feature Engineering on the dataset, the Sklearn library will be utilized. A copy of the Pandas data frame of the aforementioned data will be made. From there, the data will be passed into a function that will multiply each feature every other feature in the dataset and create new columns with this data.



Mjob
at_home
at_home
at_home
health
other
services
other
other
services
other
teacher
services
health
teacher
other
health
services
other
services
health
teacher
health
teacher
other
services
services
other
health
services
teacher
health
services
teacher
other
other
other
teacher

M_job	M_health	M_service	M_at_hon	M_teache	M_other
2	0	0	1	0	0
2	0	0	1	0	0
2	0	0	1	0	0
0	1	0	0	0	0
4	0	0	0	0	1
1	0	1	0	0	0
4	0	0	0	0	1
4	0	0	0	0	1
1	0	1	0	0	0
4	0	0	0	0	1
3	0	0	0	1	0
1	0	1	0	0	0
0	1	0	0	0	0
3	0	0	0	1	0
4	0	0	0	0	1
0	1	0	0	0	0
1	0	1	0	0	0
4	0	0	0	0	1
1	0	1	0	0	0
0	1	0	0	0	0
3	0	0	0	1	0
0	1	0	0	0	0
3	0	0	0	1	0
4	0	0	0	0	1
1	0	1	0	0	0
1	0	1	0	0	0
4	0	0	0	0	1
0	1	0	0	0	0
1	0	1	0	0	0
3	0	0	0	1	0
0	1	0	0	0	0
1	0	1	0	0	0
3	0	0	0	1	0
4	0	0	0	0	1
4	0	0	0	0	1
4	0	0	0	0	1
3	0	0	0	1	0

Figure 1. It shows the before and after result of performing One-Hot-Encoding process upon the "Mjob" feature and the resulting feature values.

After preprocessing this data, it needs to be saved as a Comma Separated Values file type (CSV). A CSV file is a delimited text file that uses a comma to separate values. From there, the data needs to be imported into the Python development environment and stored as a Pandas data frame. Pandas provides a method to easily load CSV files and automatically convert them into a data frame while also keeping the data original data structure and column names.

The next step is to plot a correlation matrix of each feature against the student grade performance. To do this, the Pandas data frames have a built-in method called .corr() which will calculate a correlation matrix of the input data. Using this method, the correlation matrix can be stored in a variable for reformatting and plotting. Since the feature of interest is the student grade performance, specifically, the correlation matrix values of each feature against the student grade performance, a new variable is needed to be created to store only the values of each feature against the student grades as the current correlation matrix data contains values of each feature against each other feature.

After this, the correlation matrix data that consists only of the feature of interest needs to be sorted from greatest to least, to see which features have the highest correlation to student grades. Pandas has another easy-to-use method to sort the values of a data frame from greatest to least. To plot a heatmap of the correlation matrix values, the data can be fitted into a Seaborn heatmap method to create a heatmap of the top feature values which can then be shown using the Matplotlib package's Pyplot library.

After this, the original data can be modified using Feature Engineering using Sklearn to create features that are derived from the original data that show the relationship between different features. To do this, the original excel file needs to be broken into two different files with the new file containing only the column of the student performance grades with it being removed from the original excel file. After this, both excel files need to be saved as a CSV file and imported into Python. From there, they can both be converted into a Pandas data frame. Next, the data frame containing the student features can be fitted to a method from Sklearn that will return the feature-engineered data. After this, the previous steps can be repeated to create a correlation matrix of each feature against the student performance grade.

Once these steps are done, the original data and the feature engineered data will be split into a training and testing set with a ratio of 3:1. The testing data set will be fitted to different Machine Learning models provided by Sklearn. After the Machine Learning models have trained on the training data, the testing data set will be used to find the accuracy of the different models on the training data. Whichever models have the highest accuracy can then be further analyzed by printing the correlation coefficients of each feature in the model to see which features had the most "weight".

