Study Smarter, Not Harder: Using SHAP for Actionable Educational Insights

Finn Alberts (852685751) [1]

1. Introduction

Student success is a critical area of focus in education, as improving performance can have significant long-term impacts on individual opportunities and societal progress. For example, improving academic performance can lead to increased job opportunities, higher lifetime earnings, and greater societal contributions through innovation and knowledge sharing. With the increase of datasets and use of machine learning algorithms in recent years, one might wonder if we can use educational data to improve student success by finding what the key contributions are.

Previous research in this field has focused on trying to predict student success using this data, with a variety of machine learning algorithms. For example, research by (Ouatik et al., 2022) used k-Nearest Neighbors (KNN), C4.5, and Support Vector Machines (SVMs) to predict student success with an accuracy of 87.32% for SVMs. Other papers have been published as well, such as research by (Al Mayahi & Al-Bahri, 2020) where they focused on predicting student success based on earlier results or the research by (Cortez & Silva, 2008) which also incorporates non-academic factors, such as travel time, parental jobs, or health status.

Additional research has been conducted on finding optimal models and hyperparameters using an automated approach (Zeineddine et al., 2021). This shows that a lot of work in the prediction of student success has already been done, which is further illustrated by the literature review of (Alsariera et al., 2022).

Most of the aforementioned research discuss what factors have the highest influence on the final prediction, indicating what factors are most important for student success. SHAP, a technique to make machine learning predictions interpretable (Lundberg & Lee, 2017), has also been used to achieve this goal, as shown by the research of (Özkurt, 2024). However, these studies stop at explaining predictions or identifying important features, leaving students without clear guidance on how to improve their performance.

2. Goal

In this project, our objective is to provide actionable advice to students, based on results from the SHAP-analysis of key contributing factors. We will look into factors on which students have actual influence and can help them make a positive change. Our aim is to make clear and interpretable claims, based on the factual results by SHAP.

We can split this goal into two smaller objectives:

1. Our primary objective is to identify actionable, student-controllable factors influencing success, using SHAP to ensure our analysis is interpretable and evidence-based.
2. Translate key contributing factors from SHAP analysis into clear, actionable advice for students, empowering them to improve their academic performance.

By translating SHAP insights into actionable advice, this project seeks to empower students and educators to make data-driven decisions for academic success.

The remainder of this project is organized as follows: In Section 3, we describe the dataset used for analysis and how cleaning was done. Section 4 details the methodology applied. Results of the SHAP analysis and the impact of various features are presented in Section 5. Finally, we wrap up with the discussion and conlusion in Sections 6 and 7.

3. Dataset

3.1. Used dataset

For this project, we use a dataset of two Portuguese high schools. This dataset includes student's grades for the first, second and third periods (G1, G2, and G3) for courses in math and Portuguese, as well as a set of demographic, social and school related features. The dataset was retrieved from Kaggle[1]. A full description of all features can be found in Appendix A.

The data in this dataset was collected by Cortez and Silva, who used it to try to predict student success in period 3 in terms of both classification (binary and with five levels) as well as regression (Cortez & Silva, 2008). In their paper, they try different setups of features, where setup A includes both G1 and G2, setup B includes only G1 and setup C includes neither G1 and G2. They use a naive estimator, neural network, support vector machine (SVM), decision tree, and random forest. The experiments are conducted

[1] https://www.kaggle.com/datasets/whenamancodes/student-performance (please note that the dataset is downloaded with a .csv extension, but is actually a .xlsx file)

for Portuguese and math separately, resulting in 2 courses · 3 setups · 3 output types · 5 models = 90 experiments.

For the random forests and decision tree setup, the paper shows the relative importance of each feature as measured by the random forest algorithm (Breiman, 2001). No further analysis on contributing factors is done, which is what this project will dive into deeper using SHAP (Lundberg & Lee, 2017).

