Predicting Class Drop-Outs in Secondary School Mathematics

John Baierl and Elayne Stecher
UCLA - STAT 201B

I. Introduction

Numerous studies have explored the links between socioeconomic factors and academic performance in secondary education across a variety of outcomes of interest, such as college admission tests and classroom outcomes in core subjects. Recent work has found that performance in high school is explicitly linked to future income and occupational success and is an even better predictor of future success than other factors, like cognitive ability (Spengler et al. 2018). These and similar findings imply that dropping out of or failing a secondary school course may have large ramifications for students' future wellbeing, which raises the question: How can we identify, predict, and ideally, intervene in time to help struggling students?

In their 2008 paper, *Using Data Mining to Predict Secondary School Student Performance*, Cortez and Silva apply several machine learning techniques to predict student performance on their year-end exams across two Portuguese secondary schools. In particular, they utilize random forest classification, support vector machines, naive Bayes and neural networks to predict two different outcomes: the students' continuous score on their final exams and a binary classification of whether students scored above a threshold. The primary findings from their preferred model, the random forest, suggest that students were more likely to score below the threshold if they had previously failed a class or were absent from class; they were more likely to score above the threshold if they went out on the weekends or if their parents had higher education levels. Across all of their models, test scores in the previous period were the strongest predictors of student performance on the third exam.

While their methods demonstrated good predictive power and produced relatively intuitive findings, their choices of models lacked interpretability, preventing us from making substantive inferences about the causes of students' poor academic performance. Moreover, the authors primarily focused on predicting whether student scores fell at or above a score of 10 out of 20 (the overall median score was 11). The selected threshold is not terribly informative from a policy-making perspective, as it captures too wide a variety of students. In the math class data, approximately one-third of students scored below the 10-point threshold (130 out of 395 students) and 38 students scored zeros on the third exam. Nearly 10 percent of students dropping out of a core academic course is striking. Given the high potential for substantial academic disruption resulting from a single course drop-out, this warrants treatment as a distinct outcome from simply below-average test performance.

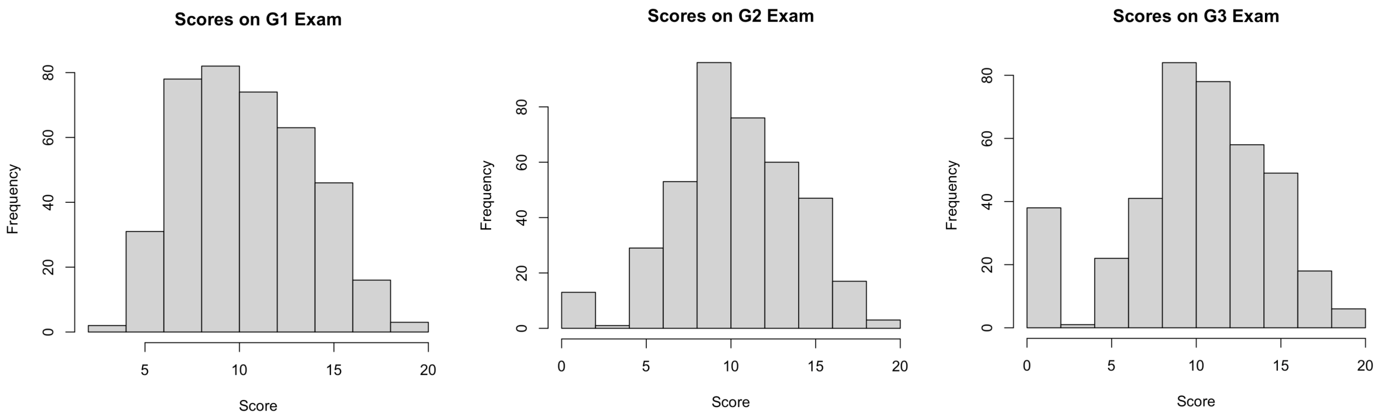
In addition, in their analysis, Cortez and Silva include the students' test grades from the previous two periods, which predictably end up being the primary predictors of students' performance on the third paper. While this may be informative in some contexts, it also prevents the authors from identifying other socioeconomic or structural factors that could predict student drop-out *before* their exam scores in previous periods start to suffer.

