A Work Project, presented as part of the requirements for the Award of a Master's degree in Economics from the Nova School of Business and Economics.

TRUST4ED: DECODING THE CONDITIONS FOR TRUST IN TEACHERS AND MACHINE LEARNING MODELS FOR 4TH-GRADE PREDICTIONS

ALEXANDRA COSTA PEREIRA

Work project carried out under the supervision of:

Patrícia Xufre

Luís Catela Nunes

10/01/2024

Abstract

Effectively predicting student failure is crucial for timely resource allocation and preventing academic difficulties. This thesis explores conditions where machine learning model predictions should be trusted over teachers' judgments, recognizing potential disparities. Beginning with model building, one investigates the impact of both teachers' and students' features on the model-teacher interplay. Contrary to expectations, teachers' features do not significantly influence this dynamic, but a closer analysis of students' reveals the significance of psychological factors and academic achievement levels. The model exceeds for high-achieving students and those with more distinctive psychological patterns. This research provides relevant insights into shaping reliable educational predictions in diverse academic scenarios.

Keywords: Machine Learning in Education, Predictive Modeling, Economics of Education, Interplay Teacher-Model

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

Education is fundamental in our society, the cornerstone for developing critical thinking and character-building, and the foundation for informed citizenship. Additionally, education is necessary for the productivity of a country and for an individual's prospects of obtaining employment in general and high-quality employment in particular. Formal education typically starts in preschool and progresses through various stages, potentially culminating with higher education institutions like universities. In this process, educators hold a crucial position as they use their skills and knowledge to evaluate and anticipate any possible academic difficulties or challenges that students may encounter.

In an era of rapid technological advancement, integrating machine learning (ML) models into education has provided an appealing opportunity to augment and, in some cases, surpass traditional methods of student performance prediction. However, this search for alternative ways of predicting students' academic achievement is not so recent, as in 1978, Dee Norman Lloyd attempted to predict school failure using third-grade data (Lloyd 1978). Using regression equations and correlation analysis, the author's procedure used students' background characteristics, school performance, and other predictors to classify later school failures.

School failure often translates to the concept of grade retention, which is based on the belief that children can better acquire and solidify their knowledge, skills, and competencies by repeating a year, as opposed to advancing with their peers. Xia and Kirby (2009) provide a detailed review of the literature, based on 91 publications, analyzing the impact of retention on various outcomes, and, in fact, some studies suggest that academic progress may be observed in the years immediately following grade retention. However, the authors also mention that these gains are only expressed in the short term and fade over time. Additionally, the researchers also state that most studies reviewed in this context reveal a negative correlation between retention and the student's subsequent academic achievements.

The previous conclusions can be reasons why there has always been this need to predict school failure so that educators can intervene early and provide targeted support to students who need it, potentially preventing them from repeating a grade (Márquez-Vera et al. 2016). Another possible reason is that school failure is usually associated with high costs. These are not limited to macroeconomic factors but also have implications on private, such as lower initial and lifetime earnings, and social levels, like increased criminality (Psacharopoulos 2007).

A study conducted by Andrew J. Martin in 2011 also finds that repeating a year is associated with both academic and non-academic consequences. Regarding educational factors, the author states that grade retention negatively impacts academic self-concept and homework completion and contributes to maladaptive motivation and higher school absences. It is also mentioned that academic motivation and engagement tend to be negative consequences of school failure. When it comes to non-academic implications, it is found that repeating a year has a significant negative impact on general self-esteem.

Grade repetition is a common practice in many countries. Based on a study conducted by PISA in 2009, it was found that 13% of 15-year-old students in OECD countries reported repeating a year at least once (OECD 2011). Portugal is one of the countries significantly above this average, with 35% of students reporting that they have repeated a grade at least once in primary, secondary, or upper secondary school (OECD 2011). Requiring students to repeat grades comes with certain expenses, such as providing an extra year of education for the student and delaying their entry into the workforce for at least one year, which has a cost to society (OECD 2011). Despite the assumption that countries with high repetition rates in education would benefit from better overall performance, PISA 2009 data indicates that this is not the case (OECD 2011). In fact, the countries with the highest repetition rates tend to exhibit lower student performance. This is consistent the previously shown conclusions made by Andrew Martin in 2011.

Having this in mind, it is imperative to recognize the value of early intervention with appropriate support to prevent the possibility of grade retention. Various interventions can be employed to assist struggling students, such as changing teaching strategies to prioritize cooperative learning, providing personalized instruction, and implementing multi-grade teaching (Schiefelbein and Wolff 1992). In view of this, it is crucial to accurately identify at-risk students due to resource constraints, as different measures may also entail different costs. Besides the direct expenses related to policy implementation, there are also additional investments necessary to make strategies operational (UNESCO 1996).

This is the motivation for developing the project *GPS4Success* carried out by NOVA School of Business and Economics and CIPES (Centro de Investigação de Políticas do Ensino Superior). This study aims to understand success and failure in the 1st cycle of primary education. It seeks to identify factors that explain school failure and propose an approach using data mining techniques to prioritize at-risk students. The goal is to create a decision support system that helps teachers and school managers intervene early to promote success and prevent school failure. This thesis is developed as part of this project, taking the issue further. It consists of not only creating a machine learning model to predict school failure in elementary schools but also comparing its predictions with teachers' perception of pupil's academic performance.

It is not consensual whether machine learning models should be used to the detriment of teachers' predictions, with both having their perks and limitations. The performance of the same model can depend on various factors, such as the analyzed courses or the course difficulty, determined by the minimum score set for passing a course (Bognár and Fauszt 2022). On the other hand, the accuracy of teacher judgments of student characteristics and task difficulties may depend on different moderators, namely teacher and student characteristics or even class-level characteristics (Urhahne and Wijnia 2021). Hence, the degree to which educators may

benefit from using machine learning techniques for predicting academic failure is not unambiguous and may even rely on the circumstances of the case at hand.

Thus, the remainder of this paper is organized as follows. In **Section 2**, a literature review is undertaken, focusing on what predictors have been used to identify at-risk students and what has yet been discovered about the comparative effectiveness of machine learning and teacher subject expertise. **Section 3** covers data used in the study. The methodology applied to conduct the study and the discussion of the results are both included in **Section 4** so that this work presents a clear and logical chain of information for readers to follow. **Section 5** presents the final considerations of the research. Appendix sections are also included in this work to understand the discussion further. **Appendices A to E** are respective to **Sections 3 to 4.4**, in that order. Furthermore, **Appendix F** provides technical explanations of formal techniques used throughout this work, while **Appendix G** presents necessary statistical formulas.[1]

2. Literature Review

Prediction of students' academic performance

The topic of predicting students' academic performance has been extensively explored in the past. More than a few studies reviewed by Ameen, Alarape, and Adewole (2019) suggest that many factors are responsible for students' academic achievement and focus on its prediction, using different data mining techniques. As mentioned by Rastrollo-Guerrero et al. (2020), most student performance prediction studies are conducted on high school or higher education levels (Aguiar et al. 2015; Khasanah and Harwati 2017; Asif et al. 2017; J. Kovacic 2010). Studies on elementary school tend to analyse whether given factors affect students' academic achievements rather than predicting their performance beforehand using models (Legkauskienė, Kepalaitė, and Legkauskas 2014; Shin, Park, and Park 2009; Hopkins et al. 2019).

Prior studies recur to a variety of prediction features, with age, gender, and birth year being

[1] For a simplification of the discourse, throughout this project, the additional material in the appendix will be referenced according to its section. For example, "Table D1" refers to the Table D1 in Appendix D.

commonly used personal attributes. One example is the case of Aguiar et al. (2015), who also included education-related features, such as students' absence and tardiness rate, to develop a predictive model that identifies those at risk of not graduating high school on time. In terms of academic features, course grades, class test marks, and assignment scores are other examples of predictors used in this context (Rovira, Puertas, and Igual 2017; Aulck et al. 2016). Moreover, using demographic, socioeconomic, macroeconomic, and academic data, Aulck et al. (2016) concluded that the GPA of base courses, like Math and English, were among the strongest individual predictors of dropout and academic success. Similarly, Omolewa et al. (2019) also concluded that test scores, quizzes, and assignments were the major factors contributing to predicting academic performance in their study. In line with the work of Aulck et al. (2016), there are plenty of studies that also consider socioeconomic factors to predict academic achievement. These are mostly parent-related features, like parents' education, occupation, economic status, and family income (Davis-Kean 2005; Suleman, Hussain, and Nisa 2014). Parents' predictors are not limited to socioeconomic features, as parent involvement at home is shown to be associated with children's academic performance (Rogers et al. 2009). From a different perspective, Enu, Agyman, and Nkum (2015) concluded that teacher-related factors, such as self-motivation, method of instruction, and teaching and learning materials, affect students' performance. The authors also suggest that students' self-motivation can impact their performance, in line with Niemiec and Ryan (2009) findings that motivation is conducive to engagement and optimal learning in educational contexts, positively influencing academic achievement. Other psychological variables have been shown to be significant predictors of academic achievement, with social competence positively impacting this variable while school anxiety negatively influencing it (Legkauskienė, Kepalaitė, and Legkauskas 2014).

As for the techniques used to predict academic performance, Aulck et al. (2016) are an example

of one of the many studies that use both non-parametric and parametric models in the context of supervised learning methods. Specifically, the authors recur to K-Nearest Neighbors (KNN), Logistic Regression, and an ensemble method, namely Random Forest (RF), to predict dropout in higher education. Kabakchieva (2012) is another example that considers KNN along with other models, namely WEKA classifiers, a Decision Tree, and a Neural Network, in predicting students' academic success at a university level. All models, except for Neural Networks, had a higher true positive rate (TPR) for the "weak" class than for the "strong" class, which is how students are classified based on academic achievement. For this reason, these models could be used to identify in advance students who require additional support, as suggested by Kabakchieva (2012). Regarding unsupervised methods, clustering techniques, such as K-Means clustering and X-Means algorithm, have been widely used in the field of predicting academic performance (Sajjadi et al. 2017; Tinuke Omolewa et al. 2019; Asif et al. 2017). Clustering algorithms are particularly useful in identifying subgroups of students with similar characteristics or learning patterns, which may help educators target interventions and support specific groups of students (Sansone 2019).

Another important aspect in predicting academic performance is which metric to use to evaluate and compare models. Most papers tend to base their analysis on more than one metric, for which one good example is the work of Kabakchieva (2012). Some authors take the issue further and acknowledge the fact that the choice of the metric should consider the school's objective, which is the case of Sansone (2019). The author justifies using recall[2] as the main metric, proving that minimizing the expected dropout rate is equivalent to maximizing recall, subject to the school budget constraint (BC). It was assumed that at-risk students would take part in a program aimed at preventing dropouts, which incurs an individual cost. The resulting total cost should not exceed the school's overall resources, which is what defines the school BC.

[2] **Recall**, also known as **True Positive Rate (TPR)**, is the percentage of data samples correctly identified as belonging to a class of interest out of the total samples for that class. This class of interest is usually referred to as the "positive class". This metric may also be interpreted as 1 – Type II Error.

Machine Learning vs Teachers on Predicting Academic Performance

As suggested in the **Introduction**, in **Section 1**, teachers' trust in the reliability of machine learning (ML) predictions is of great importance as both humans and machines are prone to errors in this task. While there is limited research specifically addressing when teachers outperform machine learning models at predicting academic performance and vice-versa, some relevant studies offer valuable insights that can be explored to address this question.

Eegdeman et al. (2022) tackle this issue by exploring the predictive abilities of teachers compared to ML models when it comes to identifying students at risk of dropping out. The authors concluded that at the start of the school year, some individual teachers could predict dropout better than some algorithms. However, after the first period, when the first grade could be used in predictions, ML models were at least as good as teachers (Eegdeman et al., 2022).

Teachers' accuracy in predicting students' performance may be influenced by numerous factors. Hurwitz, Elliott, and Braden (2007) analyze the influence of teachers' and students' characteristics on teachers' judgment accuracy. Teachers' level of education or number of teaching years do not seem to influence the performance of their judgments (Hurwitz, Elliott, and Braden 2007). On the other hand, research has revealed that their perceptions of students' social behaviors are related to student gender, and these behaviors have been demonstrated to have a significant influence on teachers' evaluations of students' academic performance (Hurwitz, Elliott, and Braden 2007). Consequently, authors infer that gender has an indirect impact on teachers' assessments of students' academic abilities, which may result in a bias in these judgments. According to the authors' research, these judgments are also influenced by the student's ability level, being less accurate for lower-performing students and more accurate for higher-performing students (Hurwitz, Elliott, and Braden 2007).

Nevertheless, teachers can observe students' behavior and engagement in real-time and have conversations with them, allowing them to make informed judgments about their performance

(Eegdeman et al. 2022). Thus, the more informed they are, the better they can predict students' achievement (Südkamp, Kaiser, and Möller 2012). This is in line with the findings of Eegdeman (2022), suggesting that teachers outperform algorithms when grade results are not used for prediction, assuming that the insights resulting from personal interactions between teachers and students are not available for the algorithms.

Furthermore, as suggested by Onyema et al. (2022), machine learning algorithms are "data hungry" since they largely depend on the data they are trained on. If the data is inaccurate or incomplete, the algorithm may produce inaccurate or biased results (Onyema et al., 2022). Just like for teachers (Südkamp et al., 2012), the prediction accuracy of machine learning models improves with larger available data (Onyema et al., 2022).

3. Data

The data used for this study was collected from 45 elementary schools in Portugal. Data collection was carried out by a team of psychologists assigned to the GPS4Success project.

The data consists of 1550 students and 72 features. These features cover information about the students, teachers, and students' guardians. Features were collected during the 3rd grade of elementary school between January and June 2021. Some of the collected information comes from the results of psychological tests carried out on the students and teachers involved in this study. Teachers' judgment of student success is defined by *Id_SucessoEscolar*, which is equal to 1 if the teacher predicts the student to succeed and 0 otherwise.

The variables in the data set provided for the development of this work are all numerical since all original categorical features were transformed into dummy variables (one-hot-encoding). Of these, 16 of them are continuous, and 56 are discrete. The features description is present in **Appendix A** for further consultation.

One of the main objectives of the developed models was to predict students at risk. Thus, the target variable in this data set is defined by *Success*, which is based on the 4th-grade Portuguese

and Math test scores information, collected between March and June 2022. *Success* equals 1, if both scores were positive, and 0 otherwise.

All the analyses of the data were conducted with *Python* programming language.

4. Methodology and Discussion

4.1 Data Exploration

The purpose of this step was to analyze the dataset to gain preliminary insights, understand its characteristics, and identify potential issues or anomalies that need to be addressed.

This process involved investigating the presence of duplicate instances and potential missing values in the dataset, but no such discrepancies were found.

To proceed further, it was crucial to comprehend how the data was distributed, visualizing the distribution of each feature conditional on the target variable. This was performed using adequate reporting tools, such as histograms and bar charts, to have an initial perception of which variables were more important in explaining the difference between both types of pupils. These are the ones that present different distributions depending on the group of students. Results can be confirmed in **Figures B1** and **B2**. These showed that the features that stand out in this regard are related to students' enjoyment of the subjects for which students who pass tend to offer higher interest in the subjects than those who fail. This can be explained by the fact that this factor may impact their motivation and engagement, which has been shown by Niemiec and Ryan (2009) to influence academic achievement positively. Other variables are related to test marks, such as Portuguese and Math test scores, *NumeroErrosIdentificados,* and the number of typos in the dictation test. The respective results align with the findings of Aulck et al. (2016) and Omolewa et al. (2019) that test scores were among the strongest predictors of academic success. These insights are very intuitive, as passing students tend to present higher scores than those who fail. Psychological variables have been shown in the literature to be important to predict school performance, as suggested in **Section 2**. In this case, features related

to students' self-emotional knowledge - *PerceaooEmocionalCorreta_PEC_E* - and self-perception of their academic achievement – *AutoEficáciaEscolar* – are also associated with different distributions of the groups of students in question. Other factors such as teachers' support for students' ideas expression, students' global self-esteem, and variables related to parents' involvement reflect similar results in the charts of *Fator1_SuporteExpressaoIdeias_Criatividade*, *EscalaAutoEstimagGlobal,* and *QEP* variables, respectively. All these latter seem to assume higher values for students who pass rather than those who fail, on average. The previous results on parent involvement corroborate the conclusions of Rogers et al. (2009). Recall that this is a univariate analysis and does not consider the interaction between variables. Hence, a good rule of thumb is not to assume the remaining features are unimportant for explaining the model predictions. This must be considered when performing other analyses of the same nature in the rest of this dissertation.

Additionally, data visualization techniques were used to transform the data into a simplified 2-dimensional representation to understand the underlying structure of the dataset better. These include t-SNE (t-distributed Stochastic Neighbour Embedding) and PCA (Principal Component Analysis). For further comprehension of these techniques, please refer to **Appendix F**. Results demonstrated that the two classes exhibit a relatively distinct separation in the two-dimensional space. This led one to believe that even a linear classifier could do a reasonably good job at distinguishing the two classes, which is why one was included in the modeling section. However, one should consider that 72 features are being reduced to 2, and much information might be lost in this process. The plots may be consulted in **Figures B3** and **B4**.

4.2 Model Estimation and Interpretability

Since the main goal is to identify students at risk of failure, a new target variable, *Fail*, was defined as 1 - *Sucesso*. This way, the performance metrics would have a clearer indication of the model's ability to classify and identify at-risk students.

The next step was to decide which metric to use to evaluate the estimated model. As suggested in **Section 2**, maximizing recall (the true positive rate) subject to the school's budget constraint is equivalent to minimizing the expected dropout rate subject to the same restriction. This result is generalizable to say that by maximizing recall, we can minimize the expected positive class (*target = 1*) rate, which, in the context of our problem, translates to the expected failure rate. Besides, we should consider that, in this case, a type II error - False Negative - is worse than having a type I error - False Positive - since failing to detect a student at risk of failure could have worse implications than wrongly predicting that some student is at risk of failure. Aside from the second scenario resulting in unnecessary costs, false negative errors also lead to other long-term private and social costs, as discussed in **Section 1**. Thus, the main metric used to evaluate the models was the TPR. Nevertheless, maximizing recall at all costs can lead to costly type I errors, as mentioned above, which is why a balance between precision[3] and recall was tried to be achieved using the f1-score[4] as a secondary metric. Statistical formulas of these metrics may be consulted in **Appendix G**.

The dataset was partitioned into three subsets: the training set, dedicated to the training of the various models; the validation set, designed to compare the models' performances; and the test set, employed to evaluate the final model. For hyperparameter optimization, including choosing the best classifier, grid search cross-validation, with 5 folds, was employed. The explanation of the involved techniques, classifiers, and tuned hyperparameters is presented in **Appendix F**.

The model architecture was configured as a pipeline to include multiple sequential steps, including preprocessing ones needed to adapt the data's characteristics. The first step of the framework involved the scaling of the features to ensure that all features play on an even field.[5] Additionally, since the dataset has 72 features, removing some of these could contribute to an

[3] **Precision** is the percentage of data samples correctly identified as belonging to the positive class out of the total samples predicted as belonging to the positive class.

[4] The **F1-score** is a combination of precision and recall, providing a balance between these two metrics.

[5] In the scaling step, *MinMax Scaler* and *Standard Scaler* methods were considered.

improvement of the model's predictive power, which is why a feature selection step was also included. As for the final step, classification was performed, for which three classifiers were considered: *Logistic Regression* – a linear classifier - and the ensemble methods of *Random Forest,* and *XGBoost.* In this step, some of the classifiers' hyperparameters were also included in the grid search parameter space.

The resulting model from the grid search is the pipeline combining the following steps: *MinMaxScaler*, RFE with 30 features[6], and RF classifier. To gain a more comprehensive understanding of the model and the selected features, please refer to **Figures C1** and **C2**, respectively. Most variables identified in the univariate analysis as potentially significant to explain the target variable were part of the selected features for the best model. As an example, consider test scores and *AutoEficáciaEscolar*. The evaluation of the model on the validation set resulted in an f1-score of 92%, a recall equal to 90%, and a precision of 93%, for which the respective classification report and confusion matrix can be found in **Figures C3** and **C4**.

These scores were computed with the default threshold of 0.5 for the relevant model. Thus, threshold tunning was also performed, to attain a greater emphasis on recall than precision with still a favorable equilibrium between these metrics. The goal was to find a threshold in these conditions, resulting in an f1-score of at least 90%. The precision-recall curve is displayed in **Figure C5**, where we can visually understand the trade-off between the default threshold and the best one in this context, which was found to be 39%. The outputs for the evaluation of the new model on the validation set can be found in **Figures C6** and **C7**, where a recall = 92% and an f1-score = 90% were reported. Given that we prioritize recall, this was selected as the final model, which, when evaluated on the test set, allowed for a TPR of 96% and achieved an f1-score of 91%. The relevant outputs are once again shown in **Figures C8** and **C9**.

To conclude, feature interpretability methods were applied to the final model. These methods

[6] For the feature selection step, Recursive Feature Elimination (**RFE**) with Decision Tree as the base model was considered.

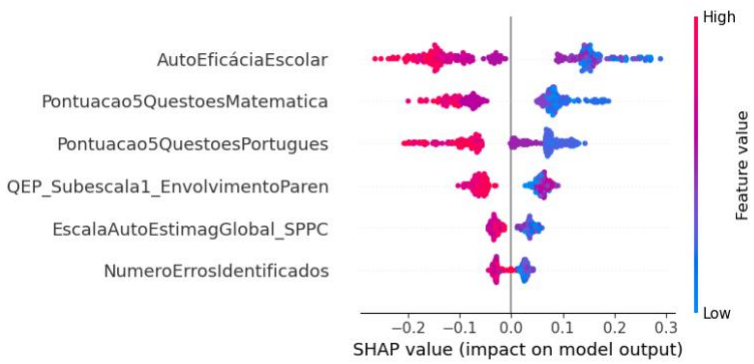

***Figure 1** - SHAP values plot of the top 6 features of the machine learning model.*

aim to explain how the model obtains its predictions, which is useful given that the model we have at hand, a Random Forest, is a "black box" model. For this purpose, both SHAP (Shapley Additive exPlanations) and PDP (Partial Dependence Plots) methods were used, illustrating how changing the features' values affects the model's prediction[7]. The results of the SHAP analysis respective to the most important features of the model may be found in **Figure 1**, which are the features the analysis will be focused on. Nevertheless, the SHAP values for all the features can be found in **Figure C10**, if needed. Looking at the plot, one can conclude that for these features, generally, lowering their values increases the probability of the student belonging to the positive class, meaning it increases the probability of failing, while increasing their values has the opposite effect[8]. To complement the analysis, PDPs for the top 6 features are also displayed in **Figures C11** to **C16**. These support the conclusions of the SHAP method, even though in a univariate setting, for all features. These model interpretability results seem to corroborate the ones found when interpreting the conditional distribution of the features.

4.3 Formal analysis of teachers' features

In this section, one aimed to formally explore how good teachers' predictions were and how they compare to the machine learning model predictions. Furthermore, one tried to understand if it was possible to distinguish which teachers benefit from using the ML model, as opposed to trusting their own perception. To ensure a fair comparison between the teachers' performance and the ML model's, these were evaluated using the test set exclusively. This approach helps

[7] **Appendix F** provides a framework for the use of these feature interpretability techniques, which were applied on the test set.
[8] SHAP values might end up being incorrectly estimated when features are correlated or interact with one another. For the sake of simplification and given that the SHAP results are consistent with the PDPs and the analysis of the distribution of the features, we will assume that this does not compromise the validity of the SHAP results. This is valid across the whole work.

prevent data leakage[9]. Concerning this matter, it is worth noting that 82 teachers were subject to this analysis, corresponding to the ones who teach students who take part in the test set.

In this section, a new variable was computed, *unsuccess_percep* = 1 – *Id_SucessoEscolar*. This reflects the teachers' perception of students' potential failure so that this analysis is consistent with the one in the previous section, where the target variable was defined as *Fail*.

As a first step, one tried to assess the performance of these teachers when predicting students' academic achievement. The table that summarizes these performance results is displayed below, as well as the respective metrics (for the positive class) in **Figure 2**. For comparison, a similar summary of the model's performance is also displayed in this Figure.

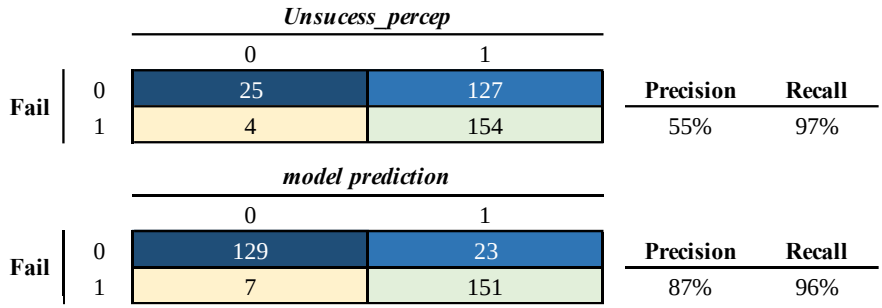

***Figure 2** – Confusion matrix and metrics of teachers' performance (**upper panel**) and model's performance (**lower panel**).*

At first glance, one could infer from these results that the teachers outperformed the model, as the TPR is our primary metric. However, taking a closer look, we understand that the model is almost as good at correctly predicting the most students out of the positive class as the teacher, with significantly better precision. From the 152 students who passed, teachers only predicted 25 of them to succeed, which roughly represents only 16% of this class[10]. Teachers overestimate the number of students who will fail, resulting in a higher accuracy in predicting academic achievement for lower-performing students than high-performing ones. This seems to counteract the findings of Hurwitz, Elliott, and Braden (2007) about the influence of students' ability level on teachers' accuracy. This level could potentially be a determining factor of the

[9] **Data leakage** occurs when information from the training set unintentionally affects the model's performance on the test set. In short, data leakage can lead to overestimating the model's accuracy and undermine the reliability of its predictions.

[10] These 16% may be interpreted as the teachers' true negative rate (the "recall" with respect to the negative class).

comparative performance of teachers and machine learning models. Given these results, one can confidently affirm that the ML model proved to be better than the teacher at predicting students' academic performance, according to the defined criteria.