3.2. Data cleaning

The dataset does not include any missing values or outliers and we can thus follow the same data cleaning setup as the original paper. This consists of three steps:

1. Creation of dummy variables
2. Normalize all features (excluding target variables G1, G2, and G3)
3. Create target variable for binary classification (*pass* if $G3 \geq 10$) and five-level classification[2]

4. Methodology

The methodology of this project contains six main steps:

1. First, we try to reproduce the results from the original paper.
2. Second, we calculate the SHAP-values for each of these models.
3. Based on these SHAP-values we take the average over all models.
4. For each feature where the normalized SHAP-value is ≥ 0.05 we loop over all possible values for that feature and set that value as a threshold to split the data into two groups.
5. We conduct a t-test of independence.
6. Based on the results we make a translation to actionable advice for students.

In the following subsections we will dive deeper into each step.

4.1. Reproduction of results from original paper

To be able to calculate the SHAP-values for the models from the papers, we first need to reproduce the results. We therefore follow the approach from (Cortez & Silva, 2008) as closely as possible.

In their paper they compare the different machine learning models. For this project we will only look at the best performing model (excluding the naive predictors, as it was used as a baseline measurement in the original paper) for each combination of course, setup and output type, meaning we will have 18 models. See Table 1 for which model performed best for each combination.

	Math	**Portuguese**
Binary Classification		
Setup A	Random Forests	Decision Trees
Setup B	Decision Trees	Random Forests
Setup C	SVMs	Random Forests
Five-Level Classification		
Setup A	Decision Trees	Decision Trees
Setup B	Decision Trees	Decision Trees
Setup C	Random Forests	Random Forests
Regression		
Setup A	Random Forests	Random Forests
Setup B	Random Forests	Decision Trees
Setup C	Random Forests	Random Forests

Table 1. Best model per combination of course, setup, and output type

We run all of these models using Python and the SciKit-learn library[3], and set the hyperparameters to the same values as mentioned in (Cortez & Silva, 2008). For decision trees we thus use the default parameters. For random forests we use the default parameters, but set the number of estimators to $T = 500$. For SVMs, we use a polynomial kernel. For the degree of the polynomial and for the value of γ we do an internal grid search with $degree = \{0, 2, 4, 6, 8\}$ and $\gamma = \{2e^{-9}, 2e^{-7}, 2e^{-5}, 2e^{-3}, 2e^{-1}\}$.

Similarly to (Cortez & Silva, 2008), we run (stratified) K-fold cross-validation with $k = 10$ folds, where we train each fold $m = 20$ times, resulting in $k \cdot m = 10 \cdot 20 = 200$ runs/model. For each of these models we calculate the accuracy for classification or root mean squared error (RMSE) for regression. We average these to get a final evaluation metric. In the case of hyperparameter tuning, we calculate the best hyperparameters for each of these 200 models separately, where we use 20% of the training set for evaluating the hyperparameters (validation set). The best hyperparameters are then used when evaluating the model.

[2]For five-level classification the levels are based on the G3 score as follows: level 1: 16-20, level 2: 14-15, level 3: 12-13, level 4: 10-11, and level 5: 0-9.

[3]https://pypi.org/project/scikit-learn/

4.2. Calculating SHAP-values for each model

During the training and evaluation of the model, we also calculate the SHAP-values. We use the Python SHAP library[4] to accomplish this. For all models we use the default explainer, except for SVMs where we use the kernel explainer as SVMs do not work with the default one.

The final result is a $p \times q \times r$ matrix containing the SHAP-values, where p is the number of rows for that fold, q the amount of features, and r the amount of classes (for regression 1, binary classification 2 and for five-level classification 5).

4.3. Averaging SHAP-values

After calculating all SHAP-values we average them multiple times:

1. First, we average over all rows per matrix.
2. Second, we average for the $k \cdot m = 200$ models per combination of setup, course, and output type.
3. Third, we average over all classes. In the case of binary classification, we simply drop one of the two values, as the SHAP-value for class A is the negation of the SHAP-value of class B. This results us in a vector for each model with one SHAP-value per feature.
4. Lastly, we take the average over all 17 different models[5]. We take this average over the normalized (per row) SHAP-values. By normalizing the SHAP-values, we make sure the row's sum always equals 1, ensuring compatibility between the different models. For G1 and G2, we exclude rows where they were not inputted as feature (setups A and B).

This results in a q-dimensional vector with a (normalized) SHAP-value per feature.