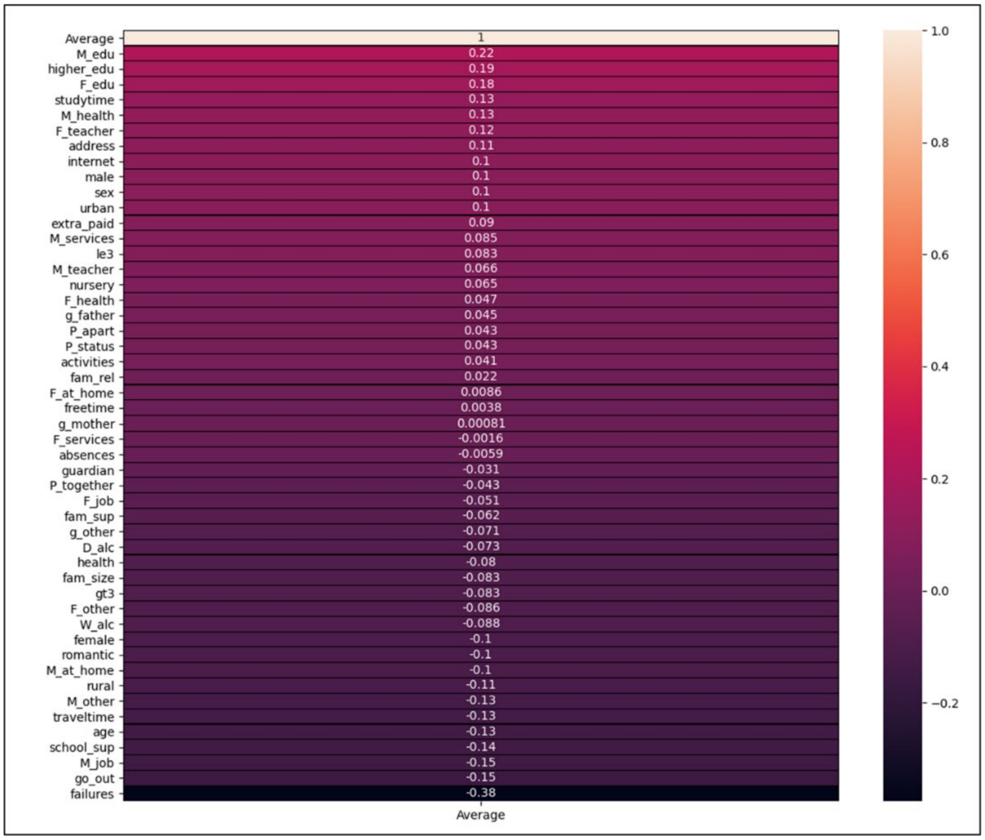


Figure 2. It shows the correlation matrix values of the raw dataset plotted against the student's average academic performance in descending order.