The next procedure was identifying which teachers would benefit from using the model developed in the previous section rather than just evaluating teachers' overall performance. To do this, we need to assess the recall and f1 scores for each teacher's predictions applied to their group of students in the test set. To determine whether the teacher had any advantages from using the model, the metrics were compared with the ones obtained using the model's predictions for the respective group of students. Specifically, if the recall of the teacher was lower than the one obtained with the model or, in the case of the same TPR, if the teacher's f1-score was lower than the one achieved using the model, it was considered that the teacher would benefit from the machine learning tool. A new variable called *benefit_or_not* was introduced, with a value of 1 assigned to teachers who belong to this latter group and 0 for others.

A total of 39 teachers were identified as benefiting from using the ML tool. **Table D1** illustrates these results for each of them for exploration of more details. It is important to notice how working with the test set limits the substance of these conclusions, given that the metrics were computed for a very low number of students, reaching a maximum of 11 pupils per teacher. These results could hold greater significance with a larger data set, but for now, caution should be exercised when considering their generalization.

Having the teachers distinguished between these two groups, the next step was to understand if there were any discernible patterns among the teachers who benefit from the model, considering teachers' features only. To do so, a clustering analysis was conducted with K-Means, an unsupervised method that clusters similar points together. Two methods were employed. Firstly, K-Means was performed using two clusters to determine if these could potentially be differentiated by the same characteristics that also distinguish the two groups of teachers. With

the same goal, this process was applied again using the optimal number of clusters for this data set, checking if any of them contained only one of the two types of teachers. This number was determined based on three criteria: the Sum of Squared Euclidean distances of each point to its closest centroid (SSE), the Silhouette Coefficient, and the Davies-Bouldin Index. Please refer to **Appendix F** for details on how these methods and the K-Means can be used for this purpose. The plots for the relevant metrics on the choice of the number of clusters are present in **Figure D1** and **Table D2**, which present the respective values for a different number of clusters, ranging from 2 to 15. Based on these results, K-Means was performed with 9 clusters.

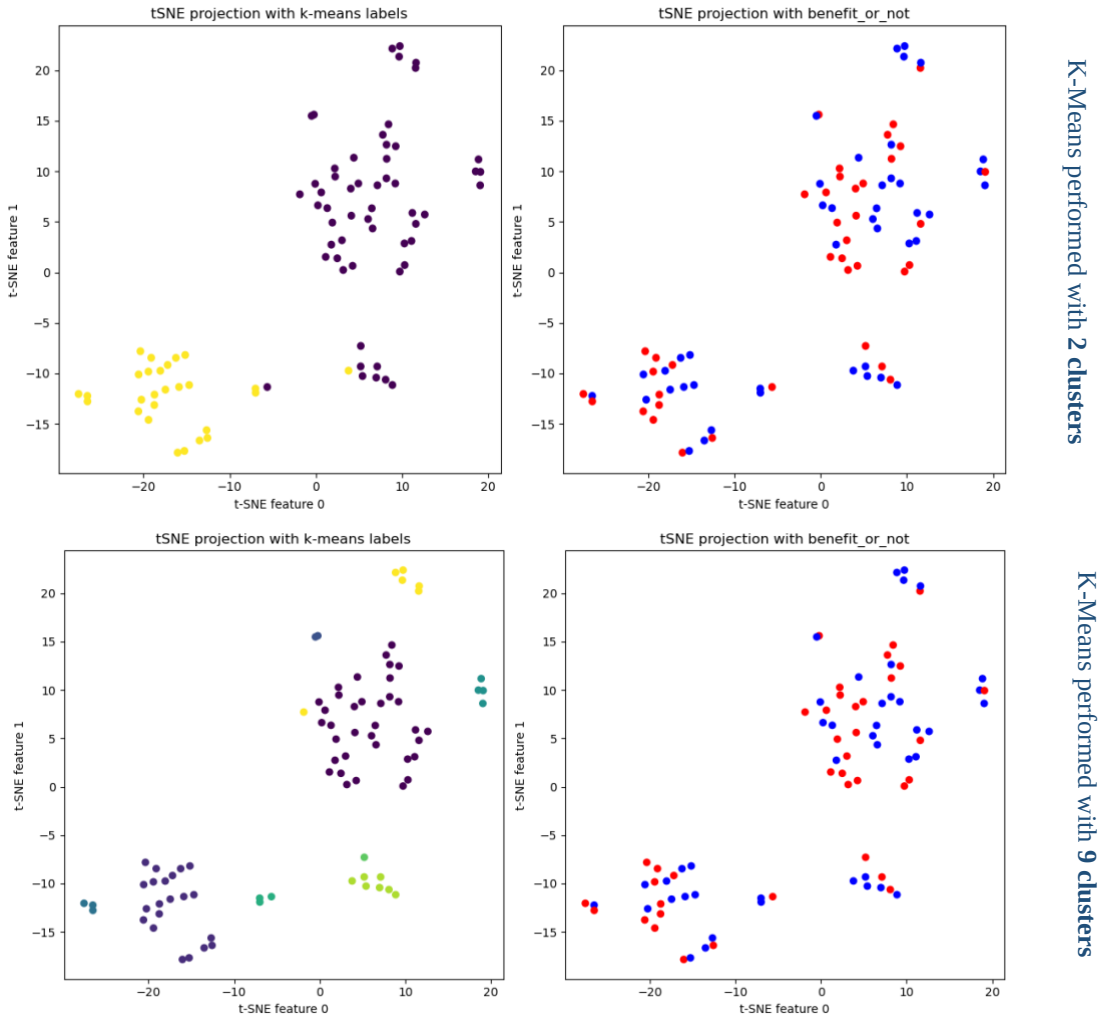

*Figure 3 - t-SNE projection of the teachers' data set considering labels based on K-Means analysis (**left panel**) and the variable benefit_or_not (**right panel**). On the right panel, a point is represented in red if benefit_or_not = 1 and blue otherwise.*

The results of the clustering analysis are displayed in **Figure 3**. We can conclude that for both methods, the two types of teachers coexist in all clusters when comparing the panels with K-

Means labels and *benefit_or_not* labels side to side.

The difficulty in finding a relation between the clusters and the binary variable shows that defining a clear pattern of features that reliably distinguishes the two groups of teachers may be challenging. The complexity of this task might be influenced by the fact that we're working with a small dataset of 82 teachers only. Additionally, it can also be related to the need for a more nuanced feature set that captures the diverse characteristics of the two groups of teachers. Finally, it was determined if any features could distinguish between the two types of educators in a univariate setting. To do so, the Chi-Square test for independence[11] was performed for each feature on the relevant binary variable, *benefit_or_not*. It is important to note that this test assumes certain assumptions, such as the variables being discrete and finite and the observations being independent. Our dataset consists of teachers' attributes that are all discrete, which confirms the validity of the first assumption. However, for the sake of simplification, we assumed that all educators in the dataset are independent, which is the second assumption. Nevertheless, it is essential to keep this simplifying assumption in mind while interpreting the results later. The Chi-Square test for independence results support the previous conclusions since no feature was found to be statistically significant for defining *benefit_or_not* at the usual significance levels. **Table D3** summarizes the results of this step.

To complement the analysis, bar charts were plotted for all teacher's characteristics for both groups. This allowed for a visual understanding of any differences in the distribution of these features based on the type of teacher, which is what the Chi-square of independence test aims to determine formally. Also in this step, no feature stands out as being potentially significant in determining *benefit_or_not*. For more details, please refer to **Figure D2**.

This motivated the next section's development. Maybe when teachers benefit from using the model is not directly explained by their observable attributes but rather their students' profiles.

[11] The general description of this test may be found in **Appendix F**, for further understanding of its application.

4.4 Formal analysis of students' features

This section aims to investigate if students' features may influence the performance of both the model and the teachers in a distinct manner, such that for each student, we can determine whether the model outperformed the teacher in predicting their performance. To conduct this section, only students' features were considered. The original dataset was used, but teachers' features, as well as their prediction for the student's success, were excluded from the analysis.

Like in the previous section, the first step was establishing a criterion that defines when the model is better than the teacher predicting the student's performance. These were considered the cases for which the model correctly predicted the student's performance, and the teacher did not. For the cases where any or both conditions were not met, we cannot consider that the teacher would benefit from relying on the ML predictor. For each student, a new variable - *model_best* - was introduced, which is 1 when both conditions are valid and 0 otherwise.

This section also included a clustering analysis of the students' features, using K-Means, with two and the optimal number of clusters. This number, obtained using the same criteria as before, was found to be 7. Outputs for this step can be checked in **Figure E1** and **Table E1**. The results reflect very distinct patterns compared to the ones described in **Section 4.3**. Focusing on **Figure 4** with 2 clusters, we can see that red points (*model_best* = 1) mostly construct the yellow cluster, whereas the second cluster mostly contains blue points (*model_best* = 0). If we consider the analysis results with 7 clusters, this pattern might not be so evident, but we can still visualize how most clusters are "dominated" by one of the two types of students.

Furthermore, a univariate analysis of the influence of students' attributes on the comparative performance of their teacher and the model was conducted. As in the previous section, this was achieved by plotting the relevant charts distinguishing between the groups and conducting formal tests to determine if there was any statistically significant difference in the distribution of the features of the two groups. In this case, we have both discrete and continuous features,

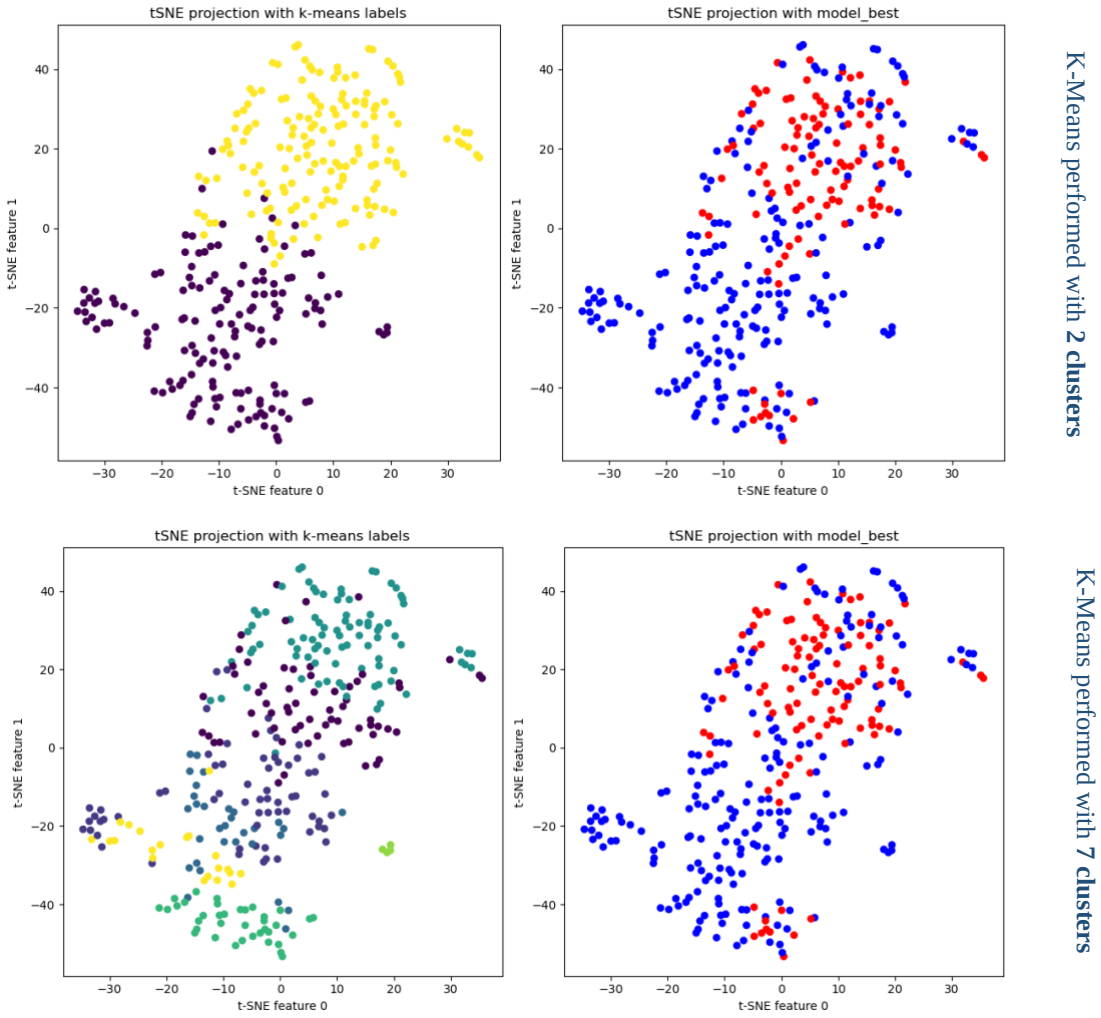

***Figure 4** - t-SNE projection of the students' data set considering labels based on K-Means analysis (**left panel**) and the variable model_best (**right panel**). On the right panel, a point is represented in red if model_best = 1 and blue otherwise.*

which is why, for the latter, another test had to be used, the Kolmogorov-Smirnov (KS) test[12]. We should consider that the test assumes that observations in the same group are also independent and identically distributed (i.i.d.). Again, these hypotheses will be assumed, but one should carefully interpret the results later.

The distribution plots are presented in **Figures E2** and **E3**, and the results of the Chi-Square and Kolmogorov-Smirnov tests are displayed in **Tables E2** and **E3**, respectively. Please refer to these resources to confirm the following conclusions. Generally, the features that stood out from the visualization charts are the ones that were also pointed out when performing a similar analysis for the prediction of failure in **Section 4.2**[13]. This could indicate that the variables that distinguish failing and succeeding students are the same as those that most differentiate the

[12] Kindly refer to **Appendix F**, where this method is described for a better understanding of its use in this context.
[13] As an exemple, consider the distribution *AutoEficaciaEscolar* and *NumeroErrosIdentificados*.

students for whom the model outperforms the teacher. Again, this suggests that students' performance might be a key factor in determining whether teachers should rely on the ML tool. Additionally, the tests performed to determine the statistically significant differences in the distributions of each variable led to results that aligned with the previous conclusions. Nevertheless, for some of the features considered to be significant by these tests, the distribution charts are not so clarifying of this significance (consider the example of students' gender and age). This disparity could be attributed to the fact that relying solely on visual inspection might be insufficient to capture this importance. Besides, discrepancies may arise due to the possibility of unmet assumptions required for conducting the tests, as suggested before.

The available evidence is suggestive that students' features may be useful in defining when the ML model should be used for predicting their academic performance. Considering this, the next step in this work consisted of performing a multivariate analysis, trying to use these features to predict *model_best*. Specifically, a second ML model was built for this purpose.

The current data set was divided between the training and test sets. For the sake of consistency and simplification, the same methods were employed to determine the optimal model, given that the data is almost identical to the original one, wherein a few features were excluded. To be specific, Grid search CV was applied, this time with the addition of a sampling method, given that the dataset is now imbalanced with respect to the target variable[14].

The f1-score was used as the evaluation metric since no type of error holds greater importance. The resulting model was the pipeline with the following steps: MinMaxScaler, SMOTE, RFE with 35 features, and RF classifier. Once again, the selection of features in the third step of the pipeline is consistent with the visual identification of the most important features in the previous step. These similarities could be more evident when considering the Chi-Square and KS tests results. Still, we should remember that what is significant in a univariate setting

[14] There are 107 students which are part of the positive class (model_best = 1), as opposed to 203 instances forming the negative class (model_best = 0). In the sampling step, two methods were considered in the grid-search parameter space – *RandomOverSampler* and *SMOTE*. A brief explanation of these techniques is presented in **Appendix F**, for further consultation.

may not be when we consider the interaction between the features and vice-versa. For a more complete understanding of the chosen characteristics, as well as the final model, kindly consult **Figures E4** and **E5**. Evaluated on the test set, the final model achieved an f1-score of 83%, resulting from a recall = 81% and precision = 85%. The respective confusion matrix and classification report can be checked in **Figures E6** and **E7**, respectively, with more details.

Note that one was working with the test set of the initial model, with a total of 310 students, which in turn had to be split between a new training and test set to build the second model. This should be considered when interpreting these and the following results, given that they are based on a sample of 62 students in the new test set.

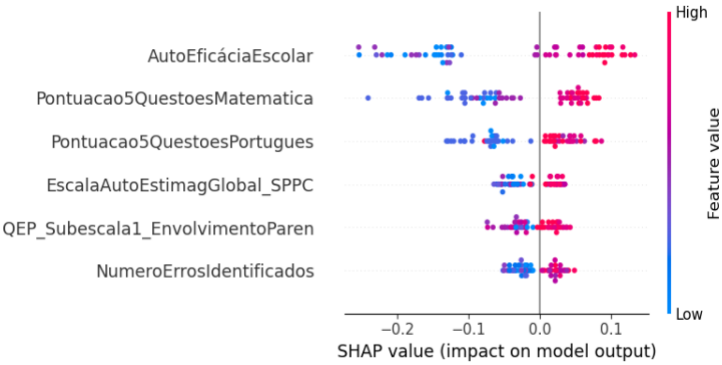

***Figure 5** - SHAP values of the top 6 features of the second model.*

Finally, to gain insights into how features affected the likelihood of a student belonging to the positive class, SHAP and PDP methods were once again employed in this section. **Figure 5** reflects the results of the SHAP analysis for the top 6 features of the second model. The full chart with the complete feature set is found in **Figure E8**. It does not go unnoticed how the most important variables to predict *Fail* in the first model are the same as the ones for predicting when teachers benefit from using the ML tool to forecast this failure. Also noted in the univariate analyses in this section, this reinforces the idea that student performance is crucial in defining the conditions for the model to outperform teachers[15]. This is also supported by the fact that three out of these top variables are related to assignment scores, strong indicators of academic performance.

Focusing on this top six, higher values of these features contribute to a higher probability of

[15] Note that one **cannot** affirm that a higher probability of success implies a higher probability of *model_best* = 1. This conclusion is not equivalent to the one stated in the text. When mentioning higher performing students, one refers to students with higher grades, not necessarily the ones most plausible to pass. As an example of why this would be wrong, consider: higher Math test score ⇒ higher probability of the student success (1st model); and higher Math test score ⇒ higher probability of *model_best* = 1 (2nd model). However, higher probability of the student success ⇏ higher probability of *model_best* = 1.

model_best = 1 and decreasing their values has the opposite effect, even though this pattern could be more evident for *QEP_SubEscala1_EnvolviementoParen*. Once again, the PDPs complement these results in a univariate context and may be consulted in **Figures E9** to **E14**.

For lower values of these features, either the model and teacher are equally accurate, or the teacher outperforms the model. This will vary depending on the specific features in question. Thus, the next step consisted of understanding which one of these scenarios applied to each feature. For matters of exemplification, only two of the top six features were included in **Figure 6**, one associated with a test score and another which is not. Nevertheless, none of the six features were shown to verify the second condition described. In this chart, the proportion of students for which the first model and the teacher accurately predicted *Fail* were plotted for each value of the features. For more details, these charts may be consulted for all six features in **Figures E15** to **E20** and complementary summary tables from **Table E4** to **E9**.

For variables directly related to students' academic performance, the teacher and model's accuracy is similar for lower values. As they increase, the model outperforms the teacher, mainly due to the lower accuracy of teachers' predictions[16]. For the variables not directly associated with students' scores, namely their perception of parents' involvement and global self-esteem, the model outperforms the teacher for the extreme values of these features, while for intermediate ones, neither prediction appears to be more accurate than the other[17].

Given these results, one could suggest that teacher accuracy tends to be lower for students with higher achievement because teachers usually base their judgments on their grades and may not always consider psychological factors affecting success. In fact, when evaluating students whose psychological features values fall within the average range, teachers' accuracy is just as good as the model's. However, if students have psychological features values that deviate

[16] For the maximum value of the Portuguese test score, their performance becomes very close again. Still, this behavior is generally present across this feature's range of values. Nevertheless, in other variables related to students' performance, such as *NumeroErrosIdentificados*, *Pontuacao5QuestoesMatematica,* and *AutoEficaciaEscolar* this conclusion is very evident.

[17] This is harder to detect in the SHAP plot. However, one should acknowledge these proportions do not necessarily reflect the model's predicted probability of *model_best* = 1. Still, it provides a hint of how this probability changes with features' values.

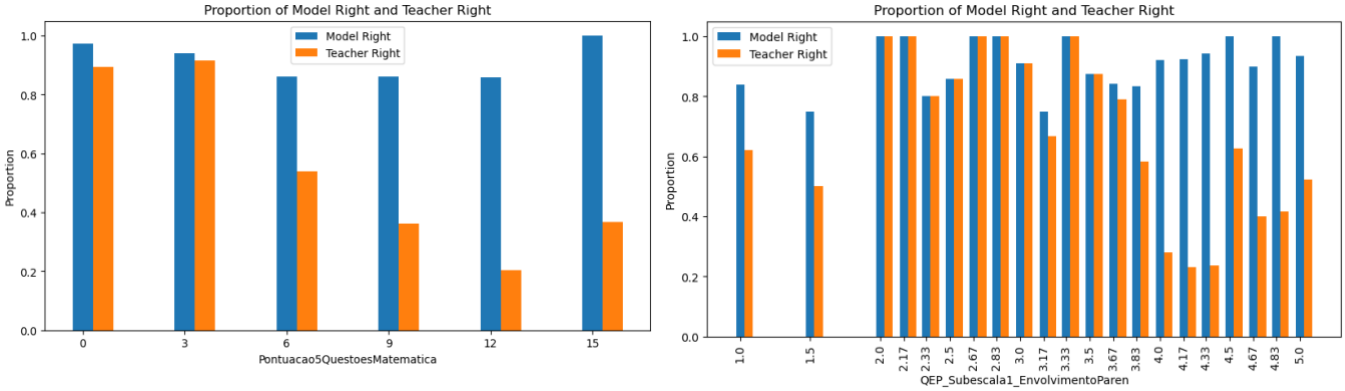

Figure 6 *– model_right and teacher_right proportions for each value of PontuacaoQuestoesMatematica (**left-panel**) and QEP_SubEscala1_EnvolviementoParen (**right-panel**). Model (Teacher) Right represents the instances for which the first model (the teacher) correctly predicted failure/success of the student.*

significantly from the norm, teachers may be unable to detect these differences, neglecting them when making judgments. On the other hand, the model considers other variables for its prediction, so it does not overestimate the number of higher-grade students who succeed.

To confirm this theory, the proportion of students whom the teacher predicted to fail is plotted for each value of the assignment scores features. For comparison, the actual proportion of students who failed and the one predicted by the model are also presented in the plot. Once again, for simplification, only one feature was included below, but **Figures E21** to **E23** and summary **Tables E10** to **E12** are available for further consultation of all three of them.

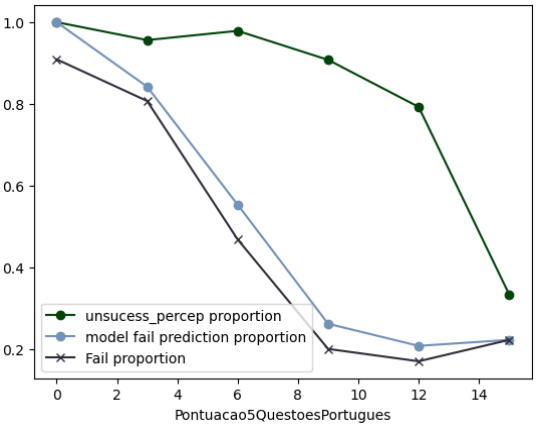

Figure 7 *– Proportion of students predicted to fail by the model and teachers (unsuccess_percep proportion) for each value of the Portuguese test score. Actual proportion is also included in the plot.*

According to **Figure 7**, the proposed explanation for the previous results is not confirmed. On the contrary, teachers tend to overestimate the number of failing students with higher grades. This is consistent with the conclusions stated in **Figure 2** in **Section 4.2**. For higher-grade students, teachers' accuracy worsens because this discrepancy between the number of students predicted to fail and the ones who fail becomes more evident. On the other hand, for lower-grade students, even if teachers overestimate the number of unsuccessful ones, this difference is not as noticeable since the failing rate for these grades is generally higher.

5. Conclusions

The integration of machine learning models in education for predicting students' success is becoming increasingly popular. However, there might be cases when teachers and the model's predictions do not coincide. In such conditions, it is crucial to identify the source to be trusted. The present work tackles this issue by using data mining techniques to identify the key factors determining when one should rely on ML algorithms rather than teachers.

A ML model to predict students' (un)success was built, and its performance was compared to the one obtained with each teacher's predictions. Afterward, an attempt was made to find any patterns that could distinguish the teachers outperformed by the model based on their attributes. One concluded the model's usefulness cannot be determined based on teachers' features alone. A second approach consisted of understanding whether this usefulness depends on students' features. According to the results, the model is shown to be useful for predicting the success of higher-performing students. Teachers tend to overestimate failing, which leads to a more pronounced underperformance for high-performing students when compared to the model's predictions. Additionally, the findings suggest how psychological factors are important in determining the model's value-added for teachers. One concluded that for students deviating most from the average levels of such features, the model tends to be more accurate.

Throughout this work, various limitations related to the size of the data set were mentioned. In future work, performing this analysis on a larger data set would ensure more significant results. Exploring the reasons for teachers to overestimate the number of non-succeeding students is crucial to uncovering student features that are either over or underrated in teachers' judgments. Some of the most important predictors of academic success for the model are of a psychological nature, which teachers might overlook. Providing teachers with this kind of information can guide them to identify at-risk students with greater accuracy. This can lead to a more effective allocation of resources, as teachers can focus on the students requiring the most attention.

References

Aguiar, Everaldo, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, and Kecia L. Addison. 2015. "Who, When and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time." *ACM International Conference Proceeding Series* 16-20-Marc: 93–102. https://doi.org/10.1145/2723576.2723619.

Agyman, Osei K, and Daniel Nkum. 2015. "Factors Influencing Students ' Mathematics Performance in Some Selected Colleges of Education in Ghana" 3 (3): 68–74.

Ameen, Ahmed O., Moshood Alabi Alarape, and Kayode S. Adewole. 2019. "Students' Academic Performance and Dropout Prediction." *Malaysian Journal of Computing* 4 (2): 278. https://doi.org/10.24191/mjoc.v4i2.6701.