Thus, in this paper, we re-analyze the 2008 data with a shifted emphasis on identifying those factors in students' school and home lives that indicate a high risk for dropping out of a class. We anticipate that redefining the outcome variable will be more informative from a policymaking or interventionist perspective and would provide school administrators and teachers with the ability to predict which students are at high risk of academic struggle early in the school year, enabling proactive measures to be taken. We also explore whether or not the factors that predict drop-out rates in mathematics class are similarly predictive of dropping out of a Portuguese language class, to see whether or not the same factors are significant and have some intuition about how generalizable our findings are.

II. Data and Methods

Our data were collected during the 2005-06 academic year from two Portuguese public schools. Results from three different exams taken throughout the year were recorded in both mathematics ($N = 395$) and Portuguese language ($N = 1044$) courses. As previously discussed, prior analysis of this data focused on predicting whether a student's performance on their final test (G3) fell at or above a threshold of roughly 10 out of 20 possible points (Cortez and Silva 2008). The median of all test takers was 11 (out of 20 points), so this threshold roughly translates to students scoring the 33rd percentile or below. The authors thus classify students as low-scoring or high-scoring and do not offer an explanation for the high number of students who score 0 on the third exam. It is worth noting that all students who scored a 0 on the second exam also scored a 0 on the third exam, which has led us to the conclusion that a score of 0 represents a student dropping out from the course. This is further bolstered by the fact that no students received a score of 0 on the first exam, suggesting that such a score is distinct from other, below-average grades.

Fig. 1: Distribution of the first (G1), second (G2), and third (G3) test scores



Given this, our primary outcome of interest is predicting whether or not students receive a score of 0 on the third exam or not, which we interpret as dropping out of the class or not.

There are a large number of covariates included in the data, aside from the scores on the other exams, which are shared across both the mathematics and Portuguese classes data sets. The data contains student characteristics such as gender and age, as well as home factors such as family size (*famsize*; a binary measure of whether the family consisted of greater than three individuals). Details about their parents were also surveyed, including whether or not their parents lived together (*Pstatus*), their parents' education level (*Medu* and *Fedu*), and job (*Mjob* and *Fjob*). Students provided their reason for choosing that school (*reason*: close to "home", school "reputation", "course" preference or "other"). Further details about the students' academic and extracurricular lives were also recorded such as whether they are in extracurricular activities, whether they intend to go to college, if they have a romantic partner, if they have internet access at home, how frequently they go out and use alcohol, their current health, and their number of absences. Most of these student-specific variables were binary, but their levels of going out, alcohol use, and health were scored on a five-point Likert scale, and the number of failures was a continuous variable.

a) Feature Selection and Engineering

In order to detect the impact of socioeconomic factors on students' performance that could predict drop-out perhaps even before students' grades dropped, we exclude the first- and second exam scores from our control variables. This further differentiates our analysis from that of Cortez and Silva, which finds these variables -- *G1* and *G2* respectively in the dataset -- to be most highly predictive of the third exam score (*G3*), effectively "soaking up" the predictive power of their models. We also removed two factor variables, *Mjob* and *Fjob*, which are mother's and father's jobs, respectively.

Aside from this, we do not engage in any other type of feature selection or dimension reduction. While we initially considered various methods for doing so, we ultimately decided not to for several reasons: 1) multicollinearity is not a major concern among these covariates (Appendix *Table 1*); 2) among those variables with moderate degrees of correlation, we lacked theoretical rationale to adjudicate between which variables should be included and which should be dropped; and 3) for several of our models, irrelevant variables would effectively be downweighted without the need to perform variable selection a priori.

As mentioned, our primary outcome is a binary measure as to whether students scored a 0 on the third exam (dropped out) or scored anything else, which got them assigned a value of 1 (did not drop out).

b) Model Selection

Given that our primary task was classification of binary outcomes, we selected four models to perform this classification task:

1) A generalized linear model with a logistic link function (logistic regression);
2) LASSO logistic regression;
3) Kernel-based regularized least squares regression (KRLS); and
4) Random forest.

We selected these models in particular for a few reasons. Logistic regression was chosen as the standard workhorse model that could handle a binary classification problem and provide reasonably interpretable results. We chose LASSO and KRLS to help us downweight variables of lesser importance and to see if this penalization gave us meaningfully different results than the logistic regression. The KRLS model also hedges against potential model misspecification in the assumption-heavy logistic model. Finally, we included random forest to give us more comparable results to the original paper (although our threshold for the outcome variable was, as stated, different from Cortez and Silva's).

