Can the scholastic and demographic situation of a student, as well as alcohol consumption, be useful in predicting his/her grades?

Demurtas Federica, Gadda Alberto, Galimberti Dario

CdLM Data Science, Università degli studi di Milano Bicocca

1. Introduction

Is it true that consuming more alcohol is a symptom of obtaining low grades? Similarly, do specific social information about the family or a high consumption of alcohol decrease grades in a student career? The dataset 'Student alcohol consumption', along with the analyses done, aim at answering to these questions. Starting from the datasets, these were chosen on Kaggle website [1] and they are two: one including Math's course data and one Portuguese's course data, both recruited on two different secondary schools. They have respectively 395 and 649 observations, on 33 common variables. The columns include some information on the student, such as age, sex, address, in which of the two schools they go, how much they study, if they do any extra-curricular activities, how much they go out with friends, health status, absences, the consumption of alcohol during the week and during the weekend (the variable is numeric from 0- very low, to 5- very high). In addition, there are some variables explaining the family composition and size, their education level and the job and if parents paid some additional support for the child. As last, the 3 most important variables show the grades of students from 0 to 20 for the first period, the second period and the final period grade. The approach chosen, after some statistical analysis, is to classify students who passed or not the final period exam, based on other characteristics. The goal starts by dividing between training and test set, and then continues by using an inducer, create an algorithm to understand how to classify. This can be done with some methods, then decide which one is the best by evaluating the performances. Other approaches to this problem could be creating a linear regression or studying deeply correlations, but they could be less efficient.

2. EDA

2.1. Choosing dataset

The first matter to deal with is the duality of the datasets: as said, one contains math's course students, while the other Portuguese course students; the peculiarity is that 382 students are in both datasets, meaning that almost all students in math's dataset are also in Portuguese's, but not vice versa. This arises the first problem: how to handle the datasets. Three solutions can be found: 1. The first one includes keeping separated the datasets, in order to develop the same operations on both datasets, and then we can also examine more deeply by creating a third dataset through the inner join of the two. This solution has that advantages that differences between humanistic and scientific subjects can be found, on the other hand 3 smaller datasets are of study. 2. A merge of the two datasets can be done at the beginning of the analysis, but a problem of null values originates: in fact, students who are in both datasets will have the grades' variables both for math's course and for Portuguese course. The benefit is that one big dataset is created. 3. The last option is to choose one of the two datasets and very likely it would be the Portuguese one, as it has highest numerosity and almost all students in math's one are in Portuguese, but still some data would be lost.

The option chosen for the study is to work with the Portuguese dataset only, so the last option, in order to get the analysis easier and clearer. The decision has been made as a consequence of the small numerosity of the Math's dataset; in fact, if we had used also the Math's dataset, the implementation of a model would have been complex and less efficient.

2.2. Correlations

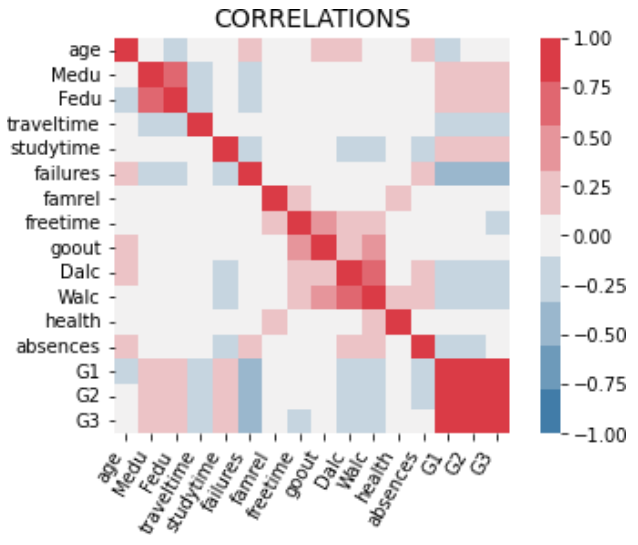

Fig. 1. Correlations plot

The second step is the analysis of correlations; the following graphs show the links between variables, but no strong evi-

dence seems to be present, except for the grades, that are obviously highly linked.
Regarding the distributions of grades, the final period grades seem to have more zeros than the other periods grades, this can be since the final exam is the most discriminant between diligent and less hard-working students. It is also possible that they did not reach enough to take the last exam, automatically taking G3 = 0.