Asif, Raheela, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider. 2017. "Analyzing Undergraduate Students' Performance Using Educational Data Mining." *Computers and Education* 113: 177–94. https://doi.org/10.1016/j.compedu.2017.05.007.

Aulck, Lovenoor, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2016. "Predicting Student Dropout in Higher Education." http://arxiv.org/abs/1606.06364.

Bognár, László, and Tibor Fauszt. 2022. "Factors and Conditions That Affect the Goodness of Machine Learning Models for Predicting the Success of Learning." *Computers and Education: Artificial Intelligence* 3 (September). https://doi.org/10.1016/j.caeai.2022.100100.

Davis-Kean, Pamela E. 2005. "The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment." *Journal of Family Psychology* 19 (2): 294–304. https://doi.org/10.1037/0893-3200.19.2.294.

Eegdeman, Irene, Ilja Cornelisz, Chris van Klaveren, and Martijn Meeter. 2022. "Computer or Teacher: Who Predicts Dropout Best?" *Frontiers in Education* 7 (November): 1–10.

https://doi.org/10.3389/feduc.2022.976922.

Hopkins, Shelley, Alex A. Black, Sonia L.J. White, and Joanne M. Wood. 2019. “Visual Information Processing Skills Are Associated with Academic Performance in Grade 2 School Children.” *Acta Ophthalmologica* 97 (8): e1141–48. https://doi.org/10.1111/aos.14172.

Hurwitz, Jason T., Stephen N. Elliott, and Jeffery P. Braden. 2007. “The Influence of Test Familiarity and Student Disability Status upon Teachers’ Judgments of Students’ Test Performance.” *School Psychology Quarterly* 22 (2): 115–44. https://doi.org/10.1037/1045-3830.22.2.115.

J. Kovacic, Zlatko. 2010. “Early Prediction of Student Success: Mining Students Enrolment Data.” *Proceedings of the 2010 InSITE Conference*, 647–65. https://doi.org/10.28945/1281.

Kabakchieva, Dorina. 2012. “Student Performance Prediction by Using Data Mining Classification Algorithms.” *International Journal of Computer Science and Management Research* 1 (4): 686–90.

Khasanah, Annisa Uswatun, and Harwati. 2017. “A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques.” *IOP Conference Series: Materials Science and Engineering* 215 (1). https://doi.org/10.1088/1757-899X/215/1/012036.

Legkauskienė, Šarūnė Magelinskaitė-, Albina Kepalaitė, and Visvaldas Legkauskas. 2014. “Importance of Learning Motivation, School Anxiety, and Social Competence, for Academic Achievement of the First Grade Pupils.” *Education in a Changing Society* 1 (0). https://doi.org/10.15181/atee.v1i0.677.

Lloyd, Dee Norman. 1978. “Prediction of School Failure from Third-Grade Data.” *Educational and Psychological Measurement* 38: 1193–1200.

Márquez-Vera, Carlos, Alberto Cano, Cristobal Romero, Amin Yousef Mohammad Noaman, Habib Mousa Fardoun, and Sebastian Ventura. 2016. "Early Dropout Prediction Using Data Mining: A Case Study with High School Students." *Expert Systems* 33 (1): 107–24. https://doi.org/10.1111/exsy.12135.

Martin, Andrew J. 2011. "Holding Back and Holding behind: Grade Retention and Students' Non-Academic and Academic Outcomes." *British Educational Research Journal* 37 (5): 739–63. https://doi.org/10.1080/01411926.2010.490874.

Niemiec, Christopher P., and Richard M. Ryan. 2009. "Autonomy, Competence, and Relatedness in the Classroom:Applying Self-Determination Theory to Educational Practice." *Theory and Research in Education* 7 (2): 133–44. https://doi.org/10.1177/1477878509104318.

OECD. 2011. "PISA in Focus."

Onyema, Edeh Michael, Khalid K. Almuzaini, Fergus Uchenna Onu, Devvret Verma, Ugboaja Samuel Gregory, Monika Puttaramaiah, and Rockson Kwasi Afriyie. 2022. "Prospects and Challenges of Using Machine Learning for Academic Forecasting." *Computational Intelligence and Neuroscience* 2022. https://doi.org/10.1155/2022/5624475.

Psacharopoulos, George. 2007. "The Costs of School Failure – A Feasibility Study."

Rastrollo-Guerrero, Juan L., Juan A. Gómez-Pulido, and Arturo Durán-Domínguez. 2020. "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review." *Applied Sciences (Switzerland)* 10 (3). https://doi.org/10.3390/app10031042.

Rogers, Maria A., Jennifer Theule, Bruce A. Ryan, Gerald R. Adams, and Leo Keating. 2009. "Parental Involvement and Children's School Achievement: Evidence for Mediating Processes." *Canadian Journal of School Psychology* 24 (1): 34–57. https://doi.org/10.1177/0829573508328445.

Rovira, Sergi, Eloi Puertas, and Laura Igual. 2017. "Data-Driven System to Predict Academic

Grades and Dropout." *PLoS ONE* 12 (2): 1–21. https://doi.org/10.1371/journal.pone.0171207.

Sajjadi, Seyed, Bruce Shapiro, Christopher Mckinlay, Allen Sarkisyan, Carol Shubin, and Efunwande Osoba. 2017. "Finding Bottlenecks: Predicting Student Attrition with Unsupervised Classifier." *2017 Intelligent Systems Conference, IntelliSys 2017* 2018-Janua (September 2017): 1166–72. https://doi.org/10.1109/IntelliSys.2017.8324279.

Sansone, Dario. 2019. "Beyond Early Warning Indicators: High School Dropout and Machine Learning." *Oxford Bulletin of Economics and Statistics* 81 (2): 456–85. https://doi.org/10.1111/obes.12277.

Schiefelbein, Ernesto, and Laurence Wolff. 1992. "Repetition and Inadequate Achievement in Latin America's Primary Schools: A Review of Magnitudes, Causes, Relationships and Strategies."

Shin, Hoy S., Sang C. Park, and Chun M. Park. 2009. "Relationship between Accommodative and Vergence Dysfunctions and Academic Achievement for Primary School Children." *Ophthalmic and Physiological Optics* 29 (6): 615–24. https://doi.org/10.1111/j.1475-1313.2009.00684.x.

Südkamp, Anna, Johanna Kaiser, and Jens Möller. 2012. "Accuracy of Teachers' Judgments of Students' Cognitive Abilities: A Meta-Analysis." *Journal of Educational Psychology* 104 (3): 743–62. https://doi.org/10.I037/a0027627.

Suleman, Qaiser, Ishtiaq Hussain, and Zaib-un- Nisa. 2014. "Effects of Parental Socioeconomic Status on the Academic Achievement of Secondary School Students in District Karak (Pakistan)." *International Journal of Human Resource Studies* 2 (4): 14. https://doi.org/10.5296/ijhrs.v2i4.2511.

Tinuke Omolewa, Oladele, Aro Taye Oladele, Adegun Adekanmi Adeyinka, and Ogundokun Roseline Oluwaseun. 2019. "Prediction of Student's Academic Performance Using k-

Means Clustering and Multiple Linear Regressions." *Journal of Engineering and Applied Sciences* 14 (22): 8254–60. https://doi.org/10.36478/jeasci.2019.8254.8260.

UNESCO. 1996. "Primary School Repetition: A Global Perspective."

Urhahne, Detlef, and Lisette Wijnia. 2021. "A Review on the Accuracy of Teacher Judgments." *Educational Research Review* 32 (October 2020): 100374. https://doi.org/10.1016/j.edurev.2020.100374.

Xia, Nailing, and Sheila Nataraj Kirby. 2009. "Retaining Students in Grade." *New York City Department of Education*. https://doi.org/10.2307/j.ctv1r4xd2h.38.

APPENDIX A

Data Dictionary

In what follows, the data dictionary is divided into fourteen sections for better comprehension, each referring to a different group of variables.

Four fields identify the features:

- **Variable Name**: The variable's name in the data set provided for this study. All names are in Portuguese since this study was conducted in Portugal.
- **Variable Type**: The variable's type, which can be categorical, numerical discrete, or numerical continuous.

  All categorical features turned out to be numerical discrete features in the provided data set, after performing One-Hot-Encoding. If this field is described as "categorical", this is indicative that this variable was originally categorical. When a student satisfies the condition described for each dummy variable, it is assigned a value of 1. However, if they do not meet the condition, the variable is assigned the value of the base group. This group is represented by the 0 value for every dummy feature belonging to the same originally categorical feature.
- **Variable Description**: A small description of the feature.
- **Base Group**: When applicable, it identifies the base group of the originally categorical feature.

All features were collected between January and June 2021 during 3rd grade, except for *Sucesso* – the target variable for the fail prediction model. This variable is related to the Portuguese and Math test scores, which were taken between March and June 2022 in the 4th grade.

In addition to the variables mentioned and described in this annex, the original data set also included the variable *IDProfessor*, which made it possible to identify the teacher of each student. However, this variable was excluded from the student success prediction model since

the data set already incorporates the characteristics of the respective teacher. Nevertheless, the variable was available for further use if needed, namely for conducting **Section 4.3** of this work.

STUDENT'S DESCRIPTION

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Genero	Categorical	Dummy variable identifying whether the student's gender is Female.	Male
Idade	Numerical Discrete	Student's age.	-
N_irmaos	Numerical Discrete	Student's number of siblings.	-
NroAtividadesInscritoForaEscola	Numerical Discrete	Number of extracurricular activities the student attends.	-

FAMILY BOSOM

The following information was obtained through a guardian's questionnaire.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
EE_Mae	Categorical	Dummy variable identifying whether the student's guardian is their mother.	Other
EE_Pai	Categorical	Dummy variable identifying whether the student's guardian is their father.	Other
Form_EE_1	Categorical	Dummy variable identifying if the guardian's education level is primary education.	No education level
Form_EE_2	Categorical	Dummy identifying if the guardian's education level is the 2nd cycle of basic education.	No education level
Form_EE_3	Categorical	Dummy variable identifying if the guardian's education level is the 3rd cycle of basic education.	No education level
Form_EE_4	Categorical	Dummy variable identifying if the guardian's education level is secondary education.	No education level
Form_EE_5	Categorical	Dummy variable identifying if the guardian's education level is higher education.	No education level

Vive_FN	Categorical	Dummy variable identifying whether the student lives with their nuclear family (parents and/or siblings).	Other than family
Vive_FA	Categorical	Dummy variable identifying whether the student lives with their extended family (It may include grandparents, aunts, uncles, cousins, and even more distant relatives).	Other than family
Vive_FNA	Categorical	Dummy variable identifying whether the student lives with their nuclear family and extended family.	Other than family
FamiliasReconstruidas	Categorical	Dummy variable identifying whether the student is part of a family with a history of a previous marriage or de facto union, where the respondent has since moved on with their life being in another marriage or de facto union.	No / non-applicable / no answer
FamiliasMonoparental	Categorical	Dummy variable identifying whether the student is part of a single parent family.	No / non-applicable / no answer
EE_participou	Categorical	Dummy variable identifying if the student's guardian has participated in the guardian's questionnaire.	No / non-applicable / no answer

SOCIOECONOMIC SITUATION

Information obtained through a teachers' questionnaire.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
MinoriasEtnicas	Categorical	Dummy variable identifying whether the student is part of an ethnic minority.	Not part of an ethnic minority
FamiliasDesestrut	Categorical	Dummy variable identifying whether the student is part of an unstable family with challenging dynamics.	Not part of part of an unstable family
PaisNaoAtivos	Categorical	Dummy variable identifying whether the student's parents are unemployed.	Parents are employed

TEACHER'S PERCEPTION OF THE STUDENT ACADEMIC SUCCESS

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Id_SucessoEscolar	Categorical	Dummy variable identifying whether the teacher recognized the student for their academic success.	Not recognized by the teacher for their academic success
Id_Dificuldades	Categorical	Dummy variable identifying whether the teacher stood out the student for having academic difficulties.	Not recognized by the teacher for their academic difficulties
Id_ AbsentismoEscol	Categorical	Dummy variable identifying whether the teacher stood out the student for their school absenteeism.	Not recognized by the teacher for their academic absenteeism

TEACHER'S DESCRIPTION

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
GeneroProf	Categorical	Dummy variable identifying whether the student's teacher's gender is female.	Male
HL_Lic	Categorical	Dummy variable identifying if the student's teacher's education level is an Undergraduate Degree.	Bachelor's Degree
HL_Mst	Categorical	Dummy variable identifying if the student's teacher's education level is a Master's Degree.	Bachelor's Degree
SituaçãoProfissional	Categorical	Dummy variable identifying whether the student's teacher is a permanent staff member.	On contract
IdadeProf	Numerical Discrete	Student's teacher's age.	-
NumeroAnosServico	Numerical Discrete	Total years of teaching experience of the student's teacher.	-
NumeroAnosServicoEscolaAtual	Numerical Discrete	Total years of teaching experience of the student's teacher at the current school.	-

TamanhoTurmaAtual	Numerical Discrete	Number of students in the teacher's current class.	-
Indiferente (**EGSA – Classroom Management Style - Questionnaire)**	Categorical	Dummy variable identifying whether the students' teacher classroom management style is indifferent.	No predominant style
Permissivo (**EGSA – Classroom Management Style - Questionnaire)**	Categorical	Dummy variable identifying whether the students' teacher classroom management style is permissive (allows students more freedom and autonomy in their behavior and decision-making.	No predominant style
Persuasivo (**EGSA – Classroom Management Style - Questionnaire)**	Categorical	Dummy variable identifying whether the students' teacher classroom management style is persuasive.	No predominant style
Autoritário (**EGSA – Classroom Management Style - Questionnaire)**	Categorical	Dummy variable identifying whether the students' teacher classroom management style is authoritarian.	No predominant style

STUDENT'S PREFERENCES

Features measuring the student's preferences for school subjects, teachers, and classmates:

1 – None; 2 – Little; 3 – Some; 4 – Enough; 5 – A lot.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
GostoLinguaPortuguesa	Numerical Discrete	Student's preference for the Portuguese subject.	-
GostoMatematica	Numerical Discrete	Student's preference for the Math's subject.	-
GostoEstudodoMeio	Numerical Discrete	Student's preference for the Environmental Studies subject.	-
GostoEducaçaoFisica	Numerical Discrete	Student's preference for the Physical Education subject.	-
GostoIngles	Numerical Discrete	Student's preference for the English subject.	-
GostoExpressoes	Numerical Discrete	Student's preference for the Art's subject.	-

GostoEscola	Numerical Discrete	Student's preference for their school.	-
GostoProfessores	Numerical Discrete	Student's preference for their teachers.	-
GostoColegas	Numerical Discrete	Student's preference for their classmates.	-

"A MINHA SALA DE AULA" QUESTIONNAIRE

A self-report questionnaire that assesses the climate of creativity in the classroom through 4 factors. For each factor, the higher the results, the higher the evaluated field.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Fator1_SuporteExpressaoIdeias_C	Numerical Continuous	Teacher support for student's expression of ideas.	-
Fator2_InteresseAlunoAprendizage	Numerical Continuous	Student's interest in learning.	-
Fator3_AutoPercecaoCriatividad	Numerical Continuous	Student's self-perception of their creativity.	-
Fator4_AutonomiaAluno_Criatividad	Numerical Continuous	Student's self-perception of their independence in tasks development.	-

"SELF PERCEPTION PROFILE FOR CHILDREN – SPPC" (HARTER, 1985) – PORTUGUESE ADAPTATION

A scale that evaluates the *self*'s characteristics or attributes that are consciously perceived/described by the individual through the language. These attributes are distributed by five domains: academic competence, social acceptance, physical appearance, athletic competence, and conduct. It also evaluates the student's perception of their global self-esteem. For each item, the higher the scale's results, the higher the student's self-perception of the respective attribute.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
EscalaCompetenciaAcademica_SPPC	Numerical Continuous	Student's self-perception of their academic competence.	-
EscalaAceitacaoSocial_SPPC	Numerical Continuous	Student's self-perception of their social acceptance.	-
EscalaAparenciaFisica_SPPC	Numerical Continuous	Student's self-perception of their physical appearance.	-
EscalaCompetenciaAtletica_SPPC	Numerical Continuous	Student's self-perception of their athletic competence.	-
EscalaConduta_SPPC	Numerical Continuous	Student's self-perception of their behavior.	-
EscalaAutoEstimagGlobal_SPPC	Numerical Continuous	Student's self-perception of their global self-esteem.	-

ASSESSMENT OF CHILDREN'S EMOTION SKILLS - ACES – PORTUGUESE ADAPTATION OF THE SCALE

It is a scale that evaluates children's emotional knowledge. For this work, only one of the respective subscales was considered – emotional behaviors. This scale aims to evaluate the ability of the children to correctly associate each stimulus (a behavior) to one of the five possible emotions: happiness, sadness, anger, angriness, and "neutral". The higher the values of this feature, the higher the student's emotional knowledge.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
PerceaooEmocionalCorreta_PEC_E	Numerical Discrete	Variable evaluating student's emotional knowledge	-

PARENT'S INVOLVEMENT QUESTIONNAIRE

This questionnaire is answered by the students and aims to assess their perception of their parents/guardians' involvement in school activities. It is evaluated on the following six different fields.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
QEP_Subescala1_EnvolvimentoParen	Numerical Continuous	Student's parents' involvement on school and volunteering activities.	-
QEP_Subescala2_EnvolvParentalAti	Numerical Continuous	Student's parent's involvement on home learning activities	-
QEP_Subescala3_ComunicaçaoEscol	Numerical Continuous	Evaluation of the school-family communication	-
QEP_Subescala4_EnvolvAtividadesE	Numerical Continuous	Student's parent's involvement on parents' meetings.	-
QEP_Subescala5_CastigoseRecompen	Numerical Continuous	Student's parent's adequation of their punishments and rewards regarding school.	-
QEP_EnvolvimentoParentalGlobal_M	Numerical Continuous	Student's parent's global involvement on their kid's school-related activities.	-

SCHOOL SELF-EFFICACY QUESTIONNAIRE

This questionnaire is formed by 20 statements, 12 related to the self-perception of academic performance ability. The remaining 8 statements relate to the perception of performance compared to peers or the evaluation of others. The following feature is the score obtained by adding up all the answers (the higher the score, the more school efficacy the student perceived).

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
AutoEficáciaEscolar	Numerical Discrete	School efficacy perceived by the student.	-

STRENGTHS AND DIFFICULTIES QUESTIONNAIRE (SDQ) – PORTUGUESE ADAPTATION

The Strengths and Difficulties Questionnaire (SDQ) is a widely used tool for assessing the psychological well-being of children and adolescents. It is designed to provide a brief and reliable measure of emotional and behavioral functioning in young people.

It comprises 25 items and covers five domains: Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Relationship Problems, and Prosocial Behavior. Each variable is respective to one of these subscales and their possible values are: 0 – Normal Values, suggesting that the student is not experiencing significant difficulties in that area.; 1 – Borderline Values, indicating that the child is not exhibiting severe difficulties but may be showing some mild or moderate signs of challenges; and 2 – Abnormal Values, suggesting that the student is experiencing notable challenges in the specified area, and intervention or further assessment may be warranted.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Sintomas_emocionais_SDQ	Numerical Discrete	A variable that assesses emotional issues such as anxiety and depression.	-
Problemas_comportamento_SDQ	Numerical Discrete	A variable that focuses on behavioral problems and conduct issues.	-
Hiperatividade_SDQ	Numerical Discrete	A variable that assesses symptoms related to attention deficit hyperactivity disorder (ADHD) and concentration difficulties.	-
Problemas_Relacionamento_SDQ	Numerical Discrete	A variable that focuses on peer relationship issues.	-
Comportamento_Prosocial_SDQ	Numerical Discrete	A variable that assesses positive behaviors such as helpfulness, sharing, and empathy.	-

STUDENT'S ACADEMIC EVALUATIONS

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Pontuacao5QuestoesPortugues	Numerical Discrete	Score of the 3rd grade knowledge test - Portuguese subject questions.	-
Pontuacao5QuestoesMatematica	Numerical Discrete	Score of the 3rd grade knowledge test – Math's subject questions.	-
NumeroErrosIdentificados	Numerical Discrete	Number of misspellings correctly identified in a misspelled text – 3rd grade assessment.	-
NumeroPalavrasBemEscritasIdentif	Numerical Discrete	Number of well-written words identified as errors by the student in a misspelled text – 3rd grade assessment.	-
NroErrosDitado	Numerical Discrete	The number of errors (misspelled words) in the dictation test – 3rd grade assessment.	-

TARGET VARIABLE OF THE FAIL/SUCCESS PREDICTION MACHINE-LEARNING MODEL

The target variable of the first machine learning model – *Sucesso* – was defined based on the scores of the Portuguese and Math tests taken at the end of the 4th grade of elementary school. If the score was positive on both tests, the student was considered to pass in 4th grade, *Successo* = 1. On the other hand, *Sucesso* = 0 if both or one of these conditions was not met.

Variable Name	Variable Type	Variable Description	Base Group (when applicable)
Sucesso	Categorical	Dummy variable which identifies whether the student was successful in the 4th grade.	Fail

APPENDIX B - E

Methodology and Discussion

This appendix presents a collection of complementary outputs mentioned throughout the work project.

These outputs have been included to help better understand the subjects discussed in the main document and provide additional insights into the analysis performed. The materials presented in this appendix include visualizations, statistical summaries, and other relevant information that may help the reader comprehend the key findings of the research.

Due to their dimension, some figures had to be split between two subfigures, namely **Figures B1** and **E2**. For this purpose, these are divided between panels *a* and *b*.

In order to gain a more comprehensive understanding of the outputs generated through the application of some techniques, reference may be made to **Appendix F**, which provides a structured framework for the use of these formal methods.

All the outputs were generated using *Python* programming language.

APPENDIX B

Data Exploration

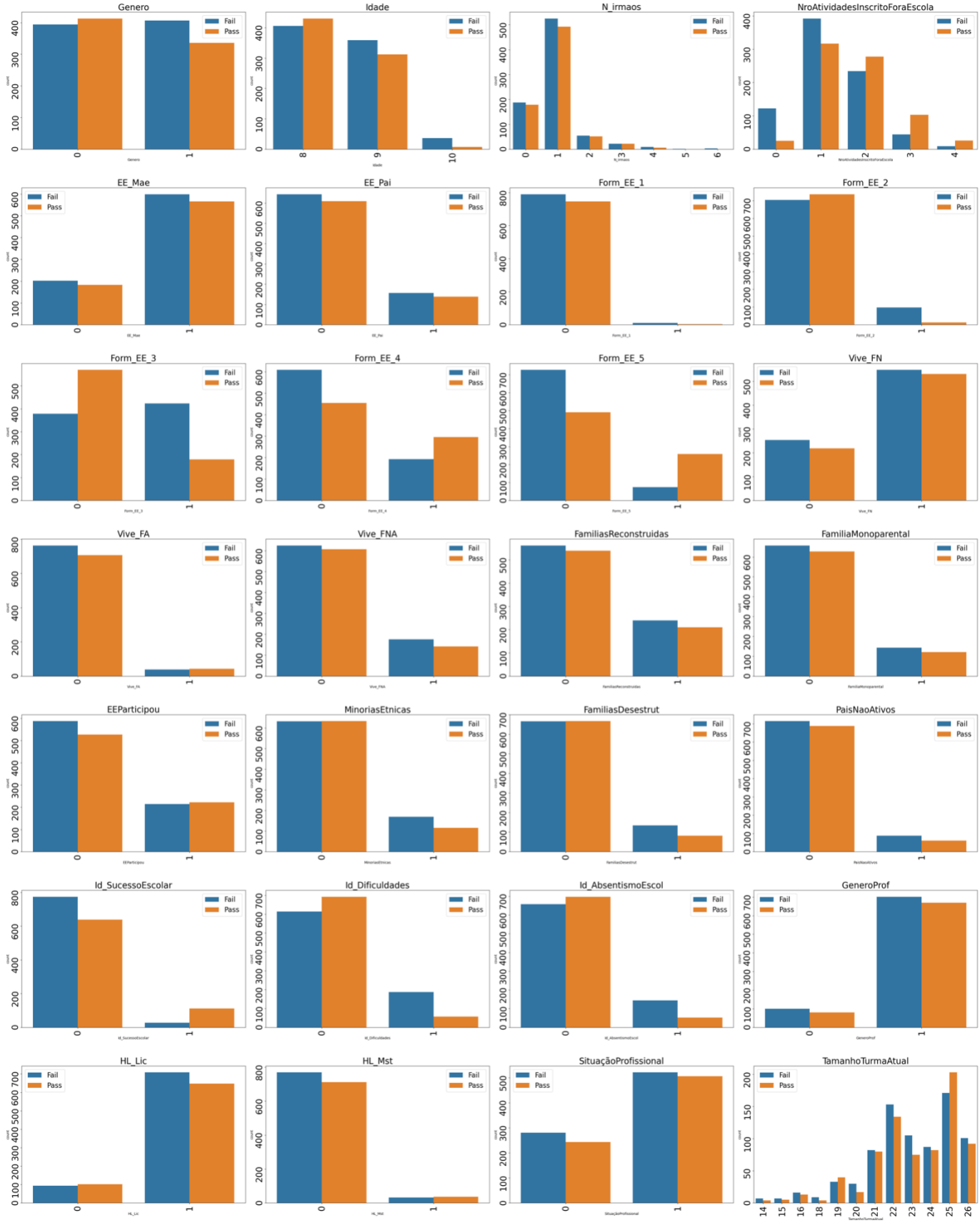

Figure B 1 a) *- Bar plots reflecting the distribution of the* ***discrete*** *features in the data set, distinguished between the students who failed (****blue bars****) and those who passed (****orange bars****). Due to its dimension, the figure had to be divided into two figures, namely* ***Figure B 1a)*** *and* ***Figure B 1b)****.*

Figure B 1 b) *- Bar plots reflecting the distribution of the* ***discrete*** *features in the data set, distinguished between the students who failed (****blue bars****) and those who passed (****orange bars****). Due to its dimension, the figure had to be divided into two figures, namely* ***Figure B 1a)*** *and* ***Figure B 1b)****.*

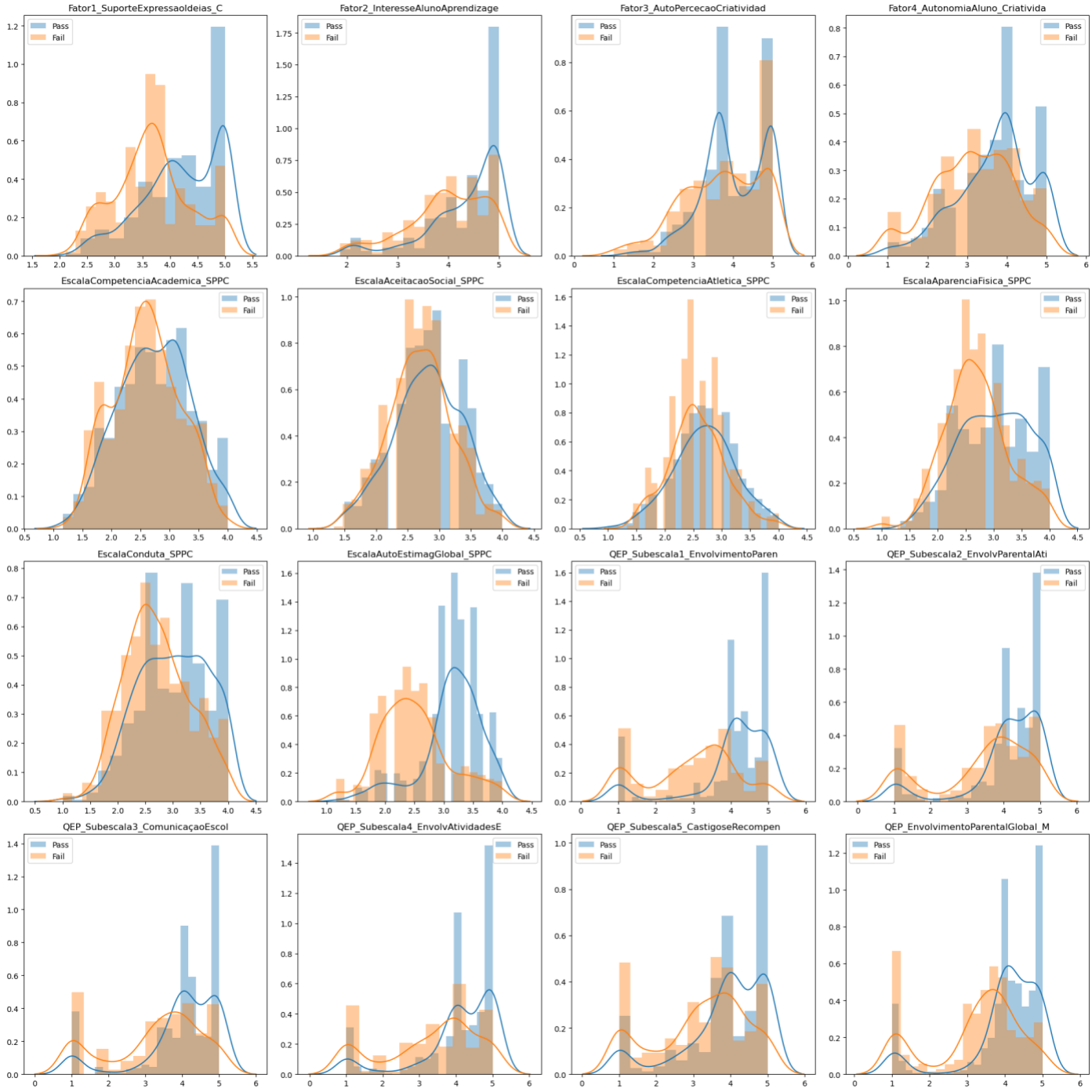

Figure B 2 *– Histograms and Kernel Density plots reflecting the distribution of the* ***continuous*** *features in the data set, distinguished between the students who passed (****blue plots****) and those who failed (****orange plots****).*

The following illustrations, **Figures B3** and **B4**, are respective to the application of visualization techniques of the data set, namely Principal Component Analysis (PCA) and t-SNE (t-distributed Stochastic Neighbour Embedding). For further comprehension of the application of these techniques, please refer to **Appendix F**.