III. Results

We first ran our models on a 90/10 training and testing split. We used 10-fold cross validation to determine our models' performance. We initially used ROC curves to determine the threshold that would optimize the tradeoff between predicting true positives and false negatives for each fold; the results of this analysis are in Column 1 of Table 1. Plots of the threshold-based classification and overall classification results from the logistic regression can be found in Appendix Figures 2 and 3. We also present the average AUC for the out-of-sample data in Column 2 for each model. For both measures of model performance, we see that logistic regression and random forest are our two favored models.

Table 1: Model Performance (out-of-sample)

	Column 1	Column 2
Model	**Prediction accuracy (threshold cutoff)**	**AUC (average)**
Logistic regression	91.0%	0.965
LASSO	85.8%	0.907
KRLS	90.0%	0.794
Random Forest	91.4%	0.939

IV. Analysis

Both the logistic regression and random forest performed quite well within our classification scheme, correctly identifying over 90% percent of students meeting our cutoff criterion in out-of-sample testing. Due to its interpretability over the random forest model, we present the results of the logistic regression model over the random forest, particularly since it outperformed random forest in average AUC.

Our logistic regression identified two significant covariates at the 0.05 significance level: previous class failures ($p = 0.026$) and family size ($p = 0.0126$). Additionally, three covariates were weakly significant at the 0.1 level: mother's education ($p = 0.077$), weekend alcohol consumption ($p = 0.059$), and health ($p = 0.059$). 95% confidence intervals for the first differences of these five factors are reported in the table below.

Table 2: First differences of significant factors

Covariate:	**First difference in probability of remaining in class:** 95% CI
(*) Previous Class Failures (0 to 3)	(-0.629, -0.226)
(*) Family Size (≤ 3 or > 3)	(-0.391, 0.037)
(.) Mother's Education (< MS: 0 to Higher ed: 4)	(-0.048, 0.412)
(.) Weekend Alcohol (Least: 1 to Most: 5)	(0.052, 0.413)
(.) Health (Worst: 1 to Best: 5)	(-0.204, 0.158)

Note: (*) denotes $p < 0.05$, (.) denotes $p < 0.1$

Previous class failures represent a significant increase in the risk of dropping out of a math class, with three previous failures increasing dropout probability by between 23 and 63 percentage points (95% confidence). This aligns with Cortez and Silva's (2008) results, identifying class failures as the strongest predictor in their binary classification. Home factors such as family size and mother's education level also show a high likelihood of increasing dropout risk, with children of larger families being more likely to drop out. While zero is included in the 95% confidence intervals of first differences for both of these factors, it is only on the very fringes, suggesting some effect is likely to exist and that further study with a larger sample might be warranted. Perhaps most surprising is the effect of weekend alcohol consumption, with increased consumption corresponding with *decreased* probability of dropping out of the class (between 5 and 41 percentage point decrease).

Despite preferring the logistic model for binary classification due to the slight improvement in performance, we will analyze our random forest results to provide a direct comparison to Cortez and Silva's results. The authors provided relative importance values given in mean decrease accuracy. These are shown in the table below along with our results.

Table 3: Comparison of mean decrease in accuracy from Random Forest models

	Relative Importance: Mean Decrease Accuracy				
Cortez, Silva:	Failures: 21.8%	Absence: 9.4%	School support: 7.0%	Go out: 6.5%	Higher ed: 6.4%
Our results:	Absences: 27.4%	Failures: 12.4%	Age: 6.31%	Higher ed: 5.09%	Go out: 4.89%

Again, we see alignment in the importance of class failures in both of our binary classification schemes. The continuing prevalence of social factors in both sets of results is also noteworthy here, with frequency of going out with friends providing predictive power in both sets of results. Comparing this with the significance of weekend alcohol consumption from our logistic model results suggests that social connectedness plays a role in predicting class failure, and that further research into this factor is likely warranted.

a) *Generalizability*

To what extent are our models only useful for predicting drop-outs from the mathematics class? We took the logistic model from our mathematics course data and used it to predict which students would drop out from their Portuguese classes. In contrast to the math students, only 15 out of 635 students dropped out of the Portuguese class. As in math class, no students received a 0 on the first exam, and all students who received a 0 on the second exam also received a 0 on the third exam, seemingly validating our assumption that a 0 score represents a drop-out.