4.4. Looping over threshold values

For all features where the normalized SHAP-value ≥ 0.05, we loop over all possible values for this feature and set it as a threshold to split the complete dataset in two. By only taking features where the normalized SHAP-value ≥ 0.05, we make sure we only look into the features with high impact.

[4] https://pypi.org/project/shap/

[5] Due to computational constraints, SHAP-values for the SVM model were not calculated, meaning the final averaged results are based on 17 models instead of 18.

4.5. Conducting t-test

We perform a t-test of independence to compare the mean G3 values between groups split by the feature threshold, determining if the difference is statistically significant. Because we are conducting multiple experiments with different thresholds we have to apply the Bonferroni correction, resulting in $alpha = \frac{0.05}{\text{amount of thresholds}}$ for 95% confidence.

4.6. Translating into actionable advice

Knowing which features make a difference and for which values, we are able to make a translation into actionable advice to students. For example, if study time is found to have a significant threshold value of 10 hours per week, we can advise students to dedicate at least this amount of time to studying for improved performance.

5. Results and evaluation

Using the setup of training the models, similar to the approach of (Cortez & Silva, 2008), we find similar results for the evaluation of the models. After computing all SHAP-values for these models and taking the average over them, we get to the final normalized SHAP-values per feature. The top seven SHAP-values are shown in Table 5 and all SHAP-values can be found in Appendix B.

Table 2. Top seven SHAP-values

Feature	Average SHAP-value	Standard Deviation
G2	0.2016	0.1202
G1	0.1398	0.1229
failures	0.1046	0.1050
absences	0.0789	0.0800
age	0.0551	0.0411
freetime	0.0331	0.0291
Medu	0.0284	0.0230

Based on our threshold we find G2, G1, failures, absences, and age to be the key contributing factors. Previous grades being an important factor for future success aligns with earlier research, and also with the findings of (Cortez & Silva, 2008).

By going over all possible values and performing the t-test as described in section 4, we find that:

- All thresholds for G2 give a significant difference.
- All thresholds for G1 give a significant difference, except for the threshold being 3 or 4.
- All thresholds for failures give a significant difference.
- None of the thresholds for attendance give a significant difference.

- Thresholds for age give a significant difference when set to 17, 18 or 19.

The p-values and a plot of the different thresholds can be found in Appendix C.

Given these results, we can give the following advice to students:

- Students should make sure to get good grades in earlier periods. Working harder there to ensure they get a good grade, can have a significant impact on later study success.

- A single failure is always unfortunate, but has impact on later results as well. Preventing failures is therefore important.

Although age can be (depending on the threshold value) an important factor, a student can of course not influence this.

6. Discussion

When interpreting the various results, some things should be taken into account. First of all, we must distinguish between correlation and causality. In this project, we have been researching correlation, and make the assumption that changing one of the factors would also impact the final result. However, this assumption requires further research to establish causality and validate these findings in practice.

Another aspect which should be taken into account is that when calculating the average SHAP-values, we average over different models in different setups. Although this will make a more robust model due to the central limit theorem, it can also obscure subtle nuances between the individual setups.

Additionally, the data used in this project comes from only two Portuguese high schools, which limits the generalizability of the results. Our findings might not extend to other Portuguese high schools, other countries, and other levels of education.

It is also worth noting that some models from (Cortez & Silva, 2008) do not perform better than naive predictors. Although improving model performance was not the primary objective of this project, better-performing models could provide more accurate SHAP analyses and yield stronger insights.

Lastly, setting the threshold for normalized SHAP-values to be meaningful at 0.05 is somewhat arbitrary. Although one can argue that features with less impact are not relevant, combining multiple improvements of these features could result in a significant improvement of the final grades.

7. Conclusion

In this project, we used SHAP-analysis to identify the main contributing factors to student success. Previous grades (G1 and G2), failures, absences, and age were identified as the top five contributing factors. The results showed significant differences when filtering students based on thresholds for G1, G2, failures, and age, indicating a strong correlation between these factors and academic performance.

By translating these findings into two actionable recommendations for students—focusing more on earlier periods and preventing failures whenever possible—we aim to provide practical guidance to enhance study success. This project bridges the gap between using SHAP for interpretability and applying SHAP insights to support students in improving their academic outcomes.