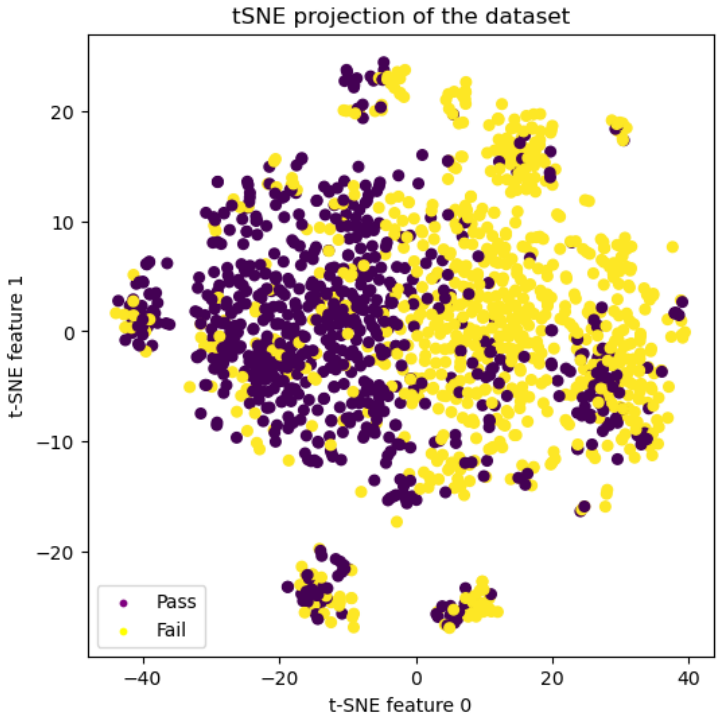

Figure B 3 *– t – SNE two-dimensional projection of the data set. Purple points are representative of instances for which Sucesso = 1 and yellow points are respective to students for which Sucesso = 0.*

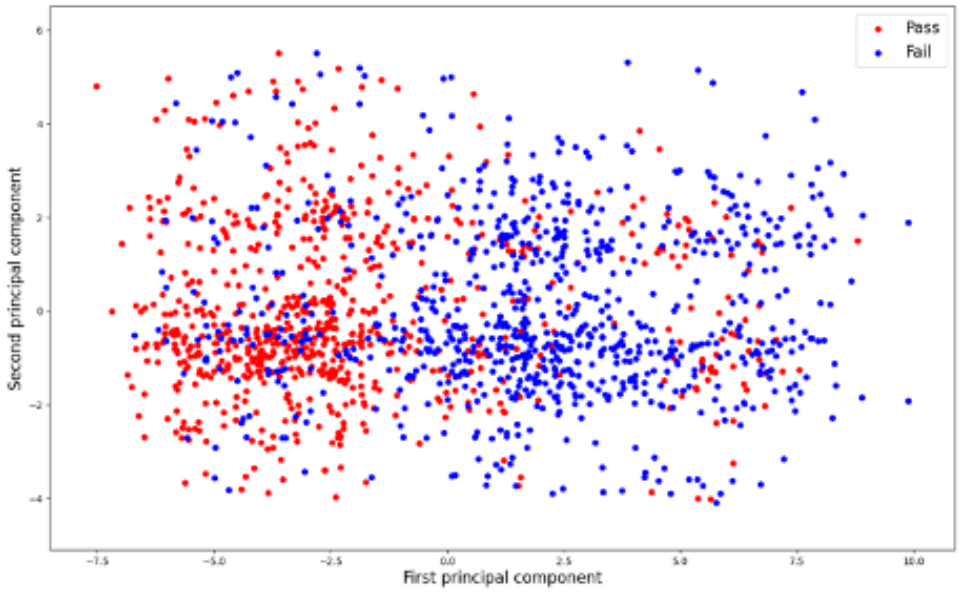

Figure B 4 *– PCA representation of the data set using two principal components. Red points are representative of instances for which Sucesso = 1 and blue points are respective to students for which Sucesso = 0.*

APPENDIX C

Model Estimation and Interpretability

```
best_estimator
```

```
Pipeline(steps=[('scaling', MinMaxScaler()),
                ('selection',
                 RFE(estimator=DecisionTreeClassifier(),
                     n_features_to_select=30)),
                ('classifier',
                 RandomForestClassifier(max_depth=7, min_samples_split=6,
                                        n_estimators=300, random_state=42))])
```

Figure C 1 *– Final model resulting from the grid-search process. The model is structured as a pipeline with a feature scaling, feature selection and classifier steps. For further comprehension of the involved steps, please refer to* ***Appendix F****.*

```
mask=best_estimator.named_steps['selection'].get_support()
chosen_feat = best_estimator['scaling'].get_feature_names_out()[mask]
chosen_feat
```

```
array(['Genero', 'EE_Pai', 'Form_EE_1', 'Form_EE_2', 'Form_EE_3',
       'Form_EE_4', 'Form_EE_5', 'HL_Lic', 'TamanhoTurmaAtual',
       'IdadeProf', 'NumeroAnosServicoEscolaAtual', 'Indiferente',
       'Permissivo', 'GostoEstudodoMeio',
       'Fator2_InteresseAlunoAprendizage',
       'Fator3_AutoPercecaoCriatividad',
       'Fator4_AutonomiaAluno_Criativida',
       'PerceaooEmocionalCorreta_PEC_E', 'EscalaAceitacaoSocial_SPPC',
       'EscalaCompetenciaAtletica_SPPC', 'EscalaAutoEstimagGlobal_SPPC',
       'QEP_Subescala1_EnvolvimentoParen',
       'QEP_Subescala4_EnvolvAtividadesE',
       'QEP_EnvolvimentoParentalGlobal_M', 'AutoEficáciaEscolar',
       'Problemas_Relacionamento_SDQ', 'Pontuacao5QuestoesPortugues',
       'Pontuacao5QuestoesMatematica', 'NumeroErrosIdentificados',
       'NroErrosDitado'], dtype=object)
```

Figure C 2 *– Selected features resulting from second step of the best model resulting from the grid-search. To be specific, the features were selected with Recursive Feature Elimination method, with Decision Tree Classifier as the base model.*

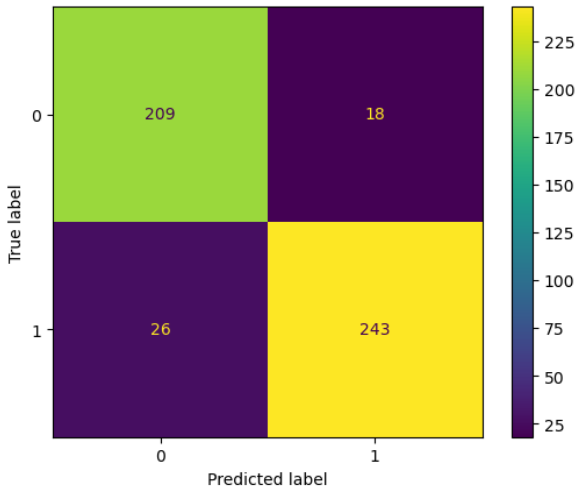

Figure C 3 *– Confusion matrix resulting from the evaluation of the grid-search best model on the validation set. These results were obtained with the default probability threshold of 0.5.*

Classification Report

	precision	recall	f1-score	support
0	0.89	0.92	0.90	227
1	0.93	0.90	0.92	269
accuracy			0.91	496
macro avg	0.91	0.91	0.91	496
weighted avg	0.91	0.91	0.91	496

Figure C 4 *– Classification Report resulting from the evaluation of the grid-search best model on the validation set. These results were obtained with the default probability threshold of 0.5.*

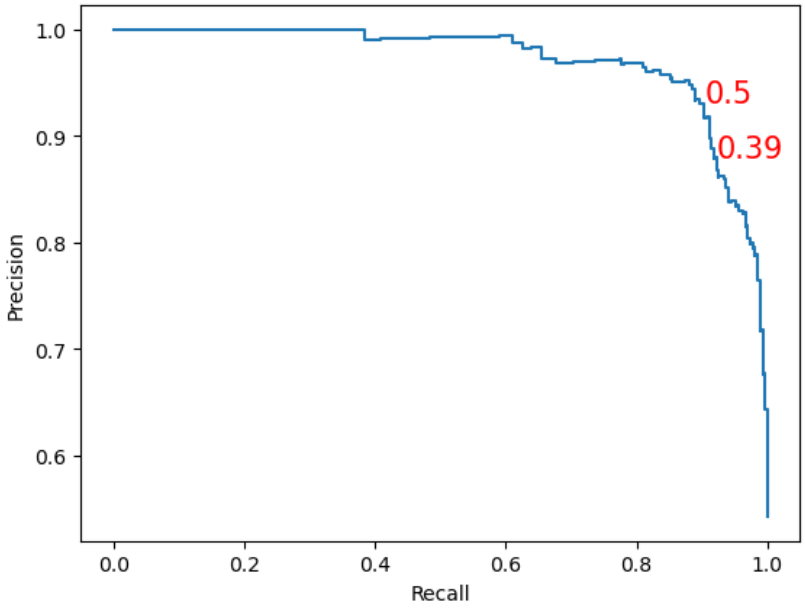

Figure C 5 *– Precision-recall curve resulting from the evaluation of the grid-search final model on the validation set. Each point in this curve is respective to a different threshold. The 39%-threshold that allows for a greater emphasis on recall rather than precision, with still a good balance between these metrics translated into an f1-score = 90%, is represented in the curve. For comparison, the default threshold is also pointed in the plot.*

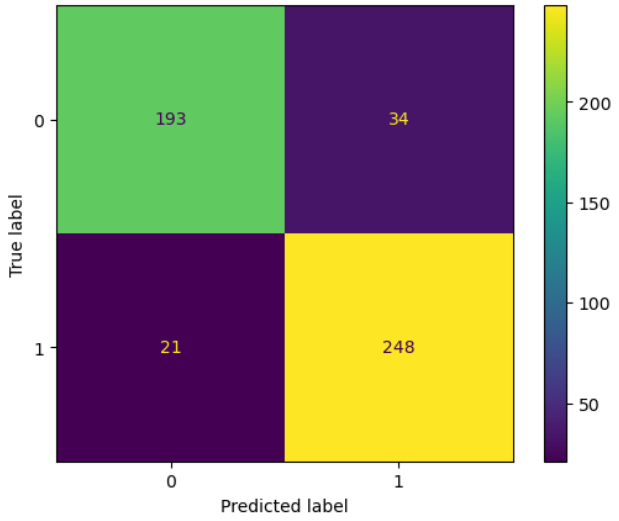

Figure C 6 *– Confusion matrix resulting from the evaluation of the grid-search best model on the validation set. These results were obtained with the probability threshold of 0.39.*

	precision	recall	f1-score	support
0	0.90	0.85	0.88	227
1	0.88	0.92	0.90	269
accuracy			0.89	496
macro avg	0.89	0.89	0.89	496
weighted avg	0.89	0.89	0.89	496

Figure C 7 *– Classification report resulting from the evaluation of the grid-search best model on the validation set. These results were obtained with the probability threshold of 0.39.*

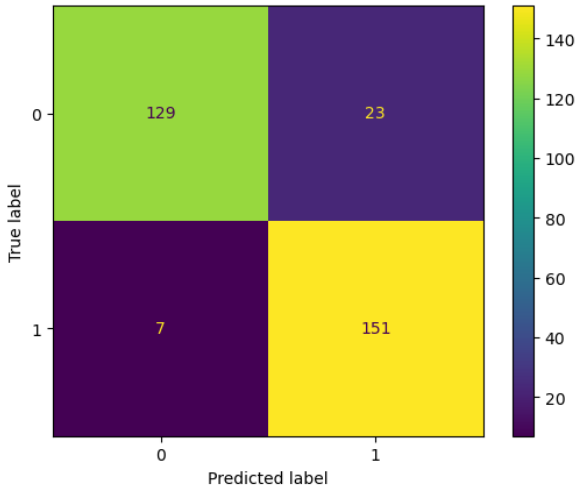

Figure C 8 *– Confusion matrix resulting from the evaluation of the grid-search best model on the test set. These results were obtained with the probability threshold of 0.39.*

Final Classification Report

	precision	recall	f1-score	support
0	0.95	0.85	0.90	152
1	0.87	0.96	0.91	158
accuracy			0.90	310
macro avg	0.91	0.90	0.90	310
weighted avg	0.91	0.90	0.90	310

Figure C 9 *– Classification report resulting from the evaluation of the grid-search best model on the test set. These results were obtained with the probability threshold of 0.39.*

The following outputs – **Figures C10** to **C16** – are respective to the application of feature interpretability methods on the model, namely SHAP (Shapley Additive exPlanations) and PDP (Partial Dependence Plots). These plots result from applying the named techniques on the **test set** to gain insights into how the model makes predictions on new, unseen data, which represents the data that the model will be applied to when it becomes available for schools to use.

- **SHAP Output Interpretation**: The position on the y-axis is determined by the feature since these are ranked in descending order and on the x-axis by the Shapley value. The colors are representative of how lowering (**blue**) or increasing (**red**) the value of the features decreases (lower SHAP value) or increases (higher SHAP value) the probability of the student being of the positive class (Fail).
- **PDP Output Interpretation:** On the x-axis, we can find the values of the feature we are trying to understand. On the y-axis, we see the average value of the prediction, meaning the probability of the student belonging to class 1 for all students for whom the feature assumes each value. The bar on the bottom represents the density of the data points for each feature value. This helps us understand the goodness of the inference. In the parts where the number of points is quite small, the model has seen fewer examples, and the interpretation can be tricky. PDPs show the average trend and empirical confidence levels, represented by the blue area.

Appendix F may be consulted for a better understanding of the use of these techniques, including when analyzing similar outputs in **Figures E8-E14**.

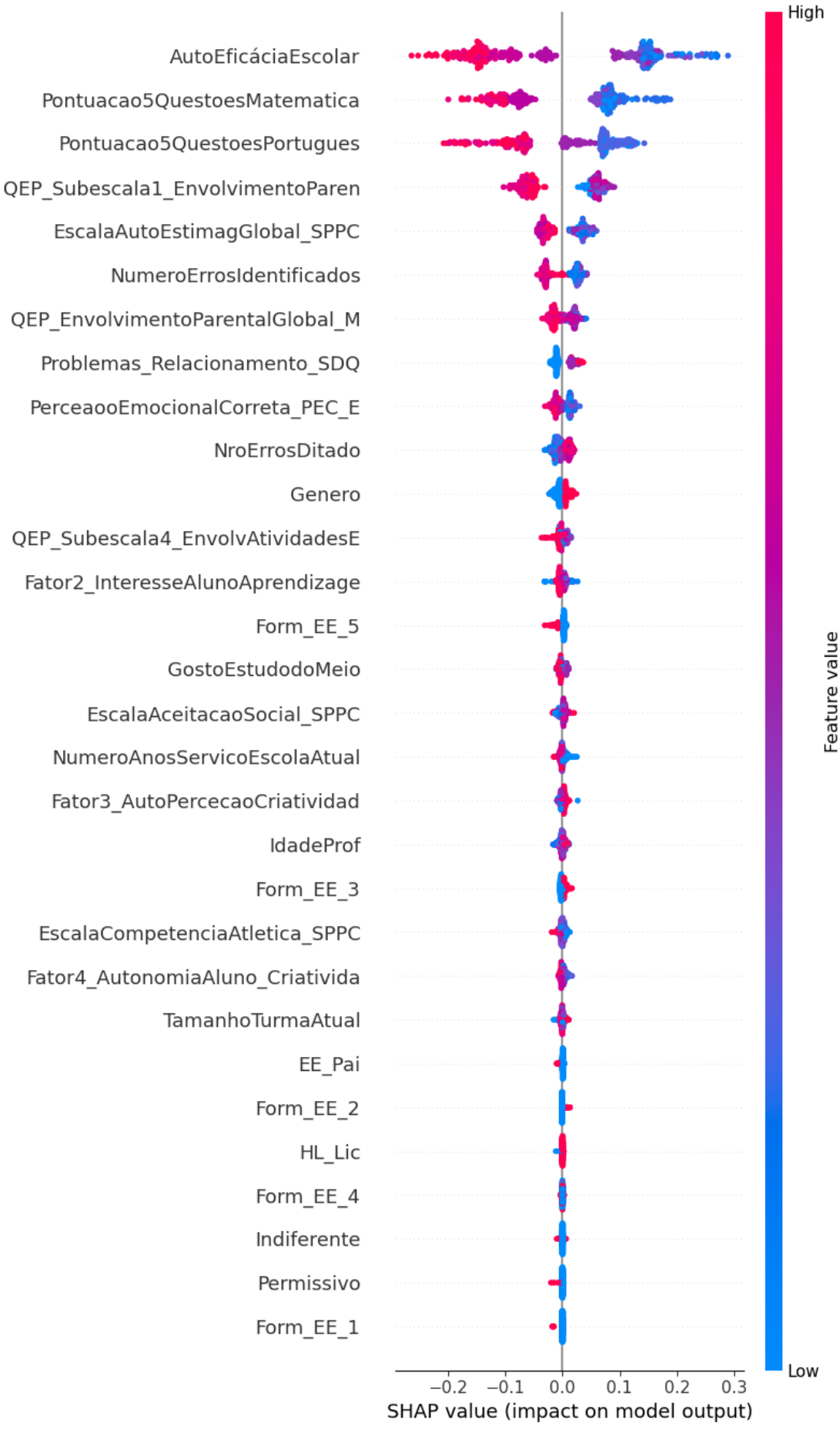

Figure C 10 *– SHAP values plot of all the 30 features selected by the machine learning model. This output results from the application of the method in the test set.*

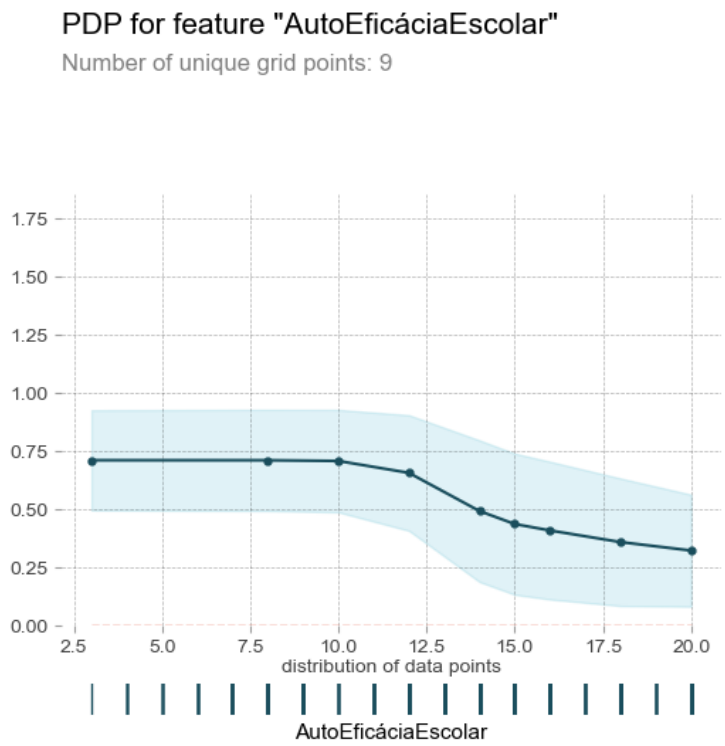

Figure C 11 *– Partial Dependence plot of* ***AutoEficáciaEscolar*** *applied on the test set.*

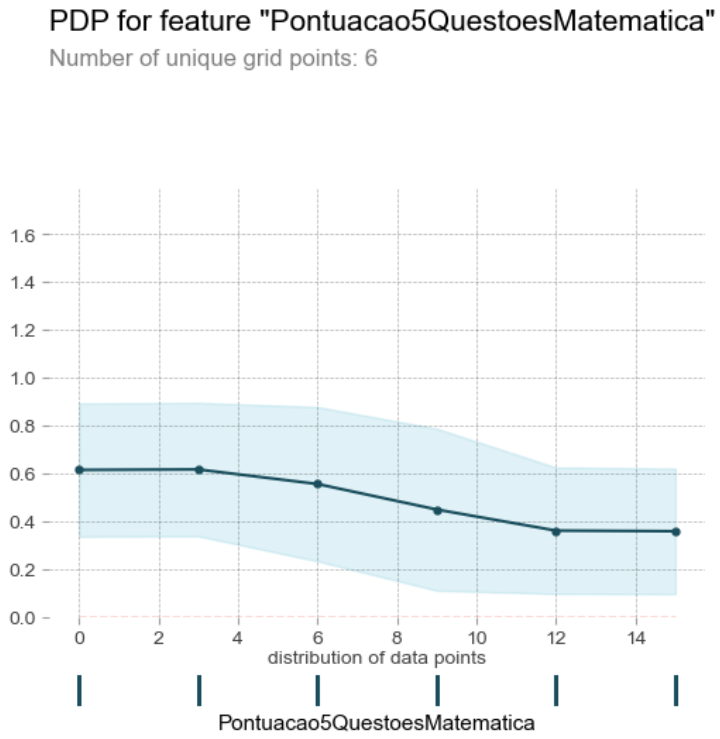

Figure C 12 *– Partial Dependence plot of* ***Pontuacao5QuestoesMatematica*** *applied on the test set.*

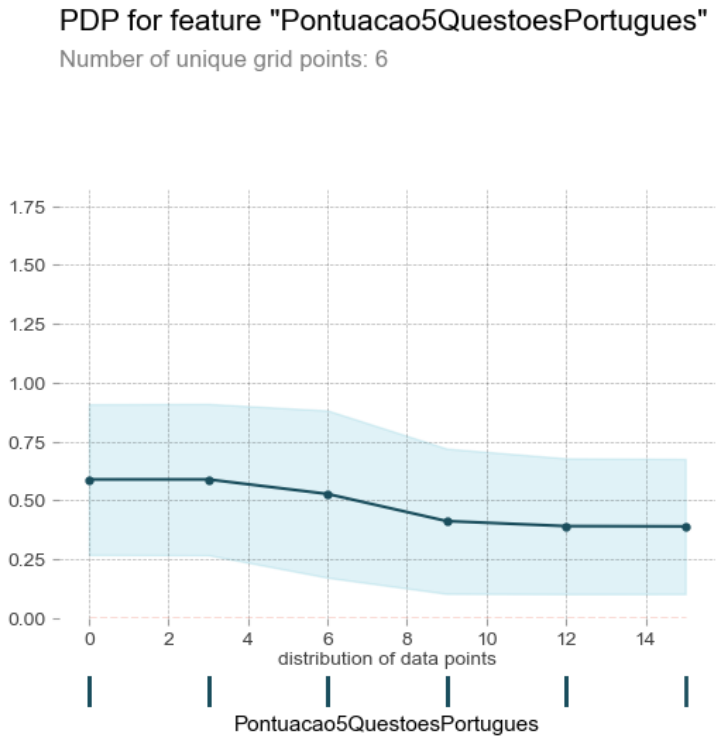

Figure C 13 *– Partial Dependence plot of* ***Pontuacao5QuestoesPortugues*** *applied on the test set.*