We find that when using our math class data as the training data and the Portuguese class data as the testing data, our model performs with 78.0% accuracy. We used the ROC curve method again to determine the threshold for our cutoff, which was 0.867 here. A plot of the results of the classification test can be seen in Appendix *Figure 4*. *Table 4* below shows the confusion matrix to assess model performance. While we can see that our model correctly predicts that 13 out of the 15 students who do drop out would do so, we also see that it incorrectly predicts that 141 additional students will drop out who did not. Thus, we see that this model has very good sensitivity at the cost of specificity and overpredicts the occurrence of dropping out from the Portuguese class. Due to the lower prevalence of dropouts in our

Portuguese data set, more data is needed to draw substantive conclusions about the differences in dropout risk factors across subjects.

Table 4: Predictions for student drop-out from Portuguese class

	Predicted		
	Drop Out	**No Drop Out**	
Actual Drop Out	13 (TP)	2 (FN)	15
Actual No Drop Out	141 (FP)	493 (TN)	634
	154	495	

IV. Conclusions and Discussion

We find that both social factors and home conditions play major roles in predicting dropout rates from mathematics class among secondary students. While some effects were not in the direction we anticipated, such as an increase in weekend alcohol use being associated with a lower probability of dropping out, other factors, like previous class failures, absences, and health, had an impact in the direction we would expect.

Our best-performing model from the math data overpredicts the dropout rate among students in the Portuguese class. While this may not be a bad thing from a policymaking perspective – we would rather offer support to too many students, rather than let a large number of at-risk students fall through the cracks – this indicates that math courses may a present unique risk for drop-out as compared to other subjects. Schools with limited resources would be better served treating these subjects independently rather than relying on a single model across both math and language.

Perhaps the most surprising result here is the relationship between weekend alcohol consumption and drop-out risk, with students consuming more alcohol being less likely to drop-out of their math classes. We suspect that this may be a proxy for the level of social-connectedness of that student. Given the limited available data on students' social lives here, future work should seek to explore the ties between social well-being and likelihood of drop-out. This also suggests that schools ought to consider students' social lives when assessing potential academic support needs.

Future data collection exploring the specific reasons why students dropped each course would add more depth to this analysis. Perhaps drop-outs resulting from overcommitment, lack of family support, or health challenges have distinct causes, warranting separate treatment and analysis. Information on family income was perhaps the most notable absence from the available data. While the surveys provided to students requested their level of income, the vast majority of families did not respond, and it was omitted from the published data and subsequent analysis. Given the extensive literature linking overall academic performance with family wealth, future data collection should seek to connect income and wealth with drop-outs. This would allow to control for the possibility that income is confounding several of the covariates considered here like academic support and student health, providing a clearer picture of any causal relationships between these factors.

References:

P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance*. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Marion Spengler et al. *How You Behave in School Predicts Life Success Above and Beyond Family Background, Broad Traits, and Cognitive Ability*. Journal of Personality and Social Psychology, 2018 DOI: 10.1037/pspp0000185

Appendix

Figure 1: Correlation plot of numerical variables

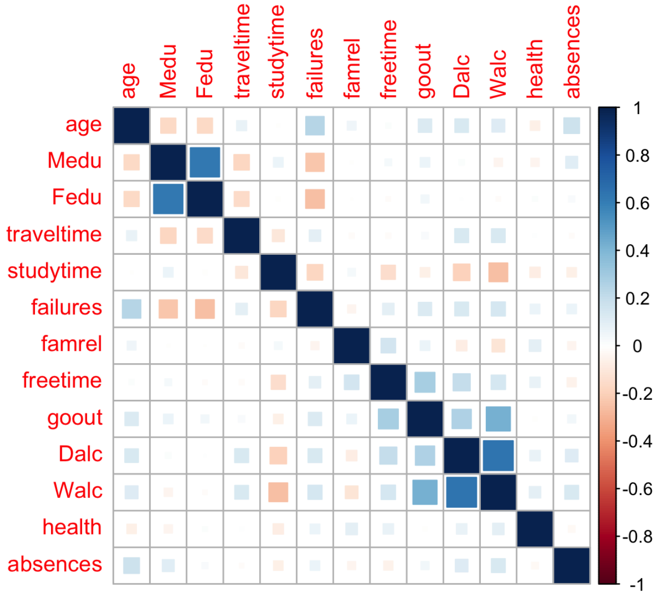


Figure 2: Example of true values (green) vs. predicted values (red) vs. threshold (dotted line)