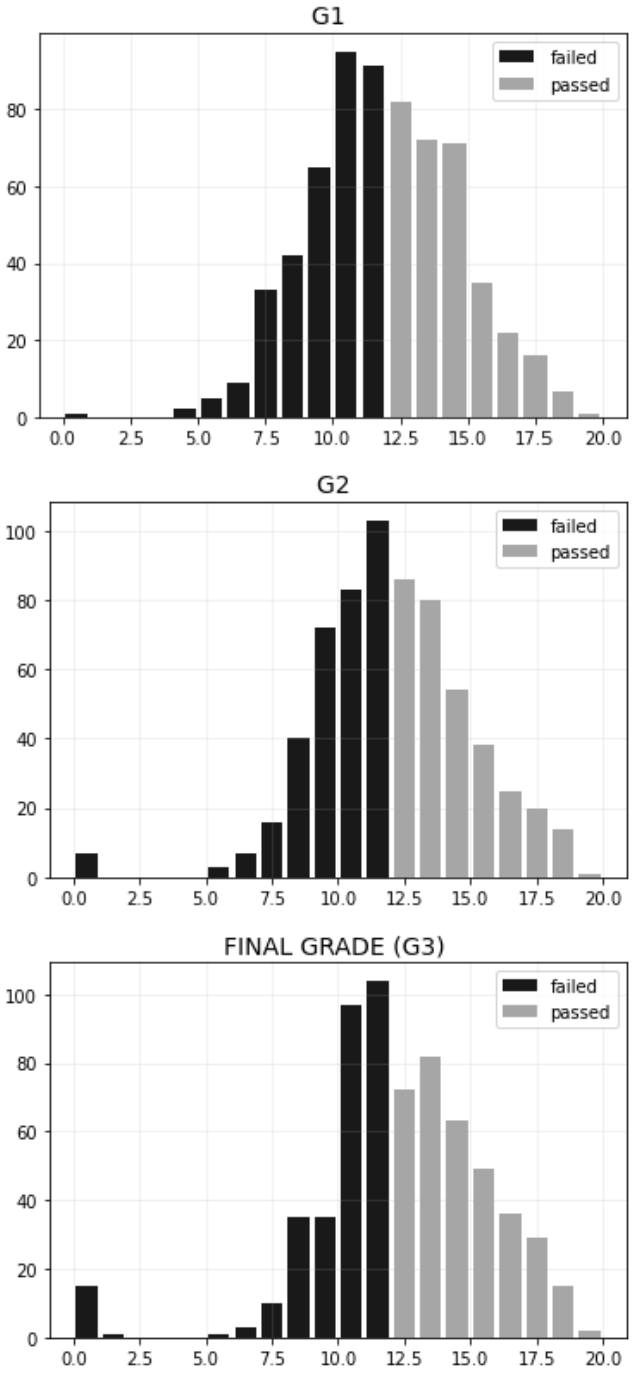

Fig. 2. Grades distribution

As it can be seen by the histograms, the values of the class variable (G3) seem to be balanced. With no further analysis, there is no need to make them perfectly balanced or of same number.

By exploring the correlation between the level of parents' education and final period grade, a positive link can be found; in fact, if the former increases, also the latter increases, even if no strong correlation is present. In this way, the variable is of interest in the analysis.