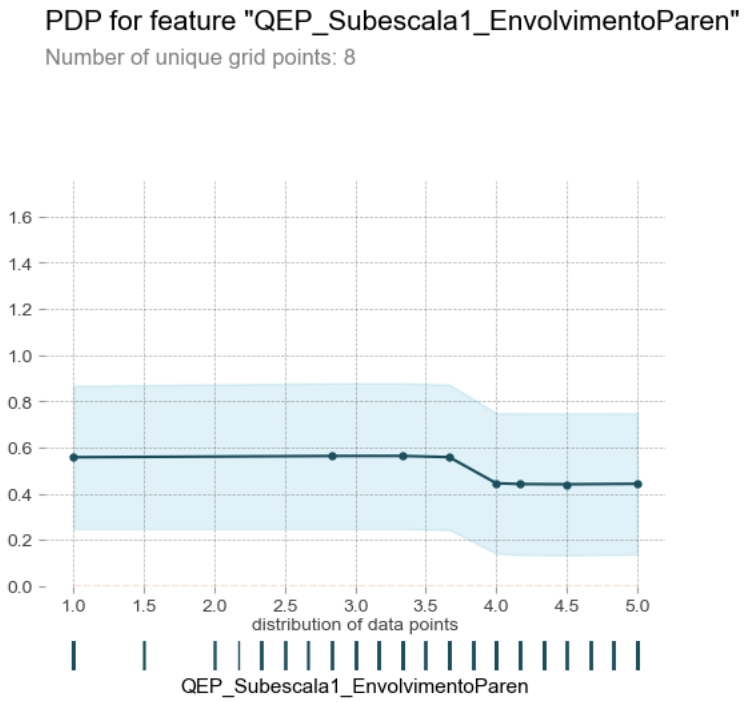

Figure C 14 *– Partial Dependence plot of* ***QEP_Subescala1_EnvolvimentoParen*** *applied on the test set.*

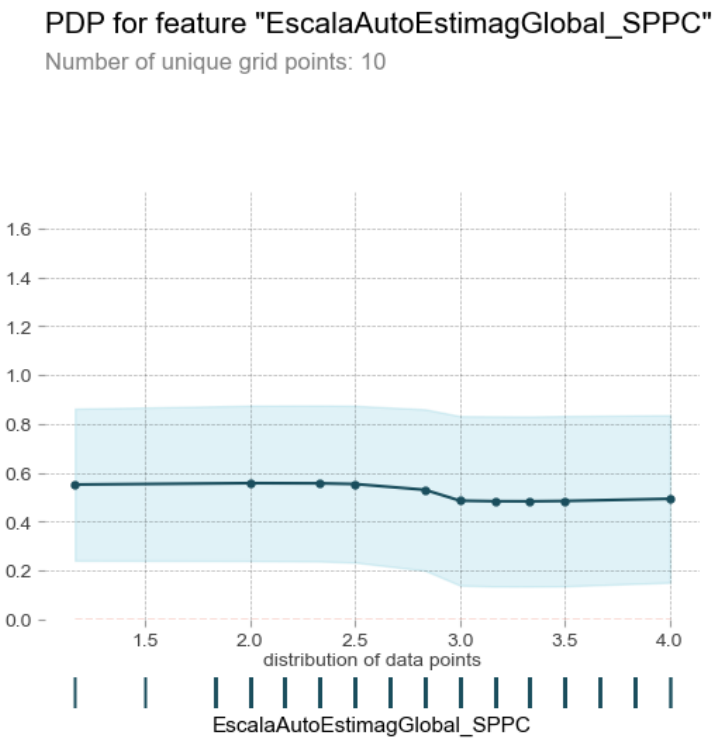

Figure C 15 *– Partial Dependence plot of* ***EscalaAutoEstimagGlobal_SPPC*** *applied on the test set.*

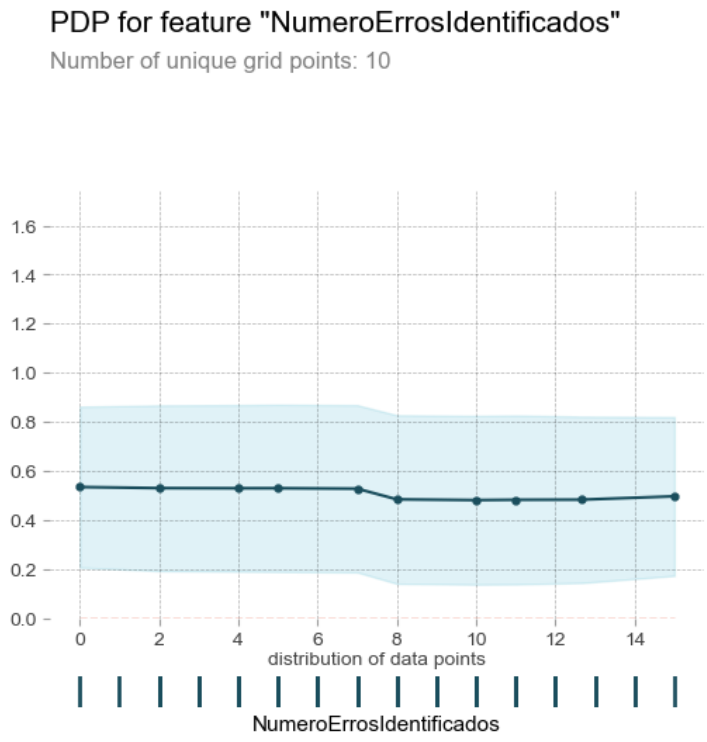

Figure C 16 *– Partial Dependence plot of* ***NumeroErrosIdentificados*** *applied on the test set.*

APPENDIX D

Formal analysis of teachers' features

	IDProfessor	Number of students in the test set	Number of students failing	Model Recall	Teacher Recall	Model F1	Teacher F1
0	p.p.a.15.24	8	2	1.0	1.0	1.000000	0.400000
1	p.p.ha.33.64	3	2	1.0	1.0	1.000000	0.800000
2	p.p.ca.11.15	6	2	1.0	1.0	1.000000	0.571429
3	p.p.ca.10.12	7	2	1.0	1.0	1.000000	0.500000
4	p.l.da.18.28	4	1	1.0	1.0	0.500000	0.400000
5	p.l.ia.36.72	5	3	1.0	1.0	0.857143	0.750000
6	p.p.ba.8.10	3	2	1.0	1.0	1.000000	0.800000
7	p.p.ja.39.78	4	1	1.0	1.0	1.000000	0.400000
8	p.l.da.17.27	8	6	1.0	1.0	0.923077	0.857143
9	p.p.a.12.17	5	2	1.0	1.0	1.000000	0.666667
10	p.l.da.20.31	5	4	1.0	1.0	1.000000	0.888889
11	p.l.ea.22.38	5	4	1.0	1.0	1.000000	0.888889
12	p.l.da.21.34	4	3	1.0	1.0	1.000000	0.857143
13	p.p.ba.4.6	6	3	1.0	1.0	1.000000	0.750000
14	p.p.a.15.25	5	2	1.0	1.0	1.000000	0.666667
15	p.p.ha.33.63	4	3	1.0	1.0	1.000000	0.857143
16	p.p.ha.35.68	4	1	1.0	1.0	1.000000	0.400000
17	p.l.da.21.35	11	5	1.0	1.0	0.833333	0.625000
18	p.l.ea.23.44	3	2	1.0	0.0	1.000000	0.000000
19	p.p.aa.2.3	3	1	1.0	1.0	1.000000	0.500000
20	p.p.a.13.20	6	2	1.0	1.0	1.000000	0.500000
21	p.l.ia.36.71	5	3	1.0	1.0	0.857143	0.750000
22	p.l.ga.31.60	5	2	1.0	1.0	0.666667	0.571429
23	p.p.ha.33.65	6	1	1.0	1.0	1.000000	0.333333
24	p.p.a.12.16	5	3	1.0	1.0	1.000000	0.857143
25	p.p.fa.27.50	4	2	1.0	0.5	1.000000	0.400000
26	p.p.ca.11.14	3	1	1.0	1.0	1.000000	0.500000
27	p.p.fa.27.49	3	2	1.0	1.0	1.000000	0.800000
28	p.l.da.21.33	6	3	1.0	1.0	0.857143	0.666667
29	p.p.aa.2.2	4	3	1.0	1.0	1.000000	0.857143
30	p.p.ja.40.79	3	1	1.0	1.0	1.000000	0.500000
31	p.p.fa.28.52	2	1	1.0	1.0	1.000000	0.666667
32	p.p.ha.35.69	6	3	1.0	1.0	1.000000	0.666667
33	p.p.a.14.22	4	3	1.0	1.0	1.000000	0.857143
34	p.l.ea.24.41	5	2	1.0	1.0	1.000000	0.666667
35	p.p.fa.29.53	2	1	1.0	1.0	1.000000	0.666667
36	p.l.ga.31.59	4	2	1.0	1.0	1.000000	0.666667
37	p.p.ha.35.70	3	2	1.0	1.0	1.000000	0.800000
38	p.l.ea.22.40	2	1	1.0	1.0	1.000000	0.666667

Table D 1 *– Model and teacher predictions of students' performance on the* ***test set****, for the teachers who* ***would benefit*** *from using the ML model to predict failure. The first column identifies the teachers in the test set; the second column is respective to the number of students each evaluation is based on; The third column quantifies the number of failing students in the test set for each teacher; the last four columns are respective to the performance evaluation of teachers and the model, using recall and f1-score.*

Figure D1 and **Table D2** present the plots for the SSE, Silhouette Coefficient, and Davies Bouldin Score, used to choose the optimal number of clusters in the teachers' data set. The explanation of these techniques and how they are used to find the optimal number of clusters may be found in **Appendix F** for further understanding of these outputs.

The same thing applies to **Figure E1** and **Table E1.**

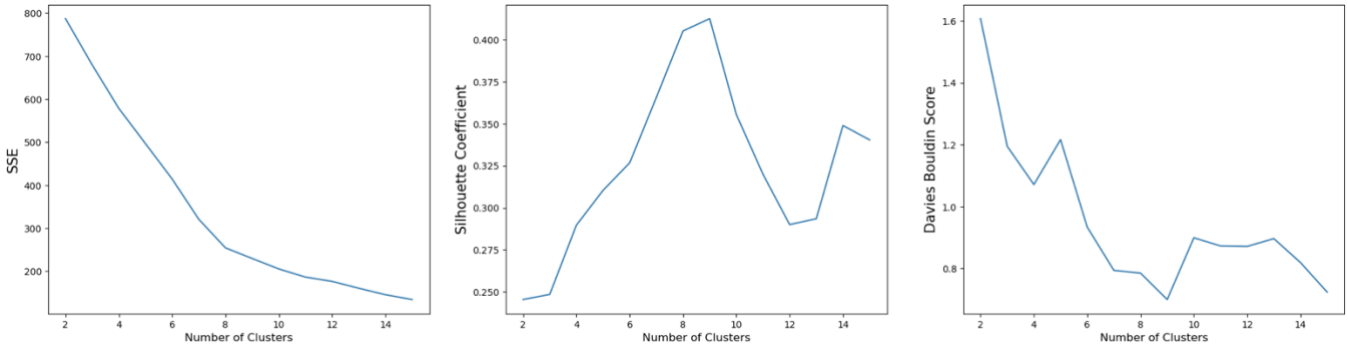

Figure D 1 *– SSE, Silhouette Coefficient and Davies Bouldin Score plots for different numbers of clusters in the teachers' data set, varying from 2 to 15 clusters.*

Number of Clusters	SSE	Silhouette Coefficient	Davies Bouldin Score
2	787.122741	0.245473	1.607295
3	679.709031	0.248527	1.194905
4	578.438567	0.289667	1.071212
5	496.524717	0.310372	1.216373
6	414.396675	0.326764	0.933945
7	320.683603	0.365658	0.793905
8	254.300664	0.405138	0.785094
9	229.692582	0.412461	0.699629
10	205.337412	0.355484	0.899274
11	186.382486	0.319983	0.872889
12	176.391620	0.290045	0.871271
13	160.486015	0.293523	0.896593
14	145.450931	0.348913	0.819094
15	134.187198	0.340372	0.723709

Table D 2 *- SSE, Silhouette Coefficient and Davies Bouldin Score summary table for different numbers of clusters in the teachers' data set, varying from 2 to 15 clusters.*

From **Figure D1**, we can visually verify that nine clusters minimize the Davies-Bouldin Index. As for the Silhouette Coefficient, nine clusters visually seem to maximize this metric. These conclusions are both corroborated by the results in **Table D2**. Since these two metrics agree, and the SSE plot finds its elbow around eight clusters (being particularly close to nine), **nine clusters** were chosen as the optimal number for this data set.

Table D3 summarizes the results of the Chi-Square Test for Independence on each feature of the teachers' data set. Once again, **Appendix F** provides more information about this test and may be used for further consultation. The same document can be referred to when analyzing **Tables E2** and **E3**.

Feature	Chi-Square Statistic	P-value
GeneroProf	0.000000	1.000000
HL_Lic	0.000000	1.000000
HL_Mst	0.007427	0.931322
SituaçãoProfissional	0.025815	0.872353
TamanhoTurmaAtual	7.150651	0.711150
IdadeProf	28.644467	0.327496
NumeroAnosServico	13.592778	0.850529
NumeroAnosServicoEscolaAtual	18.612660	0.547116
Indiferente	0.000000	1.000000
Permissivo	0.000000	1.000000
Persuasivo	0.383442	0.535767
Autoritário	0.170669	0.679518

Table D 3 *– Chi – square test for Independence performed for each feature in the teachers' data set. These test results are respective to the binary variable benefit_or_not.*

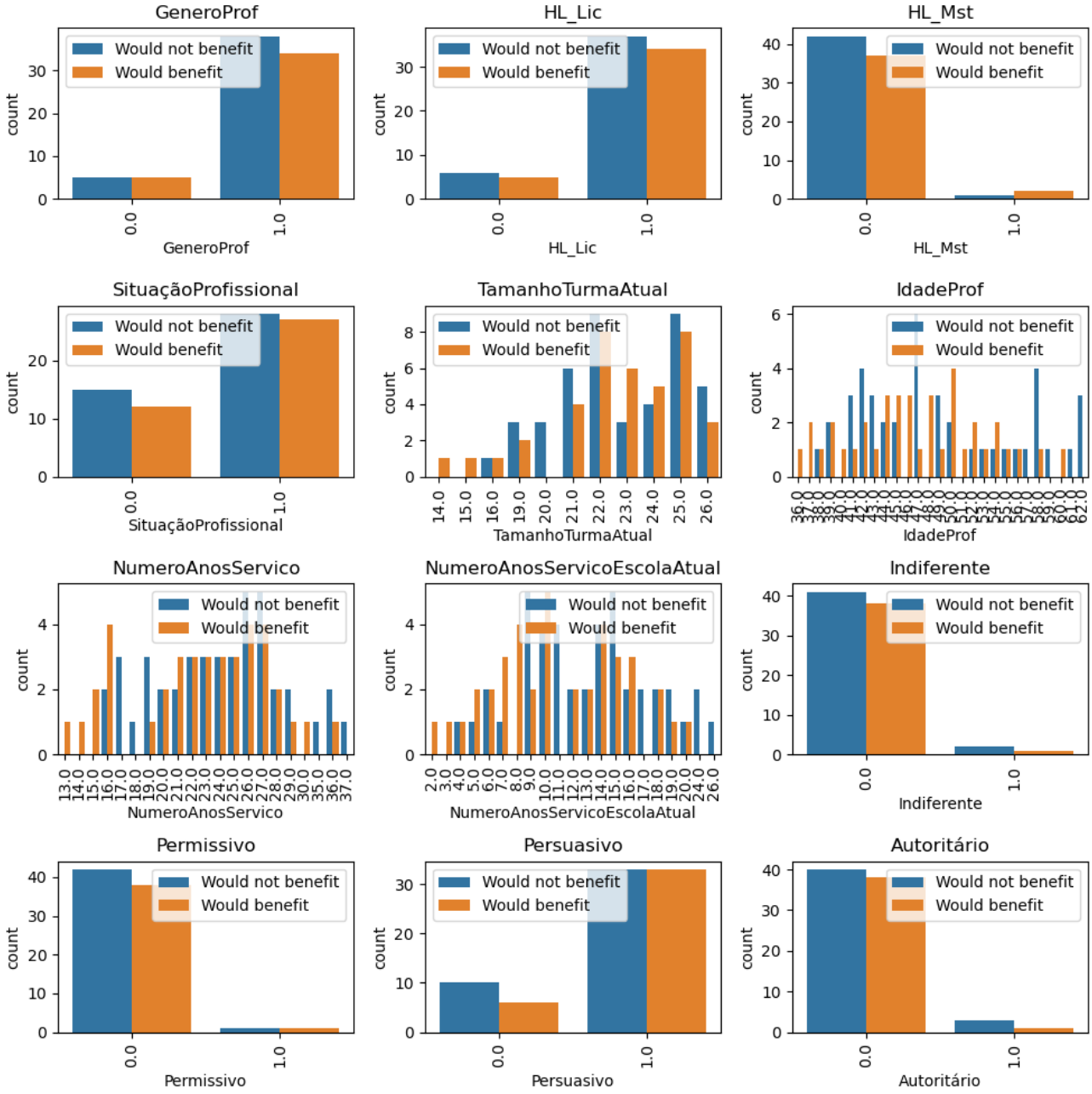

Figure D 2 *- Bar plots reflecting the distribution of the (**discrete**) features in the teachers' data set, distinguished between the teachers who would benefit from using the ML model to predict academic performance (**orange bars**) – benefit_or_not = 1 - and those who would not (**blue bars**) - benefit_or_not = 0.*

APPENDIX E

Formal analysis of students' features

Figure E1 and **Table E1** summarize the search results for the optimal number of clusters in the students' data set. As referred to in the analogous section of **Appendix D,** one can find the explanation of the use of these techniques in this context, in **Appendix F**.

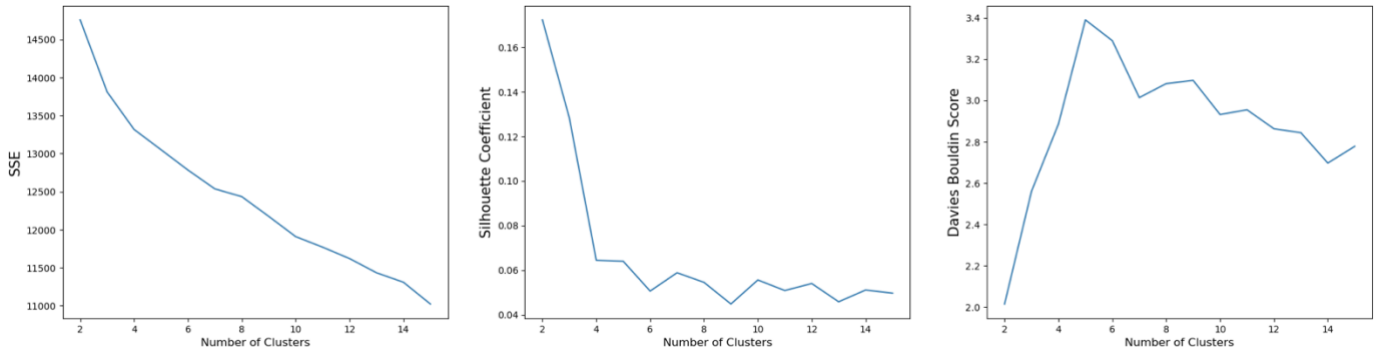

Figure E 1 *- SSE, Silhouette Coefficient and Davies Bouldin Score plots for different numbers of clusters in the students' data set, varying from 2 to 15 clusters.*

Number of Clusters	SSE	Silhouette Coefficient	Davies Bouldin Score
2	14757.687959	0.172133	2.014297
3	13811.343643	0.128163	2.559990
4	13317.303662	0.064415	2.885666
5	13051.096625	0.063990	3.389496
6	12783.494451	0.050623	3.288662
7	12535.799289	0.058838	3.013181
8	12433.568152	0.054525	3.080977
9	12174.296527	0.044834	3.096586
10	11908.093832	0.055594	2.931286
11	11769.576977	0.050865	2.954265
12	11619.573925	0.054013	2.862396
13	11433.027709	0.045830	2.843356
14	11308.468461	0.051098	2.696367
15	11022.294976	0.049685	2.777303

Table E 1 *- SSE, Silhouette Coefficient and Davies Bouldin Score summary table for different numbers of clusters in the students' data set, varying from 2 to 15 clusters.*

Defining the optimal number of clusters based on **Figure E1** and **Table E1** could be more straightforward. The elbow of the SSE plot is not very evident, but it is roughly between six and ten clusters. We can narrow this interval by looking at the other two plots. Both metrics are optimized for a number of clusters equal to two. However, this option was disregarded, considering it does not fall within the specified interval defined in the interpretation of the SSE plot. Nevertheless, we can see that the Silhouette Coefficient has a local maximum in seven clusters, which also provides a local minimum for the Davies-Bouldin Score. Therefore, **seven clusters** were considered to be the optimal number of clusters for this data set.

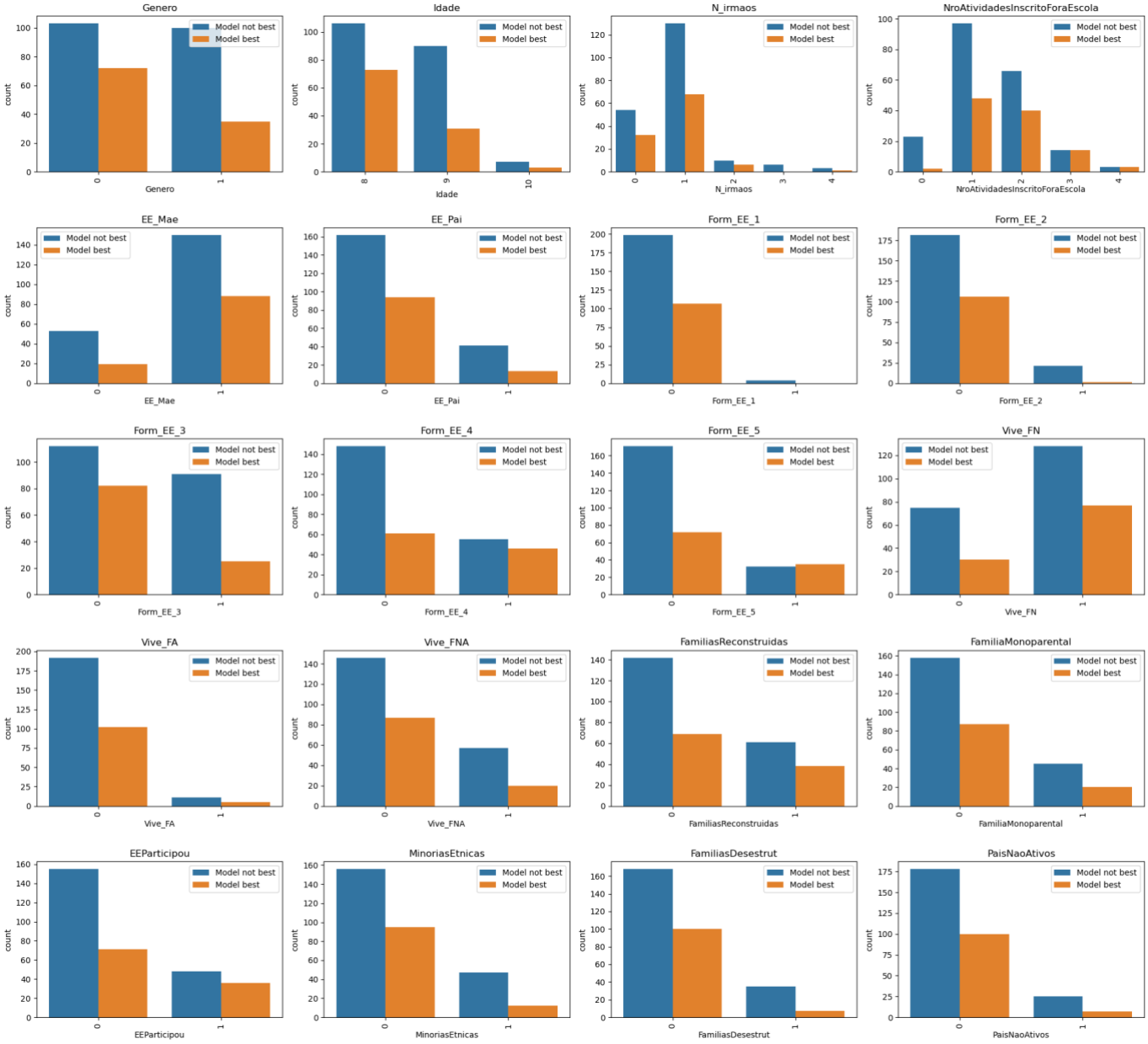

Figure E 2 a) *- Bar plots reflecting the distribution of the* ***discrete*** *features in the students' data set, distinguished between the students for which the model's prediction was correct while teacher's was not (**orange bars**) – model_best = 1 - and those for which this condition is not met (**blue bars**) – model_best = 0. Due to its dimension, the figure had to be divided into two figures, namely **Figure E 2a)** and **Figure E 2b)**.*