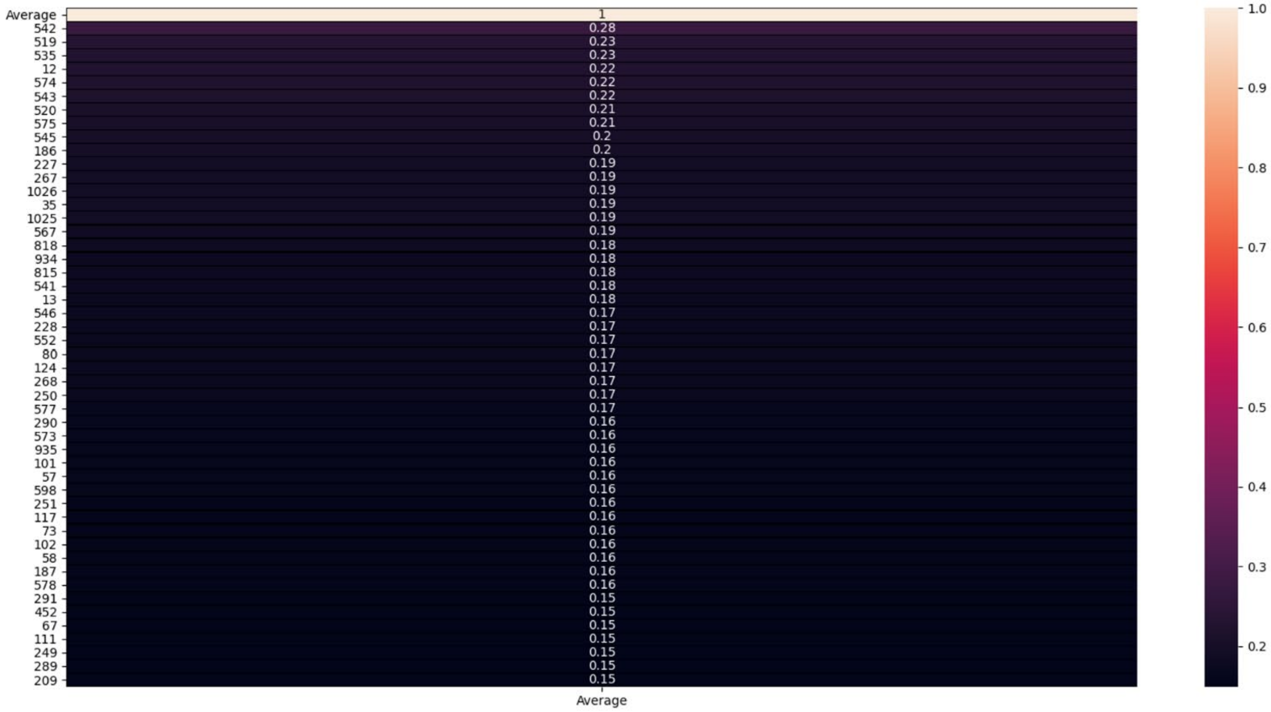


Figure 3. It shows the correlation matrix values of the raw dataset plotted against the student's average academic performance in descending order.