However, further research is required to establish causality and confirm that influencing these factors can directly improve future study success. Additionally, future studies should explore how these findings generalize to other schools, both in different countries and across various levels of education.

References

Al Mayahi, K. and Al-Bahri, M. Machine learning based predicting student academic success. In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 264–268. IEEE, 2020.

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., and Ali, N. Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*, 2022(1): 4151487, 2022. doi: https://doi.org/10.1155/2022/4151487. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/4151487.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001. doi: 10.1023/A:1010933404324.

Cortez, P. and Silva, A. M. G. Using data mining to predict secondary school student performance, 2008.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

Ouatik, F., Erritali, M., Ouatik, F., and Jourhmane, M. Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, 17(12):pp. 236–251, Jun. 2022. doi: 10.3991/ijet.v17i12.30259.

URL https://online-journals.org/index.php/i-jet/article/view/30259.

Zeineddine, H., Braendle, U., and Farah, A. Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89: 106903, 2021.

Özkurt, C. Assessing student success: The impact of machine learning and xai-bbo approach. *Journal of Smart Systems Research*, 5(1):40–54, 2024. doi: 10.58769/joinssr.1480695.

A. Dataset features

The following table is adopted from (Cortez & Silva, 2008).

Attribute	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[6])
Mjob	Mother's job (nominal[7])
Fedu	Father's education (numeric: from 0 to 4[8])
Fjob	Father's job (nominal[9])
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 $1 \leq n < 3$, 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)