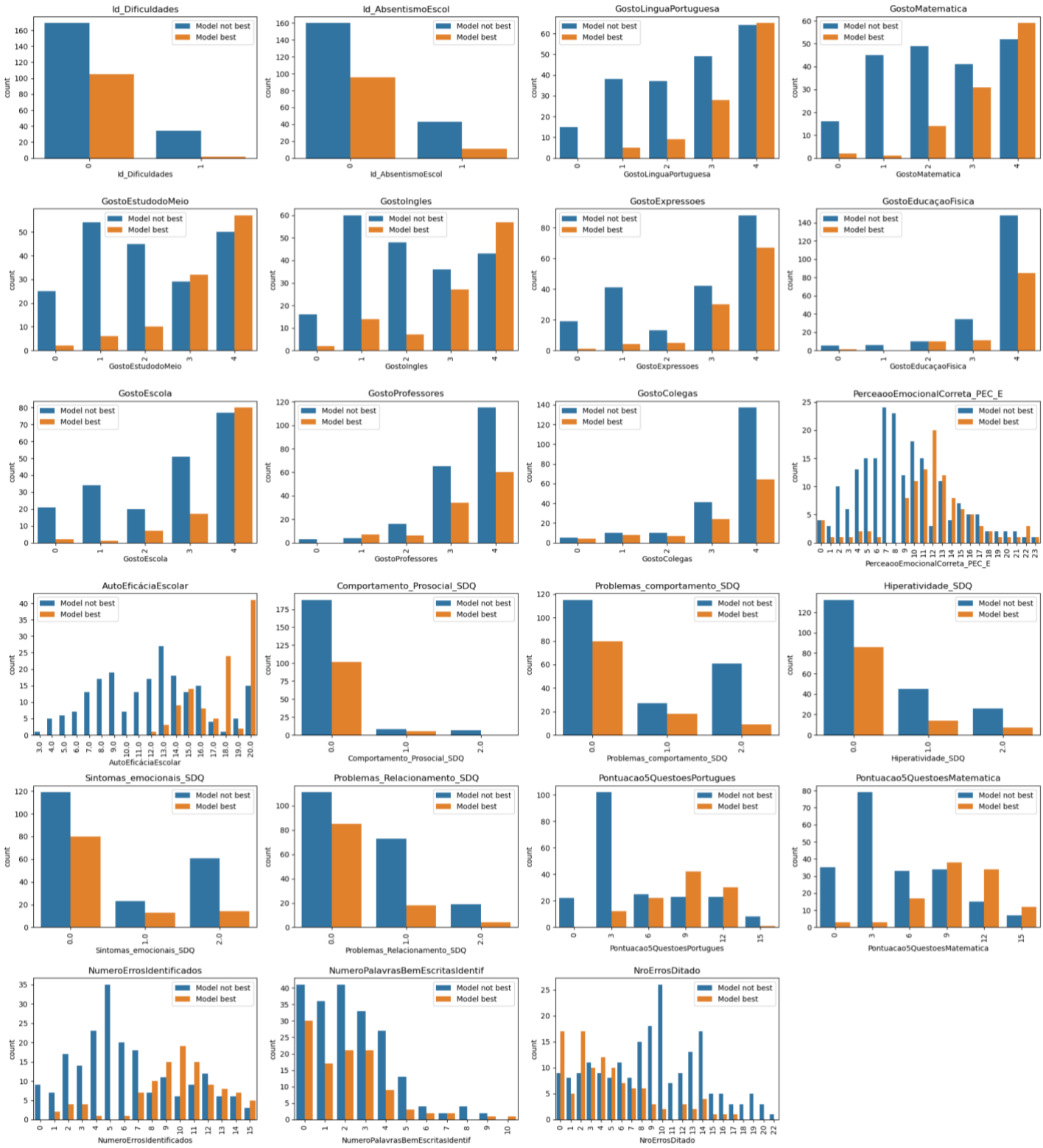

Figure E 2 b) *- Bar plots reflecting the distribution of the* ***discrete*** *features in the students' data set, distinguished between the students for which the model's prediction was correct while teacher's was not (**orange bars**) – model_best = 1 - and those for which this condition is not met (**blue bars**) – model_best = 0. Due to its dimension, the figure had to be divided into two figures, namely **Figure E 2a)** and **Figure E 2b)**.*

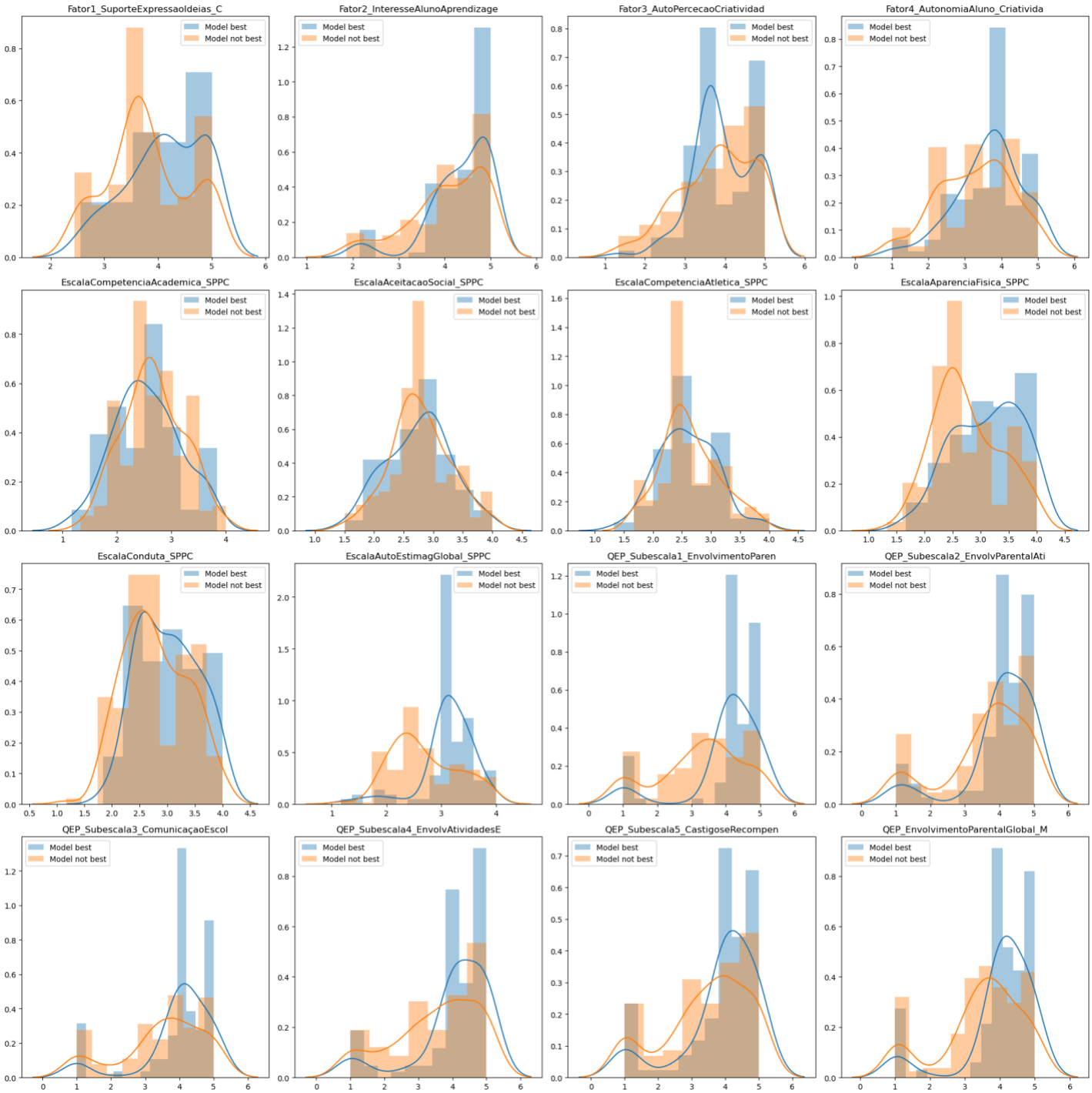

Figure E 3 *- Histograms and Kernel Density plots reflecting the distribution of the* ***continuous*** *features in the students' data set, distinguished between the students for which the model's prediction was correct while the teacher's was not (**blue bars**) – model_best = 1 - and those for which this condition is not true (**orange bars**) – model_best = 0.*

As mentioned in the analogous section of **Appendix D**, the analysis of **Tables E2** and **E3** on the Chi-Square and Kolmogorov-Smirnov tests may be complemented with the explanation of these tests found in **Appendix F**.

Feature	Chi-Square Statistic	P-value
Genero	7.148698	0.007502
Idade	7.436525	0.024276
N_irmaos	3.664436	0.453320
NroAtividadesInscritoForaEscola	11.997509	0.017370
EE_Mae	2.292267	0.130020
EE_Pai	2.619870	0.105533
Form_EE_1	0.869012	0.351229
Form_EE_2	8.037678	0.004581
Form_EE_3	12.882415	0.000332
Form_EE_4	7.353863	0.006692
Form_EE_5	10.898467	0.000962
Vive_FN	2.100771	0.147225
Vive_FA	0.000149	0.990272
Vive_FNA	2.823560	0.092890
FamiliasReconstruidas	0.727652	0.393645
FamiliaMonoparental	0.322630	0.570031
EEParticipou	3.058482	0.080317
MinoriasEtnicas	5.728271	0.016694
FamiliasDesestrut	5.965089	0.014592
PaisNaoAtivos	1.937682	0.163920

Feature	Chi-Square Statistic	P-value
Id_Dificuldades	13.698840	0.000215
Id_AbsentismoEscol	5.056046	0.024540
GostoLinguaPortuguesa	36.915226	0.000000
GostoMatematica	49.244102	0.000000
GostoEstudodoMeio	56.566506	0.000000
GostoIngles	48.184722	0.000000
GostoExpressoes	27.976894	0.000013
GostoEducaçaoFisica	8.547202	0.073469
GostoEscola	44.682551	0.000000
GostoProfessores	6.224300	0.183012
GostoColegas	2.314240	0.678178
PerceaooEmocionalCorreta_PEC_E	86.954263	0.000000
AutoEficáciaEscolar	145.436269	0.000000
Comportamento_Prosocial_SDQ	3.834448	0.147015
Problemas_comportamento_SDQ	18.782870	0.000083
Hiperatividade_SDQ	7.969163	0.018600
Sintomas_emocionais_SDQ	11.221431	0.003658
Problemas_Relacionamento_SDQ	18.520425	0.000095
Pontuacao5QuestoesPortugues	83.439792	0.000000
Pontuacao5QuestoesMatematica	90.347007	0.000000
NumeroErrosIdentificados	92.406899	0.000000
NumeroPalavrasBemEscritasIdentif	10.125865	0.429521
NroErrosDitado	63.433084	0.000004

Table E 2 *– Results of the Chi – square test for Independence performed for each discrete feature in the students' data set. This test results are respective to the binary variable model_best. Features which present* ***statistically significant*** *differences in the distributions based on model_best are* ***highlighted in blue****.*

Feature	Kolmogorov–Smirnov statistic	P-value
Fator1_SuporteExpressaoIdeias_C	0.304958	0.000003
Fator2_InteresseAlunoAprendizage	0.201878	0.005462
Fator3_AutoPercecaoCriatividad	0.182404	0.015988
Fator4_AutonomiaAluno_Criativida	0.196446	0.007454
EscalaCompetenciaAcademica_SPPC	0.114129	0.292118
EscalaAceitacaoSocial_SPPC	0.083329	0.675097
EscalaCompetenciaAtletica_SPPC	0.086782	0.626077
EscalaAparenciaFisica_SPPC	0.267069	0.000066
EscalaConduta_SPPC	0.178491	0.019588
EscalaAutoEstimagGlobal_SPPC	0.520418	0.000000
QEP_Subescala1_EnvolvimentoParen	0.544542	0.000000
QEP_Subescala2_EnvolvParentalAti	0.278164	0.000028
QEP_Subescala3_ComunicaçaoEscol	0.342203	0.000000
QEP_Subescala4_EnvolvAtividadesE	0.320473	0.000001
QEP_Subescala5_CastigoseRecompen	0.260992	0.000106
QEP_EnvolvimentoParentalGlobal_M	0.398923	0.000000

Table E 3 *– Kolmogorov-Smirnov test results performed for each continuous feature in the students' data set. This test results are respective to the binary variable model_best. Features which present* ***statistically significant*** *differences in the distributions based on model_best are* ***highlighted in blue****.*

```
best_estimator_2nd
```

```
Pipeline(steps=[('scaling', MinMaxScaler()),
                ('balancing', SMOTE(random_state=42)),
                ('selection',
                 RFE(estimator=DecisionTreeClassifier(random_state=42),
                     n_features_to_select=35)),
                ('classifier',
                 RandomForestClassifier(max_depth=6, min_samples_split=5,
                                        n_estimators=50, random_state=42))])
```

Figure E 4 *- Final model resulting from the grid-search process, for the second prediction model. The model is structured as a pipeline with a feature scaling, balancing, selection, and a classifier steps. For further comprehension of the involved steps, please refer to* ***Appendix F****.*

```
mask=best_estimator_2nd.named_steps['selection'].get_support()
chosen_feat = data_train.columns[mask]
chosen_feat
```

```
Index(['MinoriasEtnicas', 'FamiliasDesestrut', 'PaisNaoAtivos',
       'Id_Dificuldades', 'Id_AbsentismoEscol', 'GostoLinguaPortuguesa',
       'GostoMatematica', 'GostoEstudodoMeio', 'GostoIngles',
       'GostoExpressoes', 'GostoEducaçaoFisica',
       'Fator2_InteresseAlunoAprendizage', 'Fator4_AutonomiaAluno_Criativida',
       'PerceaooEmocionalCorreta_PEC_E', 'EscalaCompetenciaAcademica_SPPC',
       'EscalaAceitacaoSocial_SPPC', 'EscalaCompetenciaAtletica_SPPC',
       'EscalaConduta_SPPC', 'EscalaAutoEstimagGlobal_SPPC',
       'QEP_Subescala1_EnvolvimentoParen', 'QEP_Subescala2_EnvolvParentalAti',
       'QEP_Subescala3_ComunicaçaoEscol', 'QEP_Subescala4_EnvolvAtividadesE',
       'QEP_Subescala5_CastigoseRecompen', 'QEP_EnvolvimentoParentalGlobal_M',
       'AutoEficáciaEscolar', 'Comportamento_Prosocial_SDQ',
       'Problemas_comportamento_SDQ', 'Hiperatividade_SDQ',
       'Sintomas_emocionais_SDQ', 'Problemas_Relacionamento_SDQ',
       'PontuacaoSQuestoesPortugues', 'PontuacaoSQuestoesMatematica',
       'NumeroErrosIdentificados', 'NumeroPalavrasBemEscritasIdentif'],
      dtype='object')
```

Figure E 5 *- Selected features resulting from third step of the best model resulting from the grid-search. To be specific, the features were selected with Recursive Feature Elimination method, with Decision Tree Classifier as the base model.*

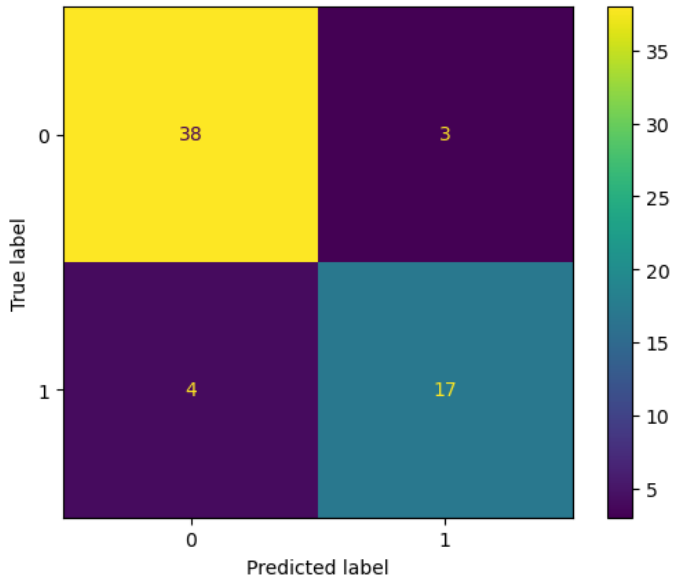

Figure E 6 *- Confusion matrix resulting from the evaluation of the grid-search best model on the test set. These results were obtained with the default probability threshold of 0.5.*

	precision	recall	f1-score	support
False	0.90	0.93	0.92	41
True	0.85	0.81	0.83	21
accuracy			0.89	62
macro avg	0.88	0.87	0.87	62
weighted avg	0.89	0.89	0.89	62

Figure E 7 *- Classification Report resulting from the evaluation of the grid-search best model on the test set. These results were obtained with the default probability threshold of 0.5.*

Figures E8 to **E14** are respective to the SHAP values and PD plots of the model predicting *model_best*. In the interpretation of **Figures C10** to **C16** in **Appendix C**, which correspond to analogous outputs but for the first model, it was provided a guide on how to interpret these results. Additionally, **Appendix F** may also be consulted with the description of these methods.

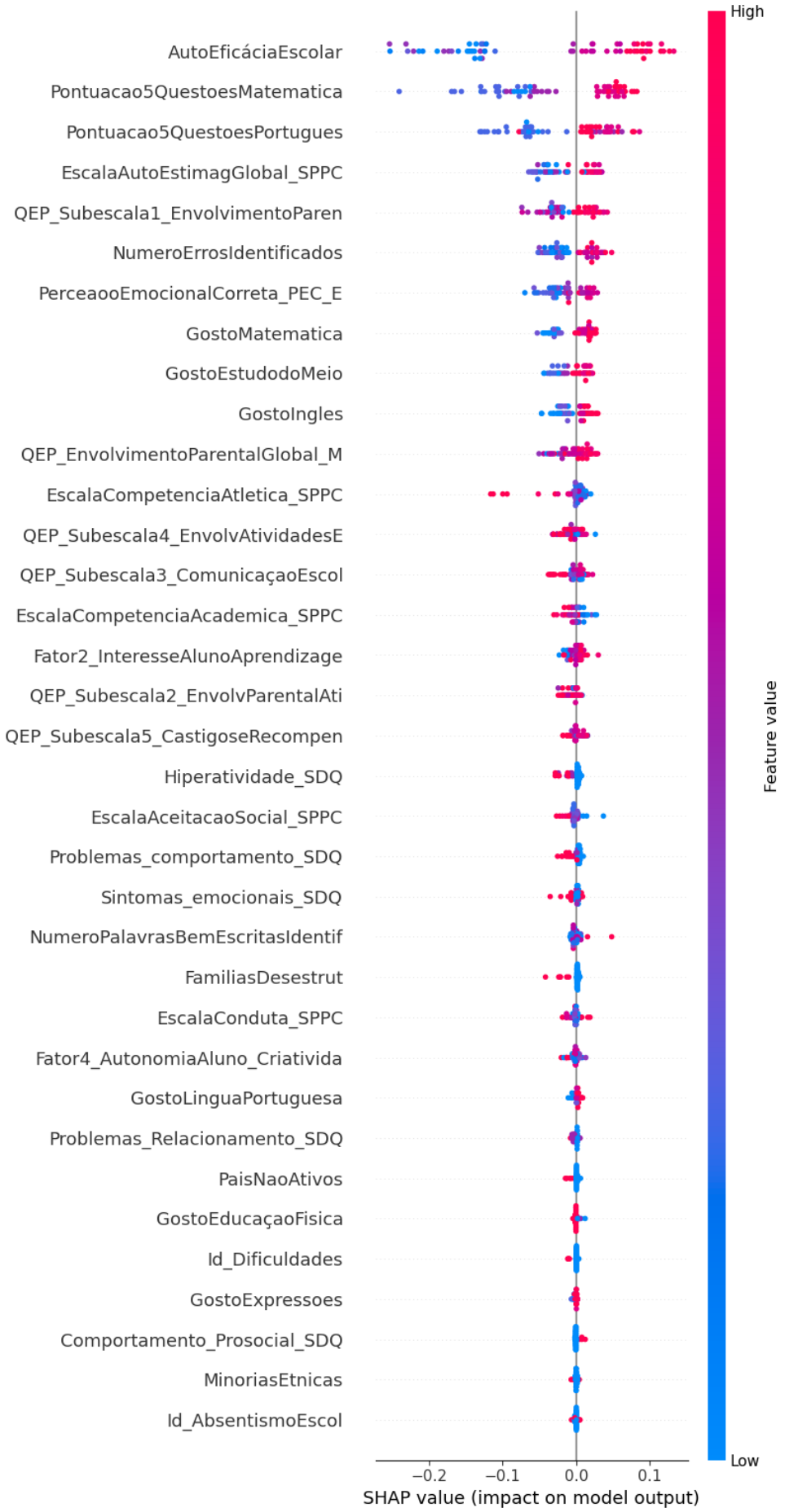

Figure E 8 *- SHAP values plot of all the 35 features selected by the machine learning model. This output results from the application of method to the test set.*

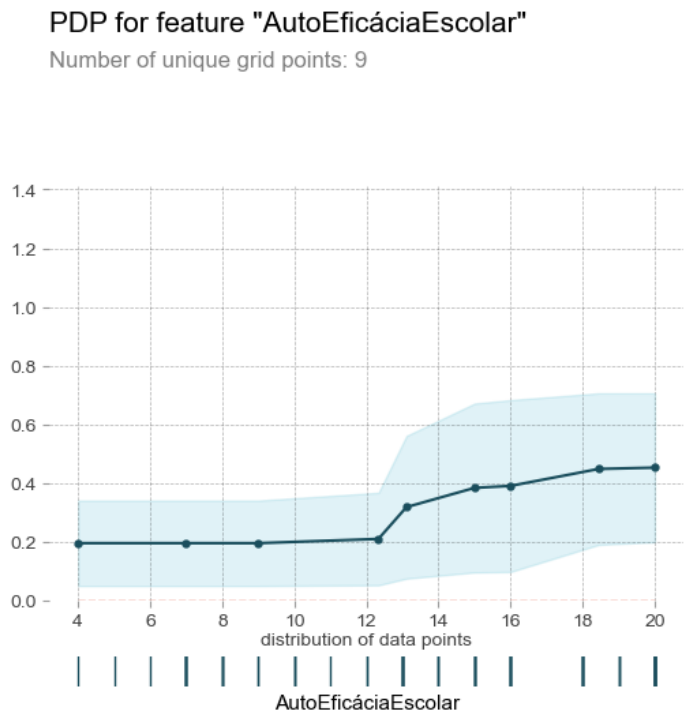

Figure E 9 *- Partial Dependence plot of* ***AutoEficáciaEscolar*** *applied on the test set.*

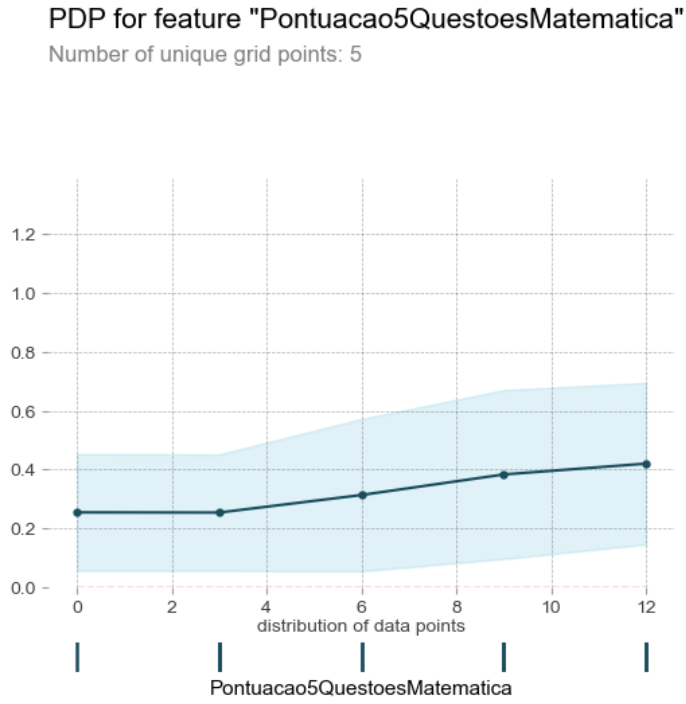

Figure E 10 *- Partial Dependence plot of* ***Pontuacao5QuestoesMatematica*** *applied on the test set.*

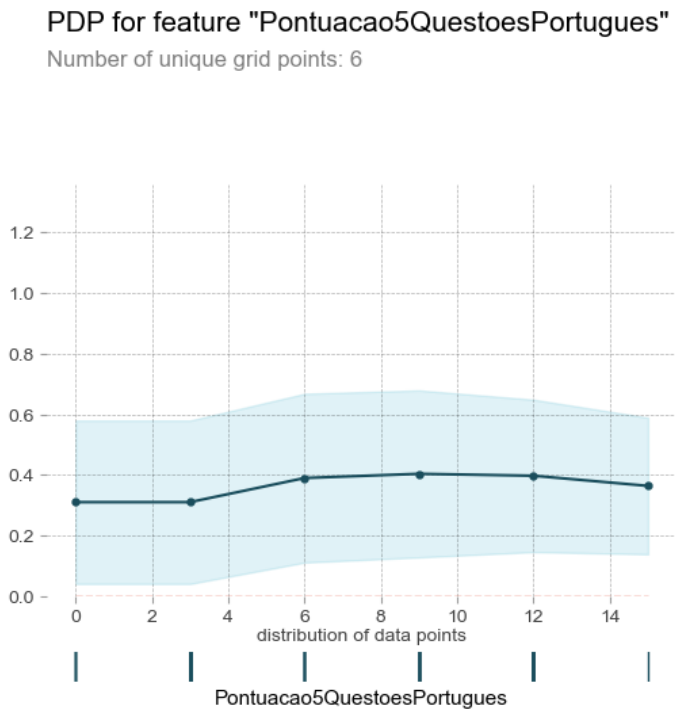

Figure E 11 *- Partial Dependence plot of P**ontuacao5QuestoesPortugues** applied on the test set.*