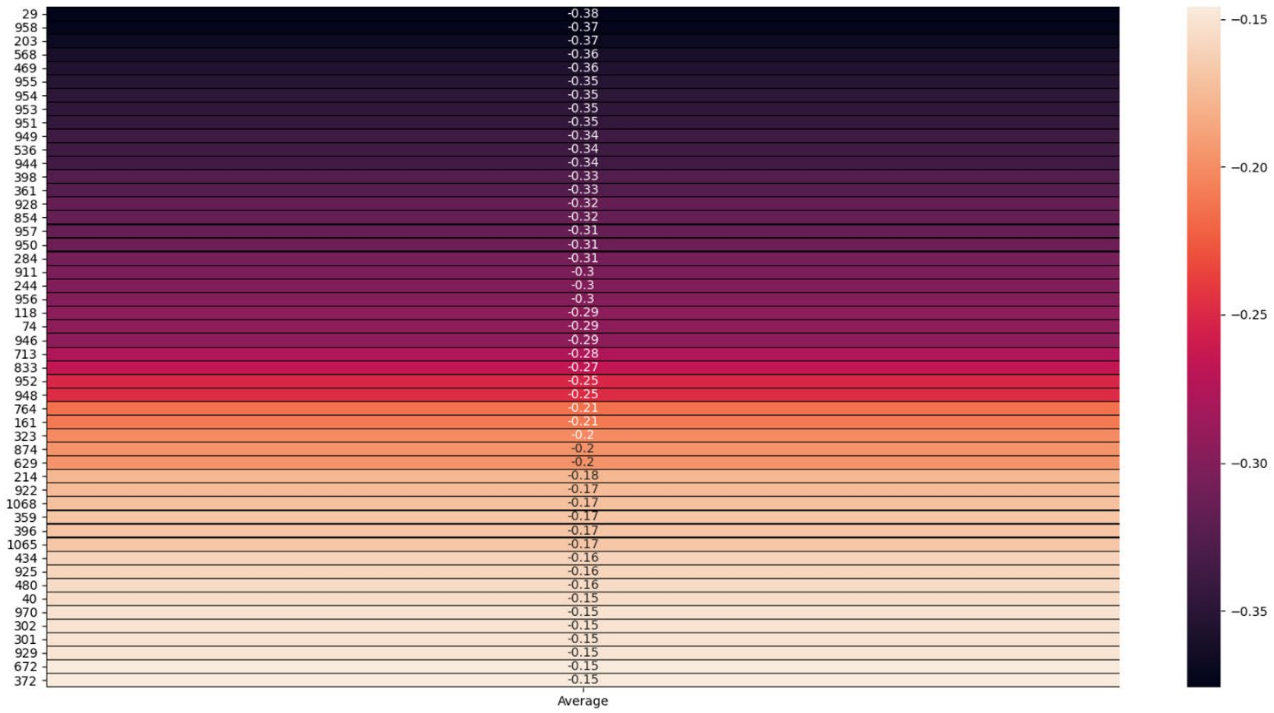
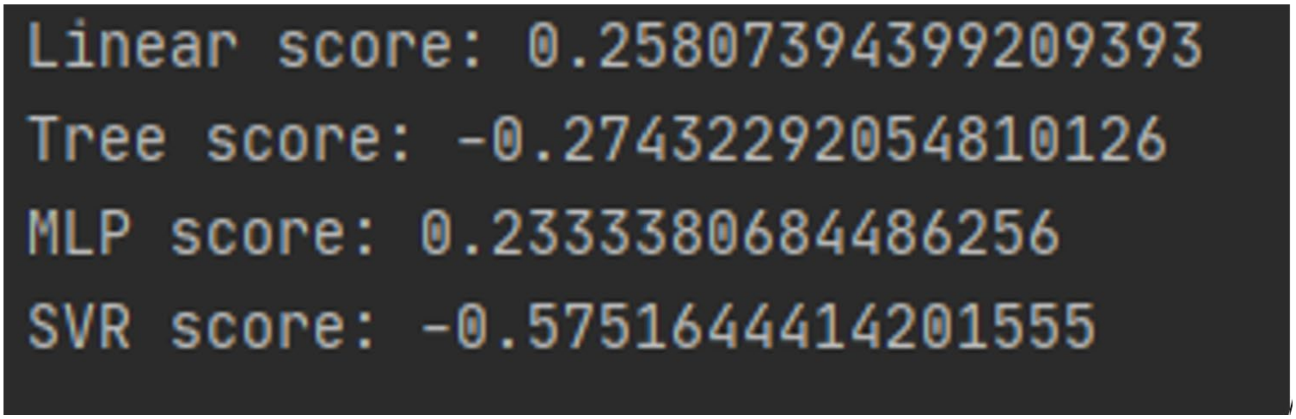
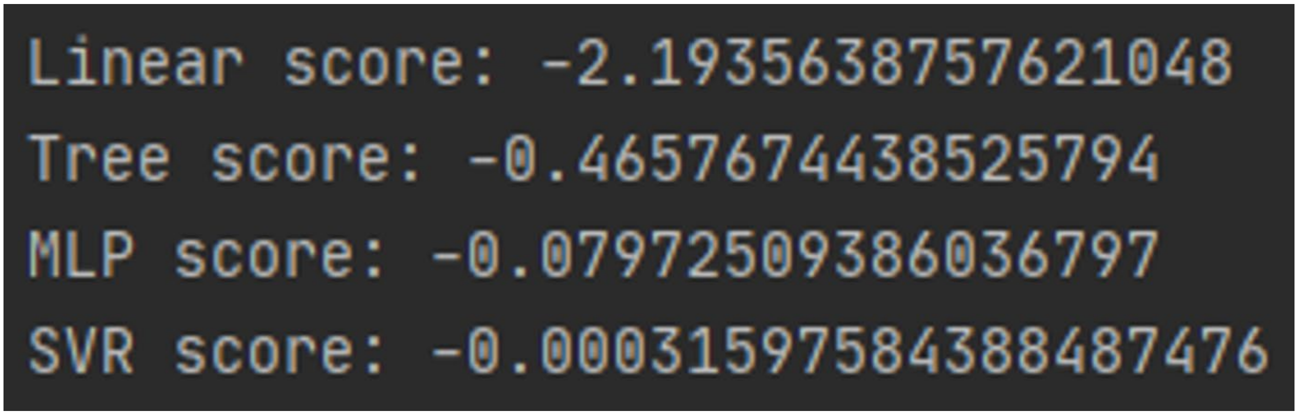