[6] 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, 4 – higher education

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

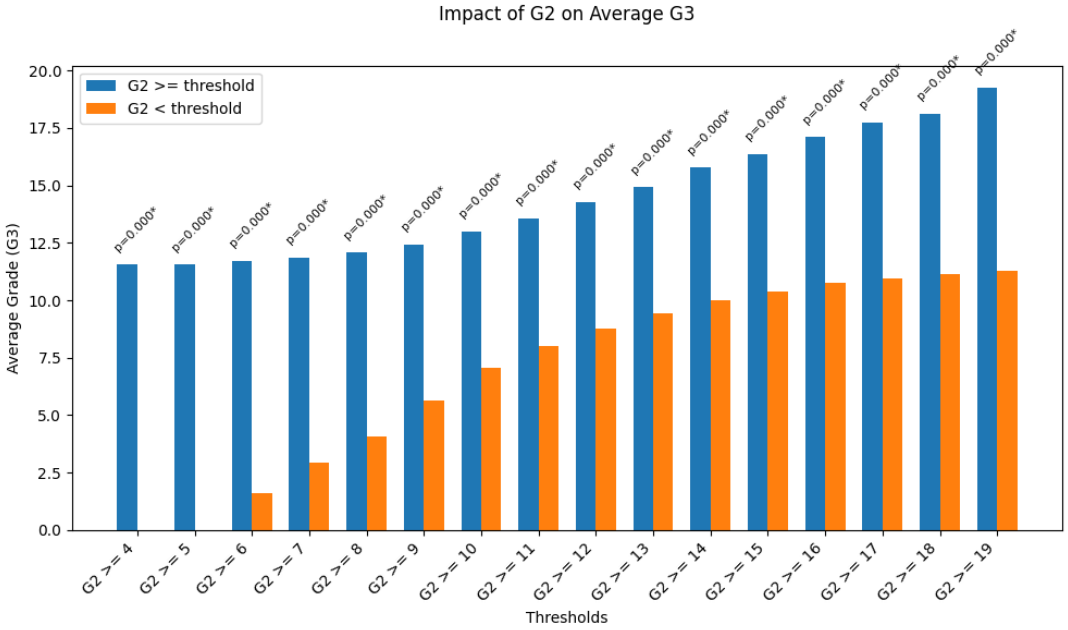
[8] Same as Medu

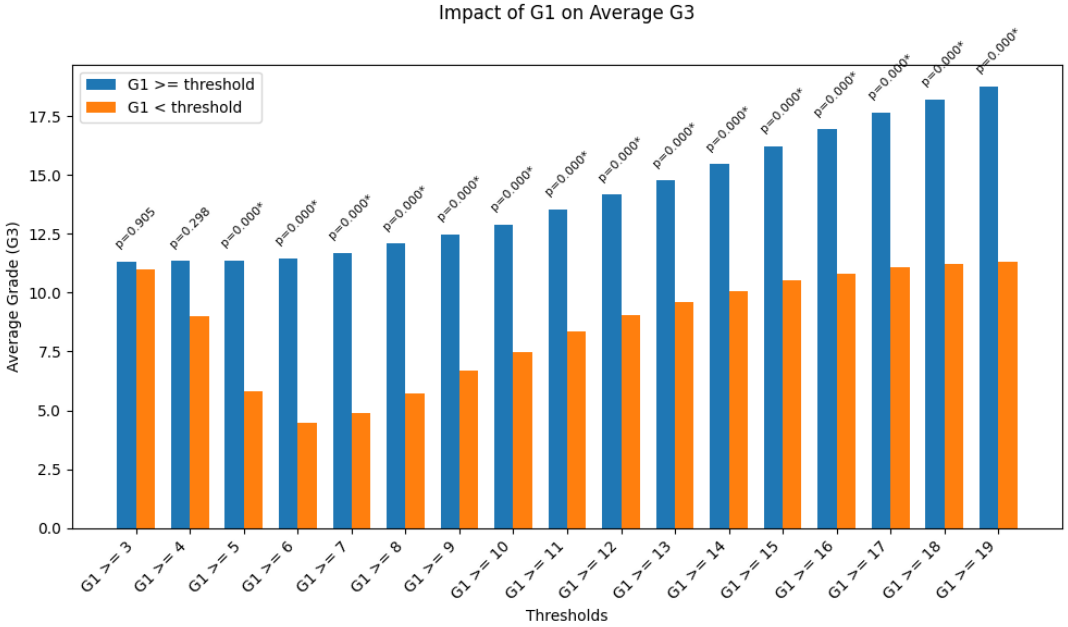
[9] Same as Mjob

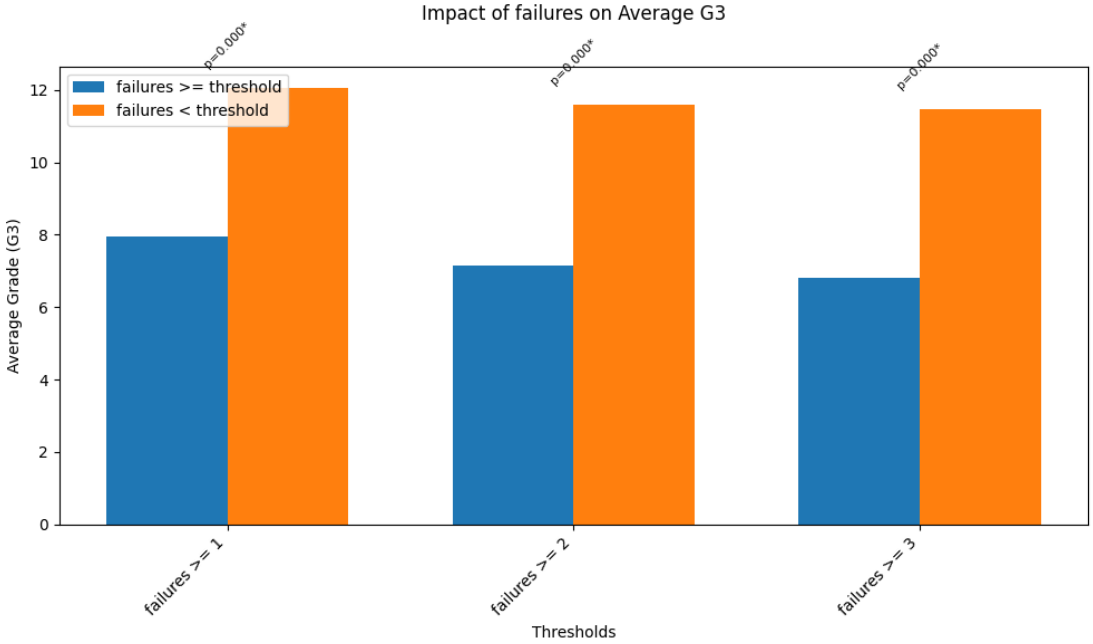
B. Average SHAP-value and standard deviation per feature