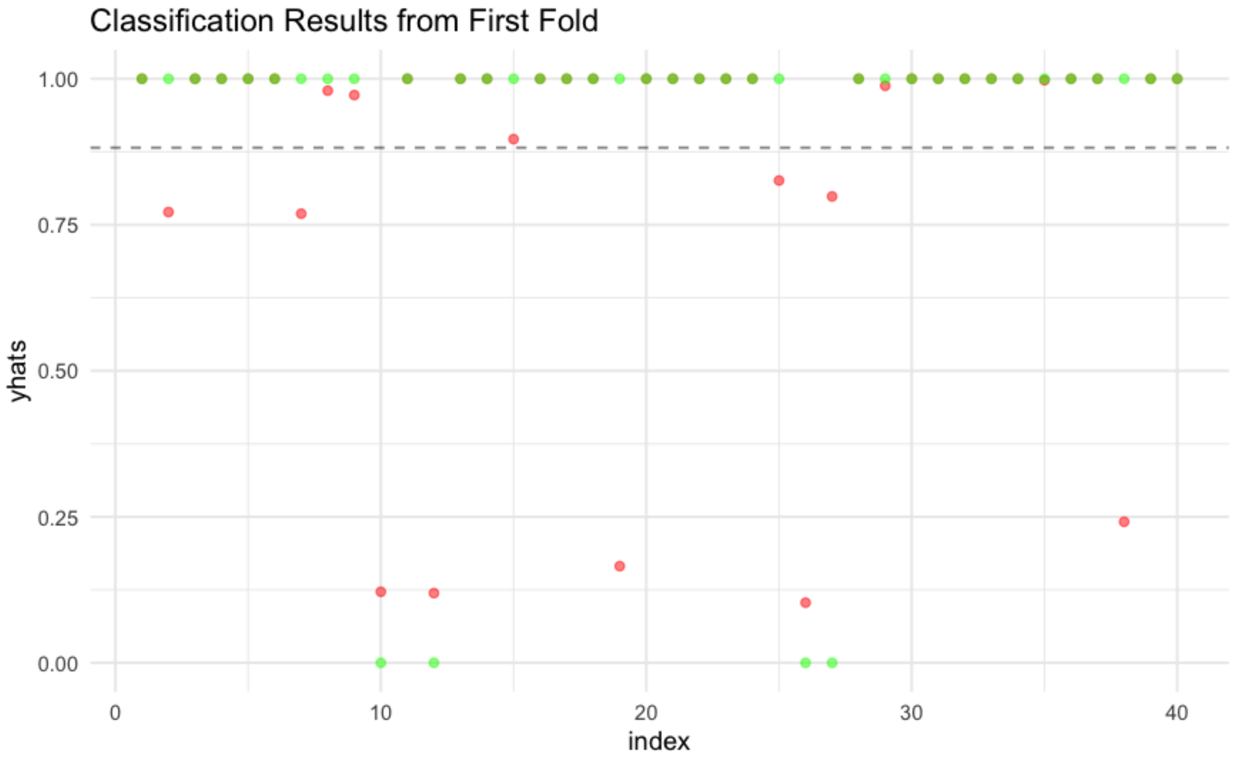


Figure 3: Overall classification of the logistic regression model

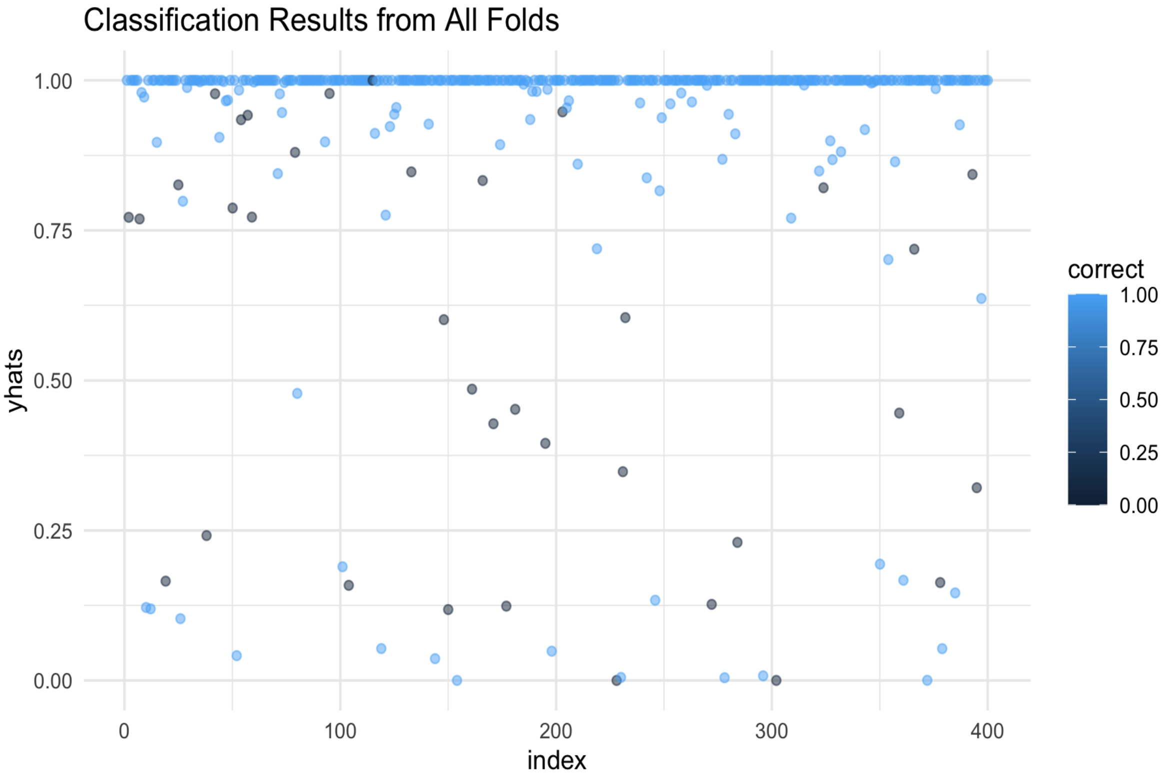