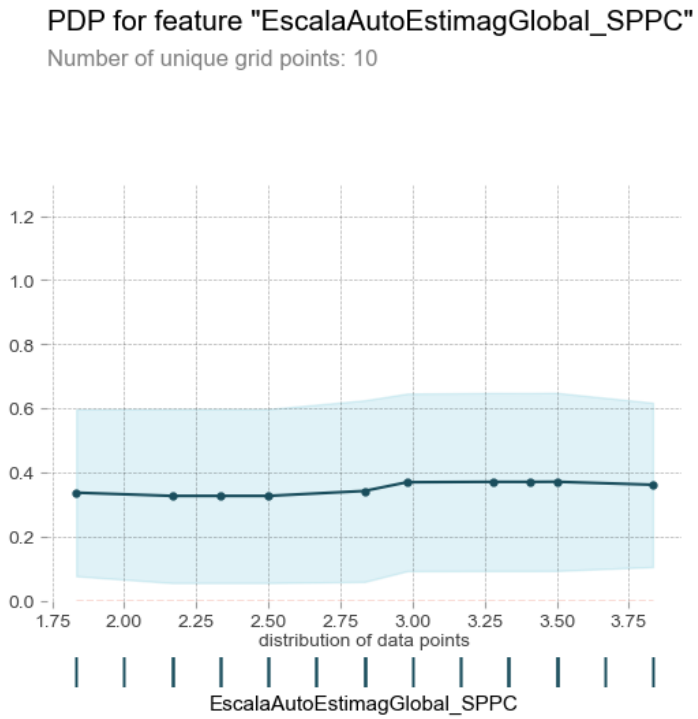

Figure E 12 *- Partial Dependence plot of **EscalaAutoEstimagGlobal** applied on the test set.*

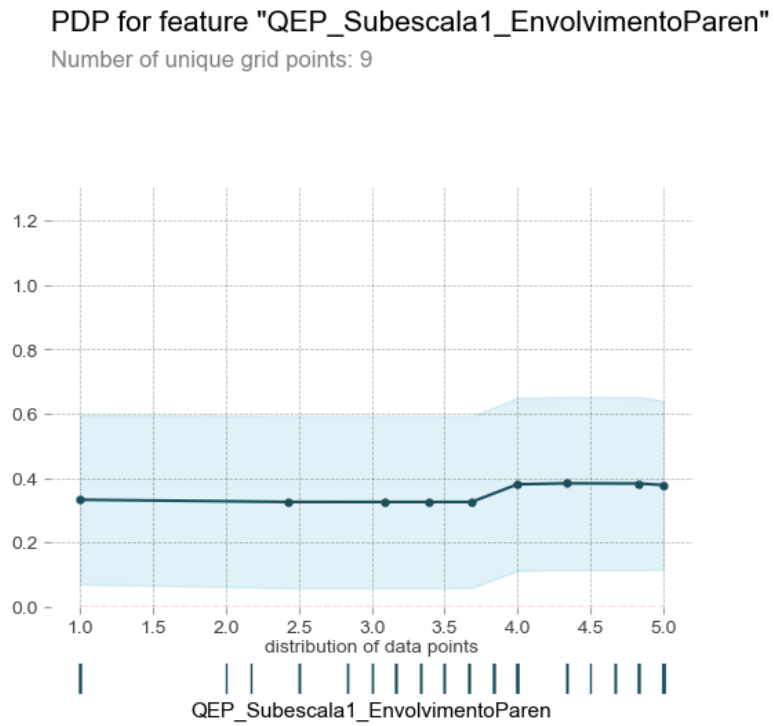

Figure E 13 *- Partial Dependence plot of* ***QEP_Subescala1_EnvolvimentoParen*** *applied on the test set.*

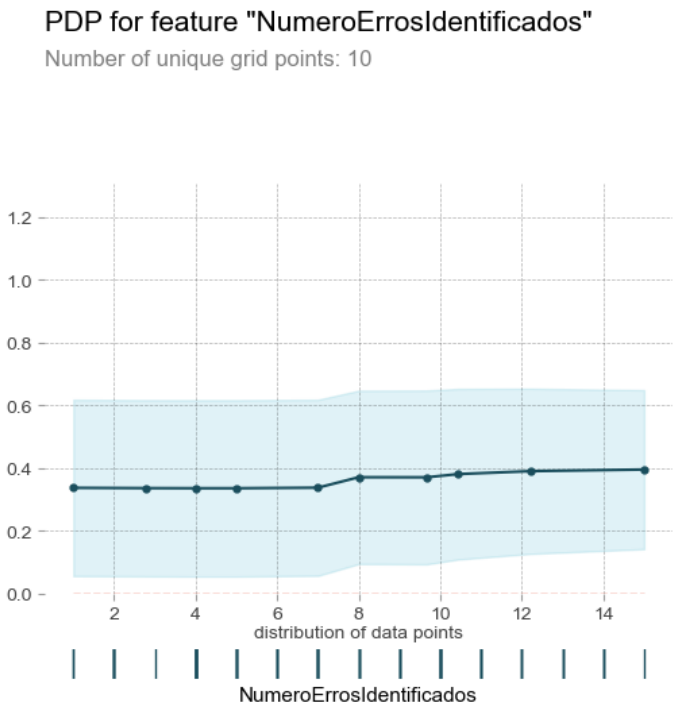

Figure E 14 *- Partial Dependence plot of* ***NumeroErrosIdentificados*** *applied on the test set.*

Tables E4 to **E9** are the values behind the visual representations in **Figures E15** to **E20**. The first and second columns of each table identify the number of students for which the model and teacher accurately predicted *Fail*, respectively, for each feature value. The third column references the number of students with the respective feature equal to each value. The last two columns are the proportion of the first and second columns on the third one, respectively. All these analyses were performed based on the **test set of the first model**.

For the variables **QEP_subescala1_EnvolvimentoParen** and **EscalaAutoEstimagGobal_SPPC**, even though they are continuous, their information was also presented with bar plots. Since they assume a finite number of values in the respective test set, representing it with bar plots allows for a more detailed comprehension of how these proportions evolve with the features' values.

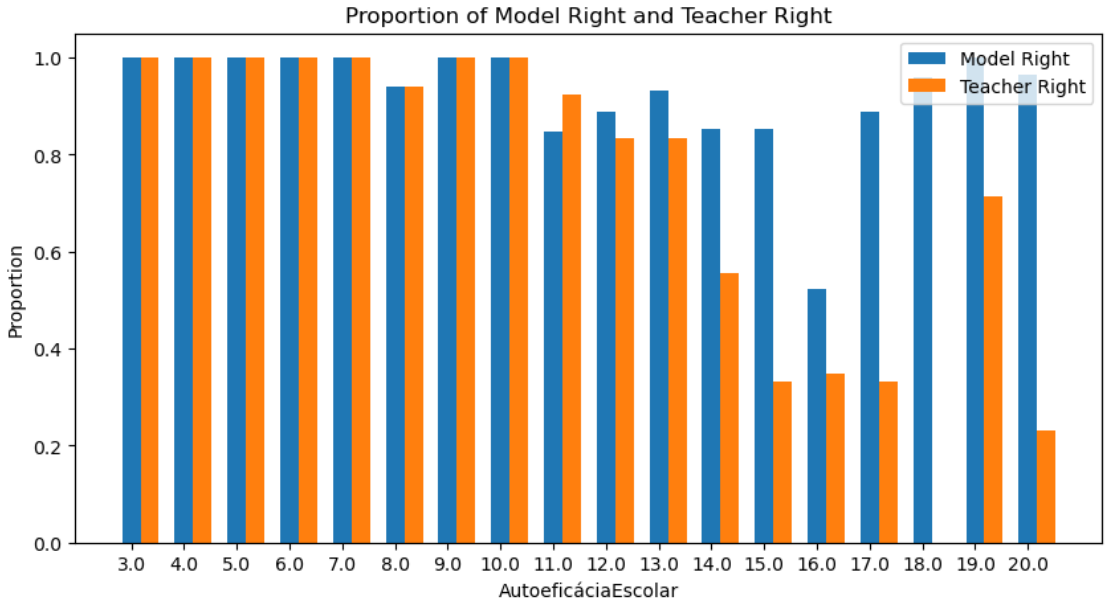

Figure E 15 *– Bar plot of the proportion of students for which the first model (**blue bars**) and the teacher (**orange bars**) accurately predicted their academic achievement (Fail) for each value of **AutoeficaciaEscolar**.*

AutoEficáciaEscolar	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
3.0	1	1	1	1.000000	1.000000
4.0	5	5	5	1.000000	1.000000
5.0	6	6	6	1.000000	1.000000
6.0	7	7	7	1.000000	1.000000
7.0	13	13	13	1.000000	1.000000
8.0	16	16	17	0.941176	0.941176
9.0	19	19	19	1.000000	1.000000
10.0	7	7	7	1.000000	1.000000
11.0	11	12	13	0.846154	0.923077
12.0	16	15	18	0.888889	0.833333
13.0	28	25	30	0.933333	0.833333
14.0	23	15	27	0.851852	0.555556
15.0	23	9	27	0.851852	0.333333
16.0	12	8	23	0.521739	0.347826
17.0	8	3	9	0.888889	0.333333
18.0	24	0	25	0.960000	0.000000
19.0	7	5	7	1.000000	0.714286
20.0	54	13	56	0.964286	0.232143

Table E 4 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of* ***AutoeficaciaEscolar****.*

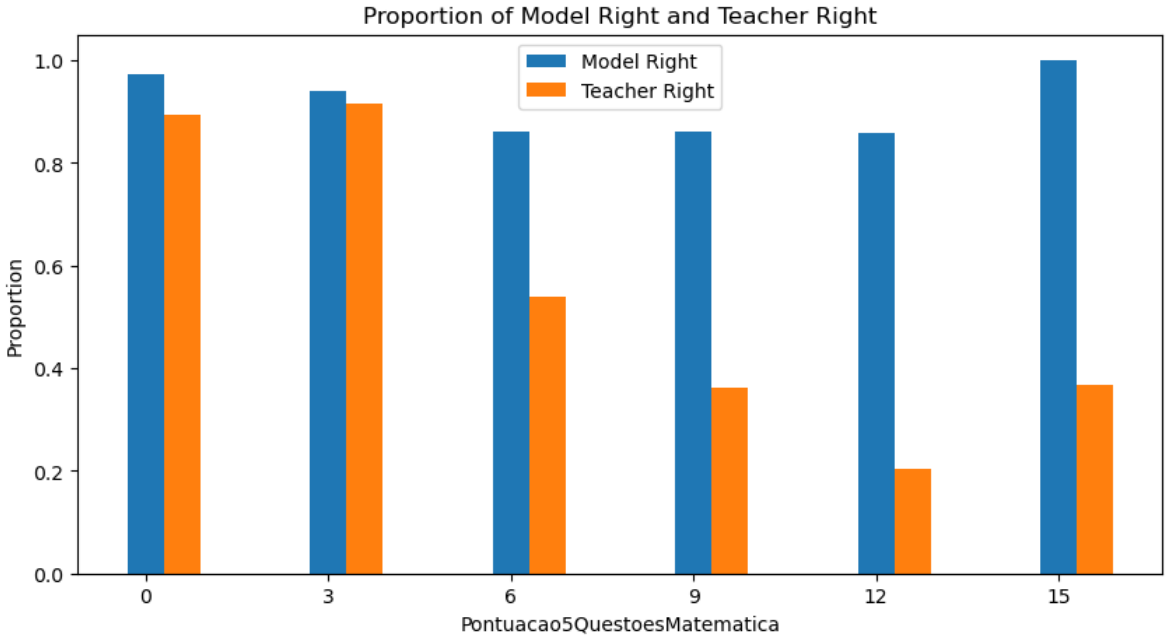

Figure E 16 *– Bar plot of the proportion of students for which the first model (****blue bars****) and the teacher (****orange bars****) accurately predicted their academic achievement (Fail) for each value of* ***PontuacaoQuestoesMatematica****.*

Pontuacao5QuestoesMatematica	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
0	37	34	38	0.973684	0.894737
3	77	75	82	0.939024	0.914634
6	43	27	50	0.860000	0.540000
9	62	26	72	0.861111	0.361111
12	42	10	49	0.857143	0.204082
15	19	7	19	1.000000	0.368421

Table E 5 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of* ***Pontuacao5QuestoesMatematica****.*

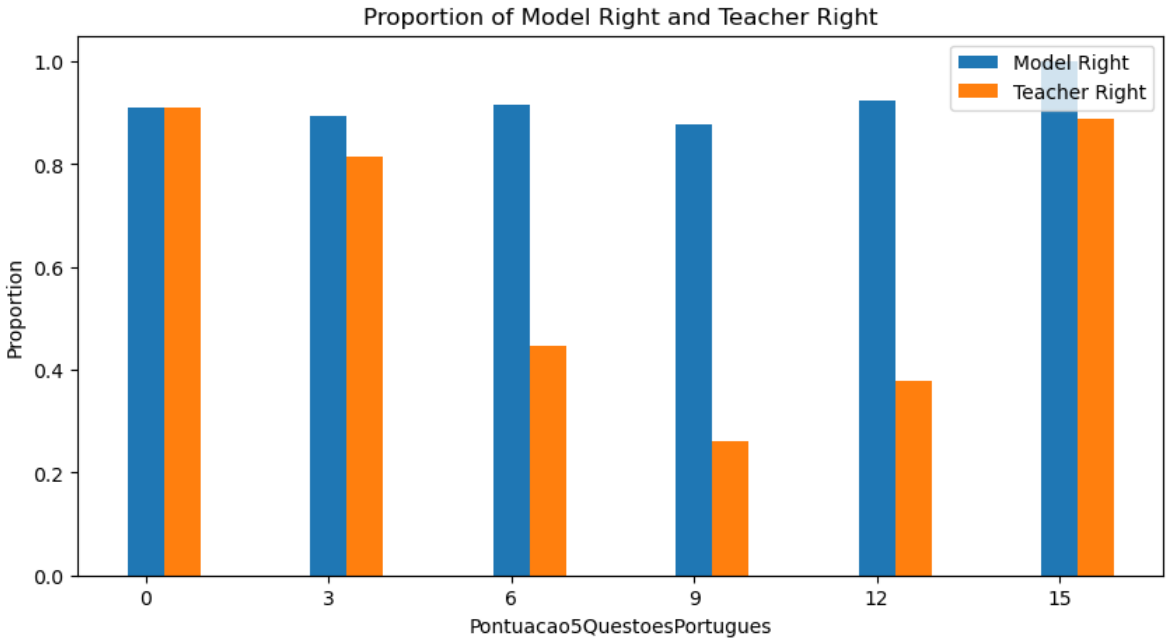

Figure E 17 *– Bar plot of the proportion of students for which the first model (****blue bars****) and the teacher (****orange bars)*** *accurately predicted their academic achievement (Fail) for each value of* ***Pontuacao5QuestoesPortugues****.*

Pontuacao5QuestoesPortugues	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
0	20	20	22	0.909091	0.909091
3	102	93	114	0.894737	0.815789
6	43	21	47	0.914894	0.446809
9	57	17	65	0.876923	0.261538
12	49	20	53	0.924528	0.377358
15	9	8	9	1.000000	0.888889

Table E 6 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of* ***Pontuacao5QuestoesPortugues****.*

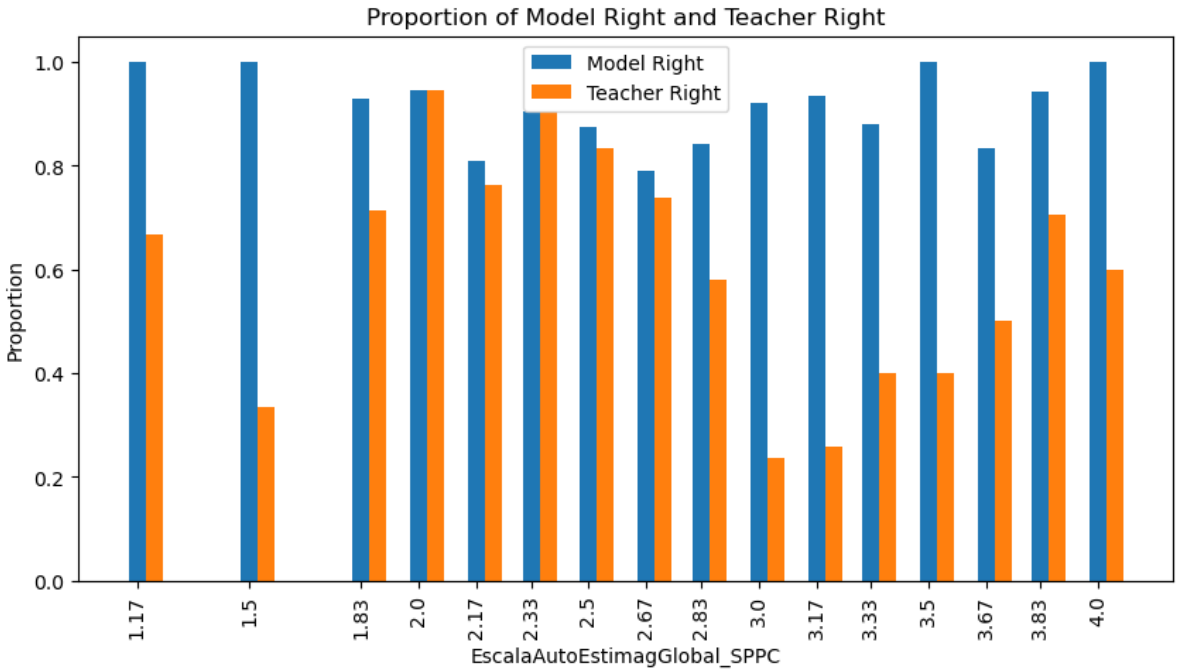

Figure E 18 *– Bar plot of the proportion of students for which the first model (**blue bars**) and the teacher (**orange bars)** accurately predicted their academic achievement (Fail) for each value of **EscalaAutoEstimagGlobal_SPPC**.*

EscalaAutoEstimagGlobal_SPPC	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
1.166667	3	2	3	1.000000	0.666667
1.500000	3	1	3	1.000000	0.333333
1.833333	13	10	14	0.928571	0.714286
2.000000	17	17	18	0.944444	0.944444
2.166667	17	16	21	0.809524	0.761905
2.333333	28	28	31	0.903226	0.903226
2.500000	21	20	24	0.875000	0.833333
2.666667	15	14	19	0.789474	0.736842
2.833333	16	11	19	0.842105	0.578947
3.000000	35	9	38	0.921053	0.236842
3.166667	29	8	31	0.935484	0.258065
3.333333	22	10	25	0.880000	0.400000
3.500000	30	12	30	1.000000	0.400000
3.666667	10	6	12	0.833333	0.500000
3.833333	16	12	17	0.941176	0.705882
4.000000	5	3	5	1.000000	0.600000

Table E 7 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of **EscalaAutoestimagGlobal_SPPC**.*

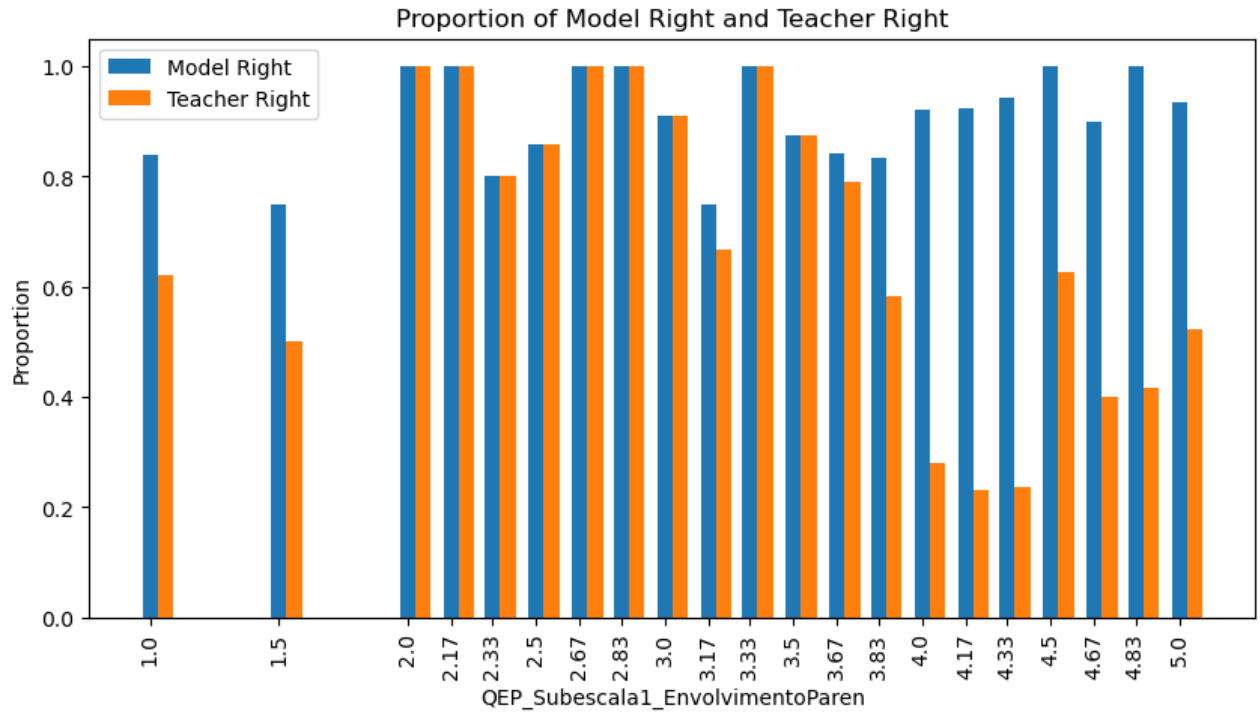

Figure E 19 *– Bar plot of the proportion of students for which the first model (**blue bars**) and the teacher (**orange bars**) accurately predicted their academic achievement (Fail) for each value of **QEP_Subescala1_EnvolvimentoParen**.*

QEP_Subescala1_EnvolvimentoParen	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
1.000000	31	23	37	0.837838	0.621622
1.500000	3	2	4	0.750000	0.500000
2.000000	5	5	5	1.000000	1.000000
2.166667	1	1	1	1.000000	1.000000
2.333333	8	8	10	0.800000	0.800000
2.500000	6	6	7	0.857143	0.857143
2.666667	4	4	4	1.000000	1.000000
2.833333	8	8	8	1.000000	1.000000
3.000000	10	10	11	0.909091	0.909091
3.166667	9	8	12	0.750000	0.666667
3.333333	16	16	16	1.000000	1.000000
3.500000	7	7	8	0.875000	0.875000
3.666667	16	15	19	0.842105	0.789474
3.833333	10	7	12	0.833333	0.583333
4.000000	46	14	50	0.920000	0.280000
4.166667	12	3	13	0.923077	0.230769
4.333333	16	4	17	0.941176	0.235294
4.500000	8	5	8	1.000000	0.625000
4.666667	9	4	10	0.900000	0.400000
4.833333	12	5	12	1.000000	0.416667
5.000000	43	24	46	0.934783	0.521739

Table E 8 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of **QEP_Subescala1_EnvolvimentoParen**.*

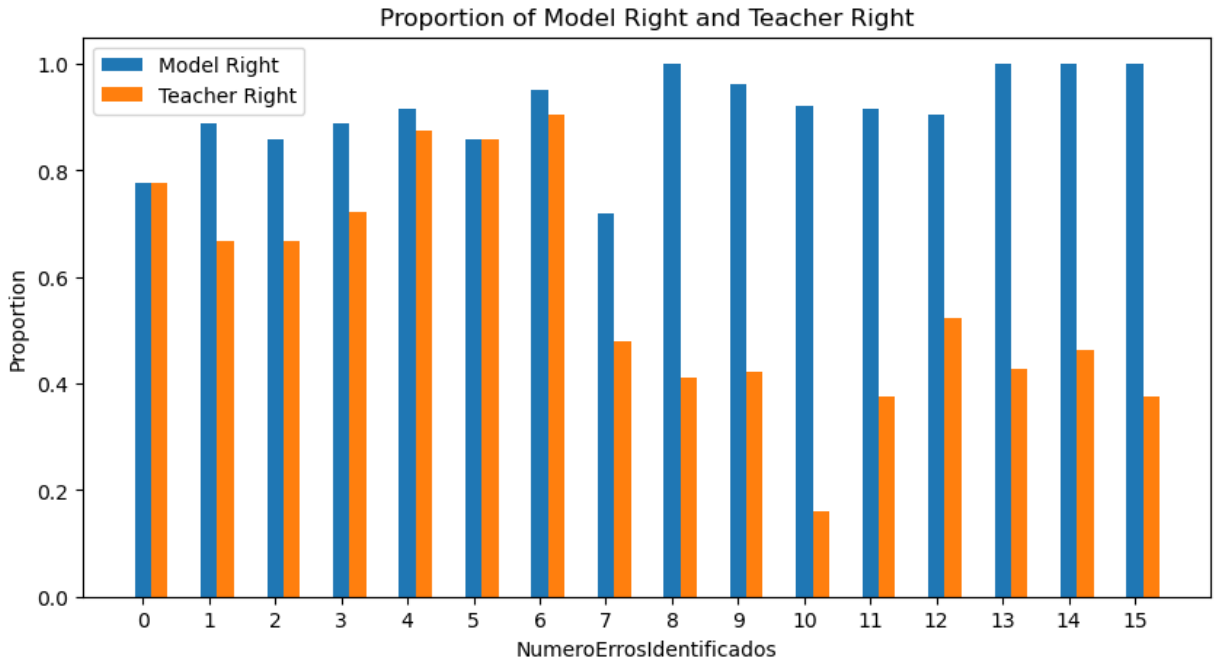

Figure E 20 *– Bar plot of the proportion of students for which the first model (**blue bars**) and the teacher (**orange bars**) accurately predicted their academic achievement (Fail) for each value of **NumeroErrosIdentificados**.*

NumeroErrosIdentificados	model_right	teacher_right	Number of students	Model Right Proportion	Teacher Right Proportion
0	7	7	9	0.777778	0.777778
1	8	6	9	0.888889	0.666667
2	18	14	21	0.857143	0.666667
3	16	13	18	0.888889	0.722222
4	22	21	24	0.916667	0.875000
5	30	30	35	0.857143	0.857143
6	20	19	21	0.952381	0.904762
7	18	12	25	0.720000	0.480000
8	17	7	17	1.000000	0.411765
9	25	11	26	0.961538	0.423077
10	23	4	25	0.920000	0.160000
11	22	9	24	0.916667	0.375000
12	19	11	21	0.904762	0.523810
13	14	6	14	1.000000	0.428571
14	13	6	13	1.000000	0.461538
15	8	3	8	1.000000	0.375000

Table E 9 *– Summary table of the proportion of students for which the first model and the teacher accurately predicted their academic achievement (Fail) for each value of **NumeroErrosIdentificados**.*

Tables E10 to **E12** summarize the values behind the visual representations in **Figures E21** to **E23**. The first column identifies the failing rate for each of the assignment scores features. The second (*unsuccess_percep* proportion) and third columns (model fail prediction proportion) specify the failing proportion predicted by the teacher and the model, respectively. The last column is the count of students whose features assume each value. These results were obtained on the **test set of the first model**.