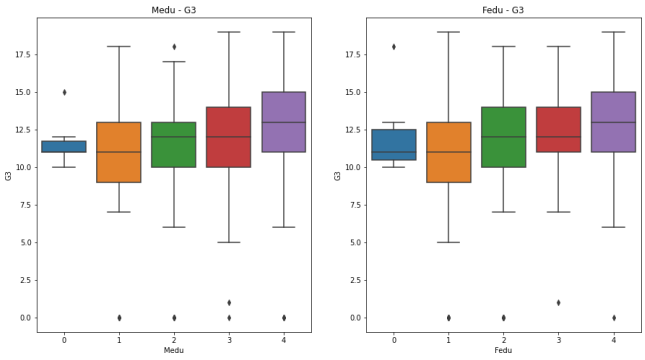

Fig. 3. Parental education

Then the same analysis is done with alcohol consumption. In this case, the grades decrease as the consumption of alcohol increases, this can be clearly seen in the graph, as the median of grades tend to be lower and lower. The tendency of negative link is more visible in weekdays, as in these days students go to school. In addition, drinking alcohol during weekdays, even more if in great quantity, can be evidence of serious problems with respect to students who consume it only during weekends.

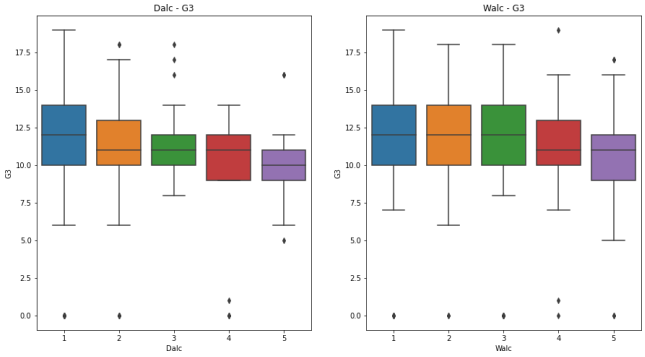

Fig. 4. Alcohol consumption

The same technique is applied to binary variables of the dataset. The variable 'higher', explaining the willingness to continue studies, has an evident correlation with the finale period grade. If a student likes studying, he will be more prone to continue studying, aiming also at higher grades.