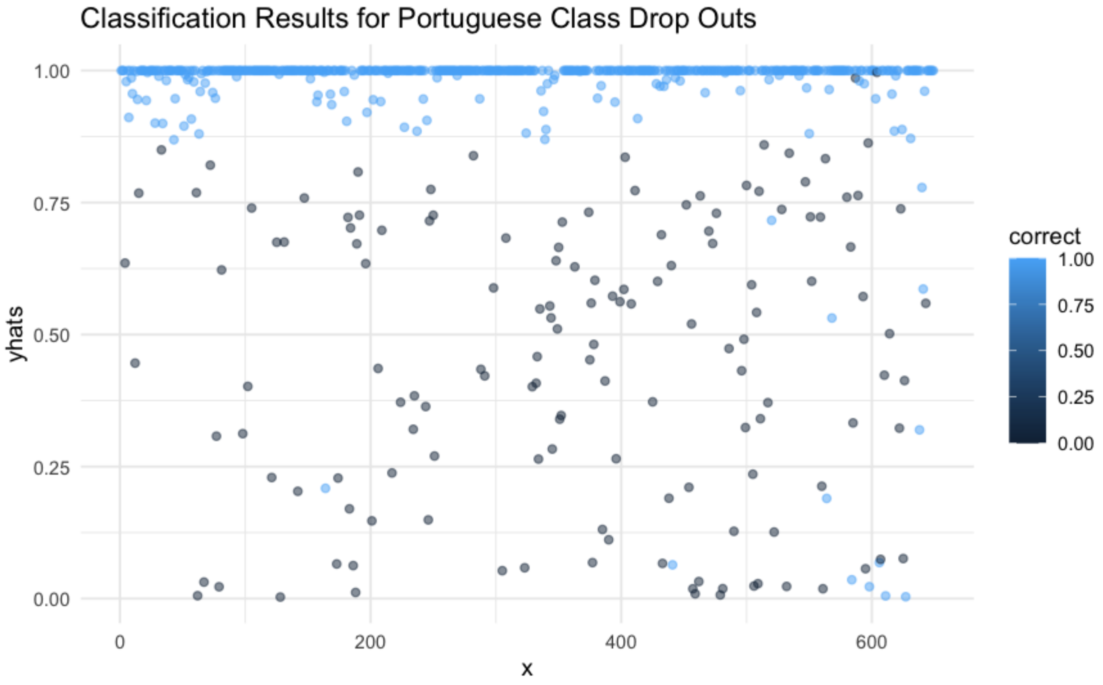


Figure 4: Overall classification of the Portuguese students dropouts


Classification Results for Portuguese Class Drop Outs
yhats
x
correct