Figure 4. It shows the correlation matrix values of the feature-engineered dataset plotted against the student's average academic performance in ascending order.



```
Linear score: 0.2580739439920939З
Tree score: -0.2743229205481012б
MLP score: 0.233338068448б256
SVR score: -0.5751644414201555
```

Figure 5. It shows the accuracy scores of the Machine Learning models on the original dataset.



```
Linear score: -2.1935638757621048
Tree score: -0.4657674438525794
MLP score: -0.07972509386036797
SVR score: -0.0003159758438848747б
```

Figure 6. It shows the accuracy scores of the Machine Learning models on the feature-engineered dataset.

Results

The cutoff for which features are deemed as having a significant correlation to student performance is a correlation value with an absolute value greater than 0.1.

As seen in Figure 2, the features mother's education level, student's desire for higher education, and father's education level all had the most positive correlation with student performance. The features of number of past class failures, frequency of going out with friends, and extra educational school support had the most negative correlation with student performance. Figure 1 shows that there was a range of correlation coefficient values from 0.22 to -0.38 (Note that the value of 1.0 is the target feature plotted against itself, which always be perfect which is why it is being ignored). These values can range from 1.0 to -1.0 with the farther a value is from 0, the greater the correlation it had on the student performance grade. The labor participation rate for women is around 10 – 15% lower than for males in Portugal (The Global Economy 2020). This fact means that more likely than not the mother is staying at home to care for their child rather than the father which means that they spend more time with their child which would make the child look and pick up habits from their mom. This idea could explain why "M_edu", which is the education level of the mother where the higher the value the longer the mother's education level is, is most highly correlated with student performance as the mother's education level increases, their lifestyle and choices are more informed and educated, which would more likely than not reflect on their child's school performance. "higher_edu" describes whether or not the student wants to pursue high-level education such as college. This would make sense as to why "higher_edu" is positively correlated to student performance as students who are wanting to pursue higher education are more likely to pay attention in class and aim to have a high performance in school compared to students who do not plan to pursue higher education. “F_edu” represents the education level of the father, with a higher value corresponding to a higher level of education. This would explain why it is correlated with student performance as the higher educated the father is, the more educated and informed the lifestyle of their child’s life will be. “M_health” is whether or not the child’s mom has a job in the healthcare-related service. This could be due to the fact that a mother with a job in the health care service industry is able to use their job experience to help care for the mental and physical wellbeing of their child which in turn would make the child less likely to have a lower performance in school. The “studytime” feature has a positive correlation as a student who studies will perform better at school. “F_teacher” is whether or not the father of the student works as a teacher. This is unsurprising as having a teacher for a parent would more likely than not boost a child’s performance. What is surprising though is while the correlation value for the feature regarding the father is 0.12, the correlation value for whether the student’s mother works as a teacher, is only 0.066 – around half that of the value of the father. No theory or reason is proposed to explain this finding. Lastly, the “address” feature is whether the student lives in a rural or urban home type. The value of 0.11 indicates that a student who lives in an urban dwelling is correlated to have a higher performance in school. This could be because students who live in an urban dwelling have more wealthy parents compared to students living in rural dwellings. And since wealth is more correlated to student performance as found by Georgetown University [6], this could also explain the value for the fact that the “rural” feature has a correlation value of -0.11.

For the negatively correlated features, the number of past class failures is the most negatively correlated feature with an unsurprisingly strong coefficient of -0.38. This makes sense as past failures in classes are one of the indicators that can be used to predict future class performance as found by Agrawal and Mavani (2015). The greater the number of classes the student has failed in the past, the more likely their performance in school is not that high. The frequency at which the student goes out and spends time with friends outside of school is negatively correlated – though not with a great magnitude – as the more time a student spends outside with friends, the less time they have for school work. This would explain the inverse relationship with the feature “studytime”. Another reason may be that students that care more about their grades are less likely to go out as they prefer to spend time studying or doing extra-curriculars. “school_sup” is whether or not the student’s school provides extra educational support. With this having a negative value, it could mean that schools that offer extra educational support have low average student performance prompting the school to host these extra opportunities. The feature “rural” indicates whether a student lives in a rural address or not. The fact that it is negatively correlated with student performance is unsurprising as it explains the correlation between “address” and student performance. The feature “traveltime” is the home-to-school travel time. A negative correlation indicates that with greater travel time, a student is more likely to perform worse in school. No explicit reason was thought to explain this but this could tie into the “address” and “rural” features as urban addresses are more likely closer to school compared to rural addresses. The last feature “M_other” is that the job of the mother is not as a teacher, health care, civil service, or as a stay-at-home mom. No reason is proposed for this finding.