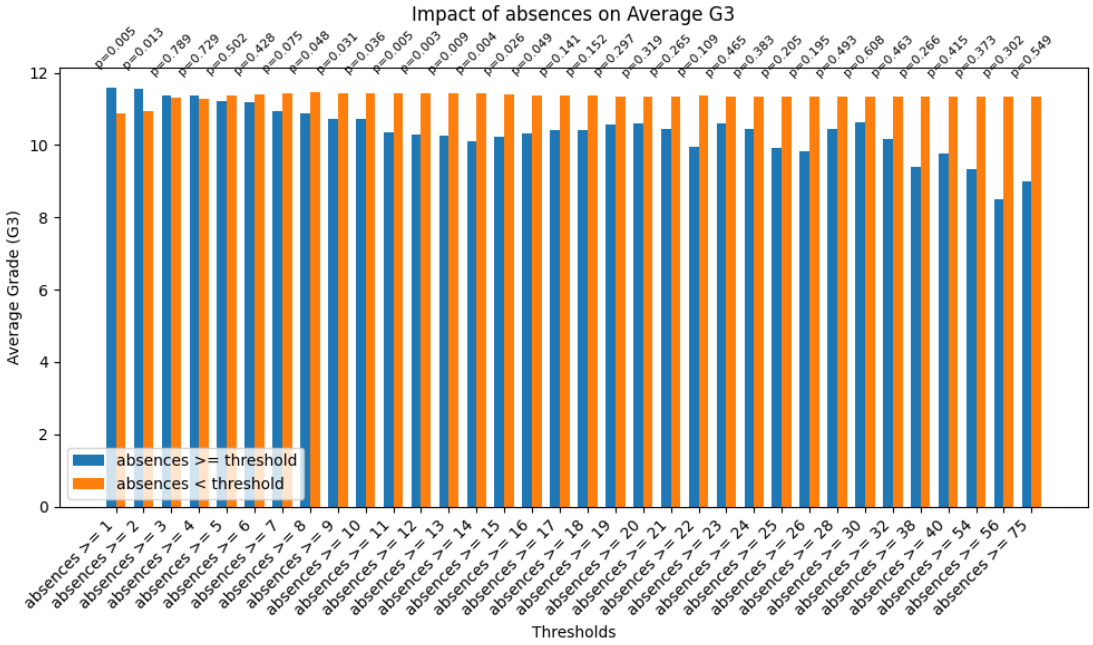
Feature	Average SHAP-value	Standard Deviation
G2	0.2016	0.1202
G1	0.1398	0.1229
failures	0.1046	0.1050
absences	0.0789	0.0800
age	0.0551	0.0411
freetime	0.0331	0.0291
Medu	0.0284	0.0230
higher_yes	0.0282	0.0322
famrel	0.0262	0.0262
Dalc	0.0259	0.0404
traveltime	0.0253	0.0235
goout	0.0251	0.0215
school_MS	0.0237	0.0255
Fedu	0.0234	0.0172
Walc	0.0234	0.0205
health	0.0206	0.0134
studytime	0.0199	0.0199
paid_yes	0.0190	0.0175
romantic_yes	0.0184	0.0261
schoolsup_yes	0.0180	0.0170
Mjob_teacher	0.0179	0.0237
Mjob_services	0.0161	0.0127
reason_reputation	0.0154	0.0107
famsize_LE3	0.0138	0.0152
reason_other	0.0132	0.0135
guardian_other	0.0117	0.0124
Fjob_teacher	0.0112	0.0125
reason_home	0.0110	0.0097
guardian_mother	0.0109	0.0109
famsup_yes	0.0109	0.0098
Mjob_other	0.0103	0.0071
activities_yes	0.0102	0.0083
address_U	0.0102	0.0096
internet_yes	0.0099	0.0093
Fjob_other	0.0097	0.0116
Fjob_health	0.0093	0.0145
Pstatus_T	0.0092	0.0080
sex_M	0.0090	0.0081
Fjob_services	0.0089	0.0062
nursery_yes	0.0081	0.0062
Mjob_health	0.0062	0.0046