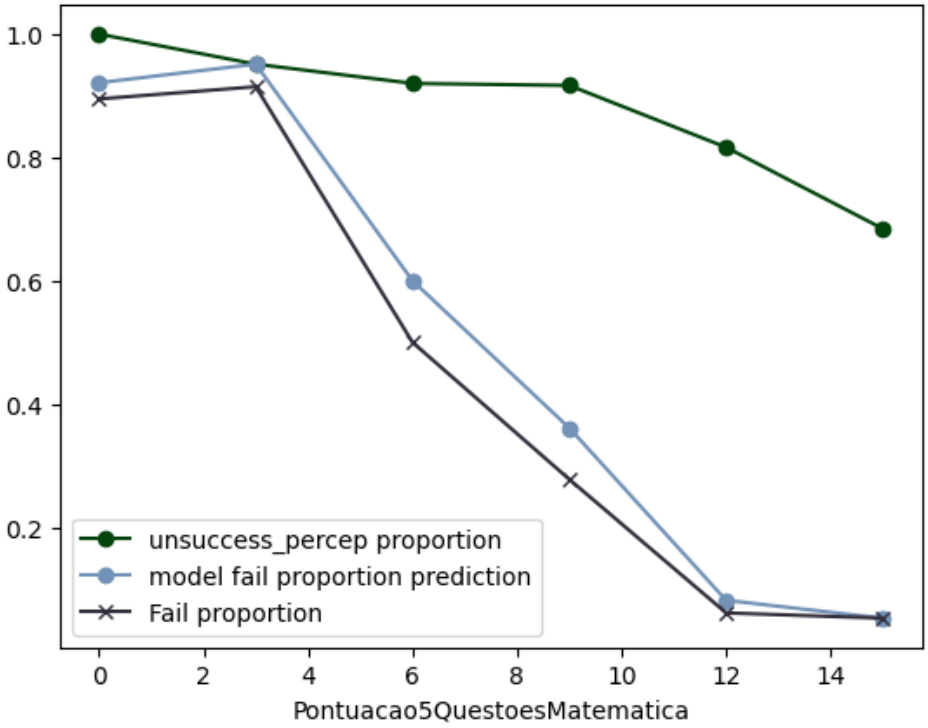

Figure E 21 *– Proportion of students predicted to fail by the model and teachers for each value of* ***PontuacaoQuestoesMatematica****. Actual proportion is also included in the plot (Fail proportion).*

PontuacaoQuestoesMatematica	Fail proportion	unsuccess_percep proportion	model fail proportion prediction	Number of students
0	0.894737	1.000000	0.921053	38
3	0.914634	0.951220	0.951220	82
6	0.500000	0.920000	0.600000	50
9	0.277778	0.916667	0.361111	72
12	0.061224	0.816327	0.081633	49
15	0.052632	0.684211	0.052632	19

Table E 10 *– Summary table of the proportion of students predicted to fail by the teacher and the model for each value of* ***PontuacaoQuestoesMatematica****. For comparison, the actual failing rate is presented in the first column.*

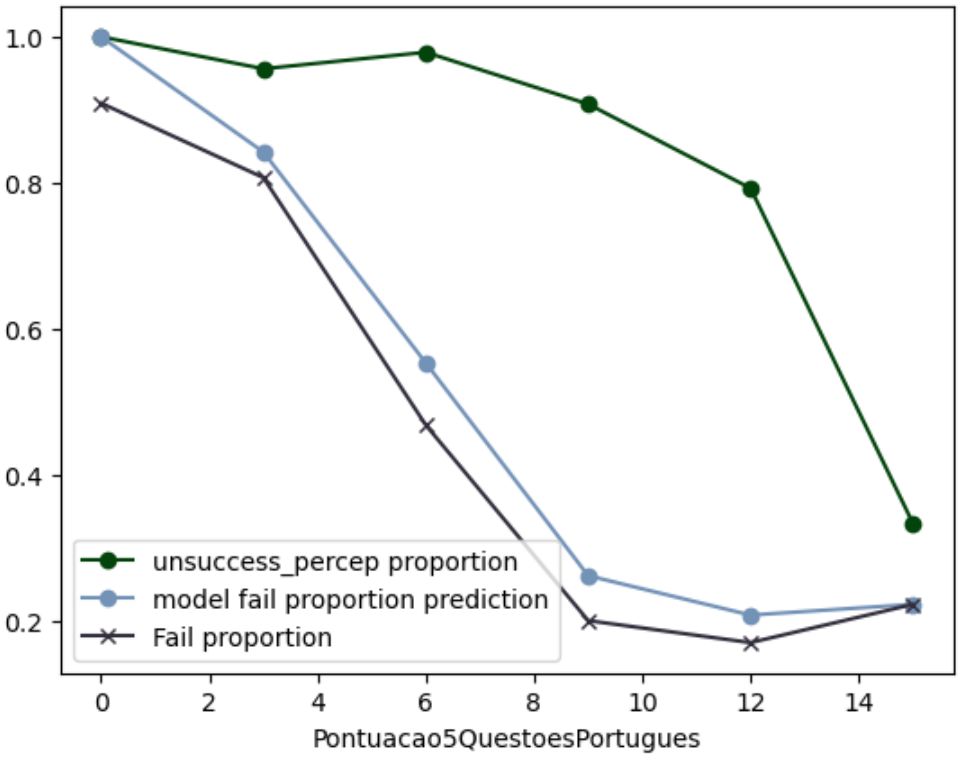

Figure E 22 *– Proportion of students predicted to fail by the model and teachers for each value of* ***Pontuacao5QuestoesPortugues****. Actual proportion is also included in the plot (Fail proportion).*

Pontuacao5QuestoesPortugues	Fail proportion	unsuccess_percep proportion	model fail proportion prediction	Number of students
0	0.909091	1.000000	1.000000	22
3	0.807018	0.956140	0.842105	114
6	0.468085	0.978723	0.553191	47
9	0.200000	0.907692	0.261538	65
12	0.169811	0.792453	0.207547	53
15	0.222222	0.333333	0.222222	9

Table E 11 *– Summary table of the proportion of students predicted to fail by the teacher and the model for each value of* ***Pontuacao5QuestoesPortugues****. For comparison, the actual failing rate is presented in the first column.*

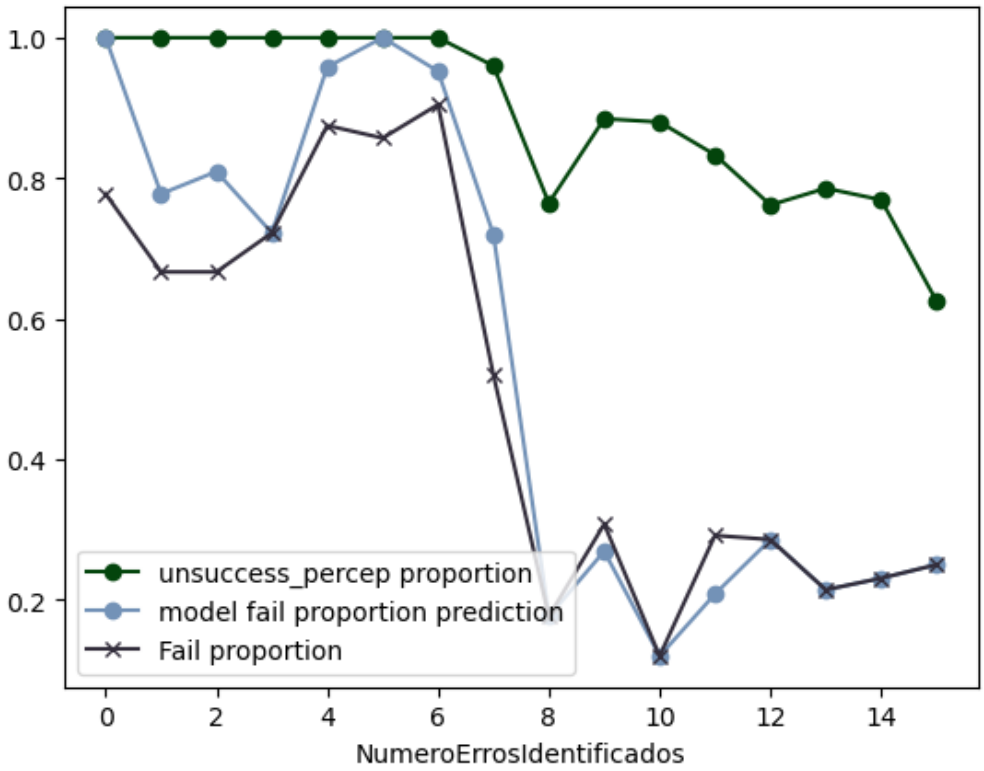

Figure E 23 *– Proportion of students predicted to fail by the model and teachers for each value of* ***NumeroErrosIdentificados****. Actual proportion is also included in the plot (Fail proportion).*

NumeroErrosIdentificados	Fail proportion	unsuccess_percep proportion	model fail proportion prediction	Number of students
0	0.777778	1.000000	1.000000	9
1	0.666667	1.000000	0.777778	9
2	0.666667	1.000000	0.809524	21
3	0.722222	1.000000	0.722222	18
4	0.875000	1.000000	0.958333	24
5	0.857143	1.000000	1.000000	35
6	0.904762	1.000000	0.952381	21
7	0.520000	0.960000	0.720000	25
8	0.176471	0.764706	0.176471	17
9	0.307692	0.884615	0.269231	26
10	0.120000	0.880000	0.120000	25
11	0.291667	0.833333	0.208333	24
12	0.285714	0.761905	0.285714	21
13	0.214286	0.785714	0.214286	14
14	0.230769	0.769231	0.230769	13
15	0.250000	0.625000	0.250000	8

Table E 12 *– Summary table of the proportion of students predicted to fail by the teacher and the model for each value of* ***NumeroErrosIdentificados****. For comparison, the actual failing rate is presented in the first column.*

APPENDIX F

Technical Explanations

The purpose of this section is to offer clear definitions and explanations of various concepts and algorithms used in machine learning. The information presented is based on notes from the machine learning course, and it is intended to provide readers with a concise and informative overview of fundamental principles in this field. The goal is to help readers better understand key machine-learning terminology and methodologies so that they can engage with the subject matter more effectively.

DATA VISUALIZATION

- **Principal component analysis (PCA)**

PCA, which stands for Principal Component Analysis, is a dimensionality reduction technique often used to visualize large data sets by transforming the set of variables into a smaller one that preserves the most information out of the original data set. PCA defines combinations of the original features, usually referred to as **principal components**, in the data space along which the data variance is maximized. The first principal component is the axis that captures the most variance out of the original data set, and each subsequent component captures the maximum remaining variance orthogonal to the previous ones. To visualize the data set in this work, the first two principal components were used.

- **t-SNE (t-distributed Stochastic Neighbor Embedding)**

This technique is a manifold algorithm that, in the context of machine learning, refers to a class of techniques designed to discover and model the underlying structure or geometry of high-dimensional data. The idea behind t-SNE is to find a two-dimensional representation of the data

that **preserves the distances between points as best as possible**. t-SNE starts with a random two-dimensional representation for each data point and then tries to make points that are close in the original feature space closer and points that are far apart in the original feature space farther apart. t-SNE puts more emphasis on points that are close by rather than preserving distances between far-apart points. In other words, it tries to preserve the information indicating which points are neighbors to each other.

MODELS ESTIMATION

General concepts

- **Grid Search Cross Validation**

Grid Search Cross-Validation (Grid Search CV) is a hyperparameter tuning technique widely used in machine learning to systematically search for the **optimal set of hyperparameters** for a model, or even search for the best classifier in a set of different estimators. Grid Search CV performs an exhaustive search over a specified hyperparameter grid, which may also include classifiers, evaluating the model's performance at each combination of hyperparameters using cross-validation. Particularly, when performing five-fold cross-validation, the data is first partitioned into five parts of (approximately) equal size, called **folds**. Next, a sequence of models is trained. The first model is trained using the first fold as the test set, and the remaining folds (2–5) are used as the training set. The model is built using the data in folds 2–5, and then the chosen metric is evaluated on fold 1. And so on until this process is repeated using all individual folds as test sets. For each of these five splits of the data into training and test sets, we compute the chosen metric on the test set. In the end, we have collected five metric values, for which the average is provided as the grid-search best score. The best estimator found (with the best average score) is trained on

the entire training set so that we can use it for prediction and, consequently, to make a final evaluation of the best model on another data set not used during training.

- **Pipeline**

In the context of machine learning, a pipeline refers to a sequence of data processing and modeling steps that are chained together. The purpose of a pipeline is to streamline and automate the process of building and training machine learning models, making it more organized and easily reproducible.

Feature Scaling

When features are on a similar scale, it helps some machine learning algorithms to perform better. For instance, if one feature has a much larger range of values than another feature, the algorithm may place too much emphasis on the larger feature and not enough on the smaller one. For this reason, features should be transformed to the same scale, so machine learning algorithms consider all features on an even playing field. In the context of this work project, two scaling methods were considered:

- **MinMaxScaler**

MinMaxScaler is a data preprocessing technique used in machine learning to scale and transform features to fall within a specified range, specifically between 0 and 1. The formula used to obtain the scaled features is the following:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Equation F 1 *– Formula used to obtain the scaled features with the MinMaxScaler method.*

The minimum and maximum are found based on the training set only, to avoid data leakage.

- **StandardScaler**

StandardScaler ensures that for each feature, the mean is 0 and the variance is 1, bringing all features to the same magnitude. Unlike *MinMaxScaler*, this scaling method does not ensure any specific minimum and maximum values for the features. It is based on the following formula:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Equation F 2 *– Formula used to obtain the scaled features of the StandardScaler method.*

Feature Selection

Usually, adding more features to our data set can make the prediction power of our model better since it has more information to learn the instances from. However, this can also make models more complex**,** increasing the chance of overfitting**.** When adding new features, or with high-dimensional datasets in general, it can be a good idea to reduce the number of features to only the most useful ones and discard the rest**.** This can lead to simpler models that generalize better.

For this purpose, a feature selection method was included in both models built for this work:

- **Recursive Feature Elimination**

This is an iterative method since it starts with all features, builds a model, and discards the least important feature according to the model. Then, a new model is built using all but the discarded feature, and so on, until only a pre-specified number of features are left. For this to work, the model used for selection needs to provide some way to determine feature importance. In both models built for this research, the Decision Tree was chosen as the base model for selection.

One of the hyperparameters of this method is the number of features to select, which was tuned during the Grid-Search CV process in this work.

Data Set Balance

The learning phase of machine learning algorithms and their subsequent prediction can be affected by the problem of an imbalanced data set. The balancing issue corresponds to the difference in the number of samples in the different classes, which, in the case of the model where this step was used, corresponds to an imbalance in the variable *model_best*. Two balancing techniques were considered in this work to overcome this issue:

- **RandomOverSampler**

This balancing technique generates new samples of the under-represented class, also called minority class. Specifically, it generates new instances by randomly sampling with replacement the current available samples.

- **Synthetic Minority Oversampling Technique (SMOTE)**

Another popular method to over-sample the minority class is the Synthetic Minority Oversampling Technique (SMOTE). While *RandomOverSampler* duplicates some of the original samples in the minority class, SMOTE generates new samples by interpolation.

Classifiers

Across this work, different classifiers were tuned, as well as their different hyperparameters. Therefore, the following section briefly describes these classifiers and the respective hyperparameters tuned in the grid-search processes.

- **Logistic Regression**

Logistic regression is a linear classifier that estimates the class probability for each instance, using the log odds of class membership, in our context meaning the logarithm of the ratio of the probability of the instance belonging to class 1 to the probability of belonging to class 0. Based on

this probability and given a threshold of 0.5 (the default value), it also provides a given class for each instance, between *Fail* = 0 or *Fail* = 1 for the first model, and *model_best* = 1 or *model_best* = 0 for the second model.

In the case of Logistic Regression, the tuned hyperparameters were the **penalty** and **regularization terms**, which are used to **control and prevent overfitting**. Specifically, these hyperparameters introduce a cost for having large coefficients in the model, encouraging it to be more robust and less prone to overfitting. The penalty was tuned based on L_1 and L_2 regularizations, which penalize absolute values of the coefficients and the sum of the squares of the coefficients, respectively. The second hyperparameter defines the strength of the regularization. A stronger regularization means less variance (fewer overfitting) in the model. On the other hand, less regularization means more variance (more overfitting) in the model.

The following classifiers are **ensemble methods**. Ensemble methods combine a group of models, often giving better predictions than the best individual model. This is based on the **Wisdom of the Crowd**, which is a theory that suggests a large and diverse group of people can collectively make better decisions and predictions than an individual expert. This is because the group brings in a wider range of perspectives and experiences, leading to a more accurate and reliable outcome.

- **Random Forest**

This ensemble method averages a set of deep **Decision Trees** (classifiers), combining it with bagging (bootstrapping). The first step - drawing bootstrap samples with replacement - consists of training different subsets, the same size as the original one, where each one is obtained by sampling with replacement. This means instances from the original training set are sampled, where ones can be repeated, and others may not even be selected for the sample. For each of the training subsets,

a decision tree is fitted. Here, randomness is also achieved by selecting a random subset of the original features to be considered for the splitting of each tree.

In this case, a different **number of trees** was tuned, as well as the **maximum depth** of the trees being trained, used to control overfitting, and the **minimum number of samples** required to split an internal node in the decision tree.

- **Extreme Gradient Boosting (XGBoost)**

This boosting method trains predictors sequentially, each trying to correct its predecessor. One of the most used methods nowadays is Gradient Boosting, which works by sequentially adding predictors to an ensemble, each one correcting its predecessor. This method tries to fit the new predictor to the residual errors made by the previous predictor. XGBoost is an extension of Gradient Boosting, designed to be highly efficient and flexible, and uses decision trees as base learners. Compared to the basic gradient boost method, XGBoost adds advanced regularization (L1&L2), which improves model generalization. The learning process in XGBoost involves minimizing an objective function that consists of two parts: a loss term, which measures the difference between predicted and actual values, and a regularization term, which penalizes overly complex models. Once again, the **maximum depth** of the trees was tuned, as well as the **learning rate** that determines the step size at each iteration while moving toward a minimum of the loss function. This latter hyperparameter plays a crucial role in controlling the contribution of each tree to the final ensemble in the boosting process.

MODEL INTERPRETABILITY

Interpretability methods in machine learning refer to techniques that help to understand, explain, and make sense of the predictions or decisions made by a model. Interpretability is particularly useful if the model one has at hand is a black-box model, meaning the internal workings of the model are complex and not easily understandable. For the development of this work project, two techniques were used:

- **SHAP (Shapley Additive exPlanations)**

Shapley values are a robust mechanism to assess the importance of each feature to reach an output, for each data point.

We can visualize feature attributions such as Shapley values as **"forces"**. Each feature value is a force that either increases or decreases the prediction for that instance. The prediction starts from the baseline, which is the average of all predictions, which in the context of classification is a value (probability) between 0 and 1. Each Shapley value, for each feature, is “a force” that pushes to increase (positive value) or decrease (negative value) the prediction. These forces balance each other out at the actual prediction of the data instance.

For the present analysis, **the average of the absolute SHAP values** for all the data points, mean(|SHAP|), is considered to conduct this method.

- **PDP (Partial Dependence Plots)**

Partial Dependence Plots (PDPs) are a useful tool for understanding the relationship between a specific feature and the predicted outcome of a machine-learning model. PDPs are univariate methods since they visually represent the marginal effect of a feature while keeping all other features constant. This marginal effect is the computed average outcome across all data points with each value of the feature.

CLUSTERING

Clustering techniques in machine learning are unsupervised learning methods that aim to group similar data points together based on certain features. The goal of clustering is to identify natural patterns or structures within a dataset without prior knowledge of the class labels.

Clustering Methods

- **K-Means Clustering**

This is an iterative algorithm that partitions a dataset into K distinct groups (clusters), where K is previously defined. The algorithm's primary objective is to **minimize the sum of squared distances** between data points and the centroids of their respective clusters. The centroid represents the center of mass of the data points in the cluster.

It starts with an **initialization step**, where the centroids are defined. This is followed by a **clustering assignment step**, where each data point is assigned to the nearest cluster, such that the sum of squared Euclidian distance is minimized. Then, in **the update centroid step**, it recalculates the centroid of each cluster by computing the mean of all data points assigned to that cluster. This process is repeated until convergence is reached. Convergence occurs when the centroids no longer change significantly or when a specified number of iterations is reached.

Choosing K, the number of clusters

Three different techniques were used to determine the appropriate number of clusters for the data sets explored throughout this work. These are often used as complementary evaluation techniques rather than one being preferred over the other.

- **The Elbow Method or Sum of Squared Euclidean distances (SSE)**

The aim of the elbow method is not to minimize the inertia, also known as SSE, but rather to look for the "elbow" close to the minimum of the inertia function. This latter is minimized at the limit considering each data point as a cluster; however, there should be a trade-off between error and the number of clusters, so one should look for the “elbow” where the inertia starts to decrease significantly slower.

- **Silhouette Coefficient**

The Silhouette coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on two factors: 1) How close the data point is to other points in the cluster; and 2) How far away the data point is from points in other clusters.

Silhouette coefficient values range between -1 and 1. Larger numbers indicate that samples are closer to their clusters than they are to other clusters, reason why one should try to **maximize** this coefficient.

- **Davies-Bouldin Index**

This metric measures the compactness and separation between clusters in a clustering result. The Davies-Bouldin Index computes the ratio of the average within-cluster distance to the distance between clusters for each pair of clusters. The index then takes the maximum of these ratios across all pairs of clusters and averages them over all clusters. A lower Davies-Bouldin index relates to a model with better cluster separation, so this index should be **minimized**.

STATISTICAL TESTS

Across this work, two tests were used to test the significant difference in the distribution of the features for two different groups. As a first instance, the tests were applied to the teachers' data set to check the significant difference between the distribution of the features across the instances for which *benefit_or_not* = 1 and *benefit_or_not* = 0. In a second approach, the tests were performed for students' features to check the difference in its distributions across the data points for which *model_best* = 1 and *model_best* = 0.

- **Chi-Square test for Independence**

The chi-square test for independence is a statistical test used to determine whether there is a significant association between two discrete variables as long as they assume a finite number of values. It assesses whether the observed frequency distribution of the variables, obtained based on the respective confusion matrices, differs from what would be expected under the assumption of independence. It is based on the following hypotheses:

Null Hypothesis (H_0): There is no association between the two discrete variables; they are independent. In the context of this work, it would mean that the variable, at least by itself, may not be useful to help predict *benefit_or_not* / *model_best*.

Alternative Hypothesis (H_1): There is a significant association between the two categorical variables; they are not independent, which means the feature has the potential to be used to define *benefit_or_not / model_best*.

To perform this test, some assumptions need to be verified, namely that features are finite discrete, as referred before, and that the observations are independent.

- **Kolmogorov-Smirnov (KS) Test**

The Kolmogorov-Smirnov (KS) test is a non-parametric statistical test used to assess whether a sample follows a specified distribution or whether two samples are drawn from the same distribution. In this work, the second approach is being used, with the two samples translating to the two groups we aim to distinguish (*benefit_or_not* = 1 and *benefit_or_not* = 0), where each feature values are obtained from. The test is used to compare the distribution of the features across the two different groups, meaning the goal is to assess whether the distributions of the features differ significantly between the two. It is based on the following hypothesis:

Null Hypothesis (H_0): There are no statistically significant differences between the feature distributions in the two groups.

Alternative Hypothesis (H_1): There are statistically significant differences between the feature distributions in the two groups.

Similar to the Chi-Square test, the KS test also requires the validity of some assumptions, namely that the observations in the same group are independent and identically distributed (i.i.d.).

APPENDIX G

Statistical Formulas

Throughout this work, various metrics are referred to in the context of the performance evaluation of different models. This appendix provides the statistical formulas and formal definitions of these metrics.

Note that in this appendix, these metrics will be referenced relative to the performance of a machine-learning model. However, these definitions are still valid if we consider teachers' predictions.

To define these metrics, four concepts must be also considered:

- **True Positives (TP):** Instances that are actually positive (from the positive class) and are correctly predicted as positive by the model.
- **True Negatives (TN):** Instances that are actually negative (from the negative class) and are correctly predicted as negative by the model.
- **False Positives (FP):** Instances that are actually negative but are incorrectly predicted as positive by the model.
- **False Negatives (FN):** Instances that are actually positive but are incorrectly predicted as negative by the model.

Accuracy

Accuracy is a metric used in classification problems to measure the overall correctness of a model's predictions across all classes. It provides a general assessment of how well a model performs by comparing the total number of correctly predicted instances to the total number of instances in the dataset.

Accuracy is also often used (informally) to mean any general measure of a classifier's performance.

$$Accuracy = \frac{Number\ of\ correct\ decisions\ made}{Total\ number\ of\ decisions\ made}$$

Equation G 1 *– Accuracy statistical formula.*

Precision

Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. In other words, precision measures the accuracy of the positive predictions, indicating how many of the instances predicted as positive actually belong to the positive class.

$$Precision = \frac{True\ Positives}{Predicted\ Positives} = \frac{TP}{TP + FP}$$

Equation G 2 *– Precision statistical formula.*

Recall / True Positive Rate (TPR)

Recall, also known as sensitivity or true positive rate, is a performance metric used to measure the ability of a model to identify all relevant instances of the positive class correctly. It focuses on the fraction of actual positive instances that the model correctly predicts as positive.

$$Recall = \frac{True\ Positives}{Real\ Positives} = \frac{TP}{TP + FN}$$

Equation G 3 *– Recall statistical formula.*

F1- score

The F1-score is a metric used in classification problems to balance the trade-off between precision and recall. It provides a single score that combines both precision and recall into a single value. It is the harmonic mean of precision and recall, and it is calculated using the following formula:

$$f1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Equation G 4 *– F1-score statistical formula.*

Threshold choice

Changing the threshold value affects the values of these metrics in a distinct matter, which is why a threshold tunning step was included for the first model.

- Lower threshold → higher number of predicted positives → higher recall and lower precision.
- Higher threshold → lower number of predicted positives → lower recall and higher precision.