Due to the Feature Engineered data having such a high number of features with the correlation values being virtually the same for most of them, only the top 3 positively and negatively correlated features will be analyzed. The top 3 most positively correlated features in the Feature Engineered Dataset are feature numbers 542, 519, and 535. Feature 542 is the interaction of features “P_together”, whether or not the parents are still together, with the feature “activities”, does the student participate in extracurricular activities. No reason is proposed to explain this finding. Feature 519 is the values of feature “P_together” squared. No reason is proposed for this finding as the original values for “P_together” are binary meaning that squaring them wouldn’t change any of the values, which should give it the same correlation score as the original dataset - not change the sign of correlation as well as the magnitude of the correlation. Feature 535 is the interaction of features “P_together” and “g_other” – the guardian of the child is the mother. No reason is proposed for this finding.

Feature numbers 29, 958, and 203 are the highest negatively correlated features. Feature 29 is “traveltime”. This would make sense as this feature was also one of the most negatively correlated features for student performance. Feature 958 is the interaction of “traveltime” and “D_alc” – the workday alcohol consumption of the student. No reason is proposed for this finding. Feature 203 is the interaction of features “male” – whether the student is male or not – and “traveltime”. No reason is proposed for this finding.

For the results of the original data being fed into the different Machine Learning models, only 2 models – Linear Regression and Multi-layer Perceptron Regressor – had a positive accuracy score. Negative accuracy scores – Decision Tree and Support Vector Regressor model scores – mean that the model is performing very badly and is not optimized for the data being fed into it. For the Linear Regression and Multi-layer Perceptron Regressor models, even though they have positive accuracy scores of 25.8% and 23.3% respectively, these scores are too low to justify further analysis. The previous studies outlined in Ofori, Maina and Gitonga (2020) had Machine Learning models with accuracy scores that were much greater – most above 80%.

For the results of the Feature Engineered data being fed into the Machine Learning models, all accuracy scores were negative meaning that they cannot be used for further analysis. Even though the Support Vector Regressor model had a score that was very close to 0, -.03% to be exact, the scores are way too low to warrant further analysis.

Conclusion

The goal of this paper was to find the correlation of different features, in the Student Performance Dataset, against the grade performance of students in the dataset. This study utilized data analysis through computer algorithms. Analysis of the original dataset found that the education level of the parents, the higher the amount of weekly study time the student engaged in, the student living in an urban address type, having the father working as a teacher, as well as the student wanting to pursue higher education (i.e., College, University), was positively correlated with student performance in school. Also, this study found that the number of failures in previous classes by the student, the longer the travel time between the school and the home, the more the student went out with friends after school, the presence of extra educational opportunities offered by the school, living in a rural address type, and having a mother working in an occupation other than the health services, civil service, teaching, or being a stay at home mom were all negatively correlated with student performance. When the data was fed into different Machine Learning models, the results were inconclusive as the models didn’t achieve a high enough level of accuracy to justify further analysis and explanation. Only the Linear Regression and Multi-layer Perceptron Regressor model were able to work with the data with the Support Vector Regressor and Decision Tree Regressor model unable to work at all on the data.

Analysis on the Feature Engineered dataset showed that having parents that live together was positively correlated with student performance. In addition, the longer the travel time between the home and the school was found to be negatively correlated with student performance in school. When this data was fed into the Machine Learning models, none of the models had an accuracy score that justified further analysis into the models.

Future Directions

With these results, this study suggests that further analysis be conducted into the relationship between the features of the education level of the parents, relationship/living conditions among the parents, occupations of the parents, travel time between the home and school, as well as the residential address type of the student to the effect on student performance in school.