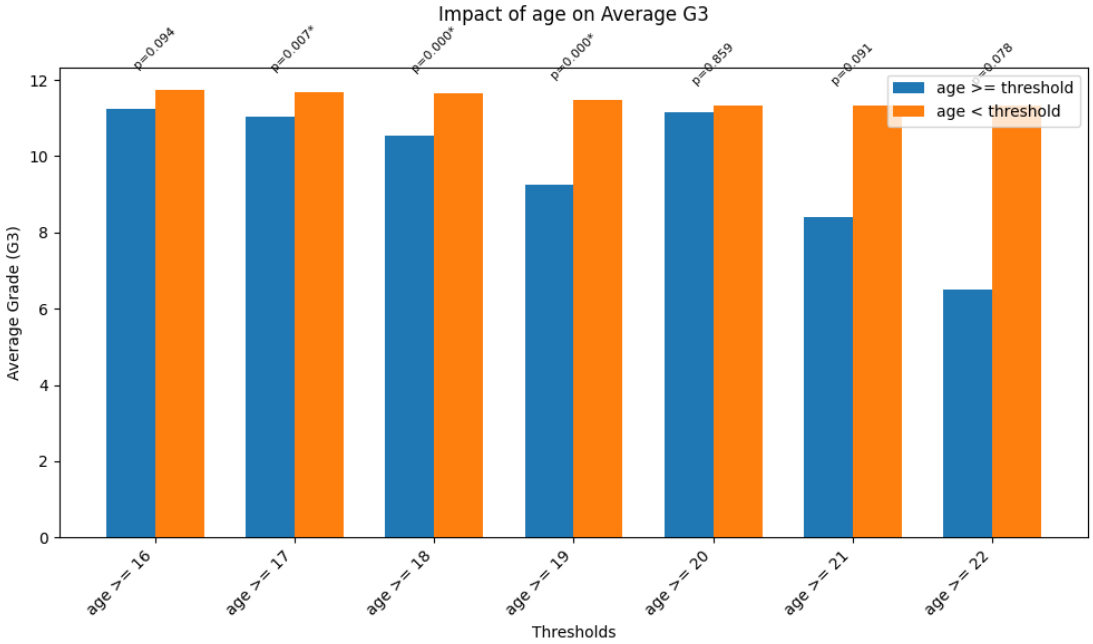
C. Plots for thresholds including p-values






Impact of failures on Average G3
p=0.000*
p=0.000*
p=0.000*
failures >= threshold
failures < threshold
Average Grade (G3)
0
2
4
6
8
10
12
failures >= 1
failures >= 2
failures >= 3
Thresholds


Impact of absences on Average G3
p=0.005
p=0.013
p=0.789
p=0.729
p=0.502
p=0.428
p=0.075
p=0.048
p=0.031
p=0.036
p=0.005
p=0.003
p=0.009
p=0.004
p=0.026
p=0.049
p=0.141
p=0.152
p=0.297
p=0.319
p=0.265
p=0.109
p=0.465
p=0.383
p=0.205
p=0.195
p=0.493
p=0.608
p=0.463
p=0.266
p=0.415
p=0.373
p=0.302
p=0.549
absences >= threshold
absences < threshold
Average Grade (G3)
0
2
4
6
8
10
12
absences >= 1
absences >= 2
absences >= 3
absences >= 4
absences >= 5
absences >= 6
absences >= 7
absences >= 8
absences >= 9
absences >= 10
absences >= 11
absences >= 12
absences >= 13
absences >= 14
absences >= 15
absences >= 16
absences >= 17
absences >= 18
absences >= 19
absences >= 20
absences >= 21
absences >= 22
absences >= 23
absences >= 24
absences >= 25
absences >= 26
absences >= 28
absences >= 30
absences >= 32
absences >= 38
absences >= 40
absences >= 54
absences >= 56
absences >= 75
Thresholds


Impact of age on Average G3
p=0.094
p=0.007*
p=0.000*
p=0.000*
p=0.859
p=0.091
p=0.078
age >= threshold
age < threshold
Average Grade (G3)
0
2
4
6
8
10
12
age >= 16
age >= 17
age >= 18
age >= 19
age >= 20
age >= 21
age >= 22
Thresholds