The results found by this study suggest that educators should pay more attention to the aforementioned factors in the “Conclusion” section for each student to determine which students may preemptive intervention to ensure that their performance at school is not hampered in any way.

More analysis will also need to be done into fitting the Student Performance Dataset to Machine Learning models. As this study performed Regression analysis, the models were not able to perform well most likely due to the fact that most of the data was of categorial nature and not quantitative values. The Linear Regression and Multi-layer

Perceptron Regressor models were the only models found to fit somewhat to the data so further analysis using these models is suggested.

Lastly, further analysis into utilizing Feature Engineering on this dataset is suggested as this study could not get useful results from the Feature Engineered dataset.

Acknowledgements

I would like to thank my AP Research teacher Ms. Claire Gueneau. She helped me to refine my initial research question as well as pairing me with other students in the same field as me so that my peer-review for the revising and editing part gave me lots of insightful feedback. I would also like to thank my cousin-in-law Mr. Pankaj Gupta for showing me how to apply ML in Educational Data Mining to complete this study.

References

Acharya, A., & Sinha, D. (2014). Early Prediction of Students Performance using Machine Learning Techniques. *International Journal of Computer Applications*, *107*(1), 37–43. https://doi.org/10.5120/18717-9939

Agrawal, H., & Mavani, H. (2015). Student Performance Prediction using Machine Learning. *International Journal of Engineering Research And*, *V4*(03). https://doi.org/10.17577/ijertv4is030127

Belachew, E. B., & Gobena, F. A. (2017). Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(2), 46–50. https://doi.org/10.23956/ijarcsse/v7i2/01219

Educationaldatamining.org, Educational Data Mining, https://educationaldatamining.org/.

Georgetown University. (2020, August 18). *Born to Win, Schooled to Lose: Why Equally Talented Students Don't Get Equal Chances to Be All They Can Be*. CEW Georgetown. https://cew.georgetown.edu/cew-reports/schooled2lose/.

The Global Economy. (2020). *Portugal Female labor force participation - data, chart*. TheGlobalEconomy.com. https://www.theglobaleconomy.com/Portugal/Female_labor_force_participation/.

The Global Economy. (2020). *Portugal Male labor force participation - data, chart*. TheGlobalEconomy.com. https://www.theglobaleconomy.com/Portugal/Male_labor_force_participation/.

Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodoru, P., Kurtoğlu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. *International Journal of Emerging Technologies in Learning (IJET)*, *14*(08), 4. https://doi.org/10.3991/ijet.v14i08.10001

Jayaprakash, S., Balamurugan E. & Chandar, V. (2018). Predicting Students Academic Performance using Naive Bayes Algorithm, BlueCrest College Accra, Ghana.

JetBrains. (n.d.). *JetBrains Delights the Python Community with a Free Edition of its Famous IDE, PyCharm 3.0: The PyCharm Blog*. JetBrains Blog. https://blog.jetbrains.com/pycharm/2013/09/jetbrains-delights-the-python-community-with-a-free-edition-of-its-famous-ide-pycharm-3-0/.

Obsie, E., & Adem, S. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*, *180*(40), 39–47. https://doi.org/10.5120/ijca2018917057

Ofori, F., Maina, E., & Gitonga, R. (2020). Using Machine Learning Algorithms to Predict Studentsâ€™ Performance and Improve Learning Outcome: A Literature Based Review. Journal of Information and Technology, 4(1), 33 - 55. Retrieved from https://stratfordjournals.org/journals/index.php/Journal-of-Information-and-Techn/article/view/480

Sikder, S. (2019, November 28). *What is the difference between support Vector regression, using a linear kernel and least squares linear regression?* Quora. https://www.quora.com/What-is-the-difference-between-support-Vector-regression-using-a-linear-kernel-and-least-squares-linear-regression.

Zimmer, C. (2016, December 21). *You're an Adult. Your Brain, Not So Much.* The New York Times. https://www.nytimes.com/2016/12/21/science/youre-an-adult-your-brain-not-so-much.html.