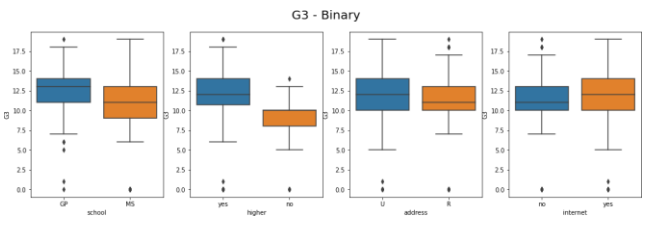

Fig. 5. Binary variables

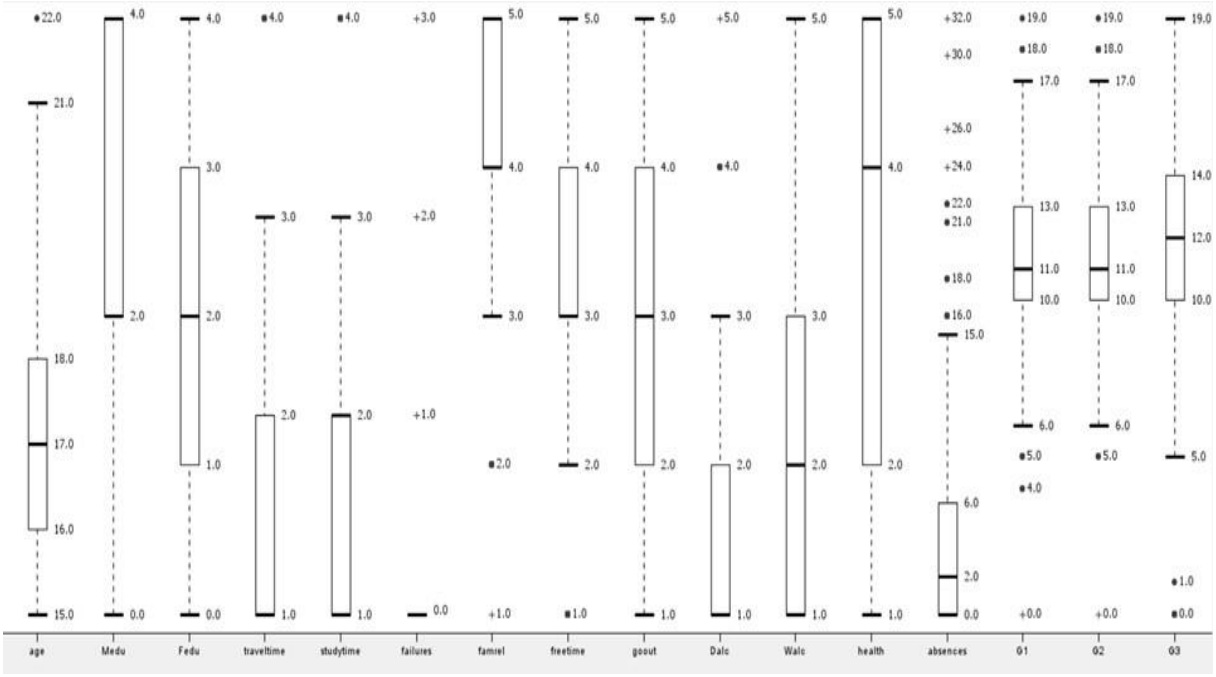

Fig. 6. Boxplot

2.3. Missing values and Outliers

It is of interest to say that no missing values are present in the dataset, so there is no need to choose how to handle them.
As first sight, it is possible to see that, especially for absences but also for some others, there are outliers. We decided to leave them where they are, as in this case they are extremely informative; if one student has more days of absences than the range, does it reduce or increase the possibility of failing the exam? So, outliers like 26,30 or 32 are extremely important in the analysis.

3. Preprocessing operations

3.1. K-Folds cross validation

K-folds cross validation is a resampling procedure to classify with higher accuracy. Basically, the dataset is divided in k groups (exhaustive and mutually exclusive), 1 group is used for test set and the remaining k-1 are used as training. Each time the test and training set change. For every couple training-test, we fit a classification model on the training set, evaluate it on test set and then retain the evaluation score. At the end, we will take the mean of all k evaluation scores; this ensures a more effective estimate of the classifier's accuracy.

3.2. Division in Train and Test set

As first step, before starting the classification algorithms implementation, it is necessary to divide the dataset into trainSet1 and testSet1, and we decided to divide in 70-30. The division is stratified with respect to variable G3, in order to have the same distribution in the two partitions. The testSet1 won't be used until the very end of the analysis, and it is needed to verify performance obtained with the final model.

The trainSet1 is again divided into trainSet2 and testSet2, so that the model can be trained with a portion of trainSet1 and results can be tested on the remaining part (testSet2). Once this operation is completed for all models chosen for the analysis, the best algorithm will be chosen, by considering various metrics. The best model then will be evaluated also on the testSet1, in order not to be influenced by a previous knowledge of data. In fact, feature selection is done only on trainSet1; so, before inducing the model on testSet1, we will filter manually the columns returned by the feature selection and only those will be used to learn the algorithm.

3.3. Feature transformation and normalization

3.3. Feature transformation and normalization It is decided to transform the variable G3 (grade of final period exam) into a binary attribute 'passed' or 'failed'. Then, the variables G1 and G2 are dropped, as the final exam is the one that decides if a student passes the year or not. In addition, all explanatory variables of the dataset are converted to numerical variables, in order to being able to use every algorithm possible. In fact, some algorithms in knime do not support categorical inputs. We choose also to normalize the dataset, as the range of values that variables assume are different with each other. Without this operation, variables with higher values (like 'age') would have had a higher weight in the model, and that would have generated influenced results. The normalization chosen scaled all variables in a range between 0 and 1.

3.4. Feature Selection

Feature selection is an important step in the preprocessing analysis, as it allows to understand which variables are significantly increasing or decreasing the alcohol consumption.

Eliminating non-significant variables means having a less costly inducer to be trained and no useless or misleading variables involved in the classification. Through the meta node 'Forward Feature Selection' in Knime, it is possible to choose to perform forward or backward elimination; we need to specify which columns should be static and which can change and choose between the 2 strategies and at the end of the loop, there will be the possibility to choose among a certain number of possible sets of variables with different accuracies. Forward selection found 8 variables to be significantly change the class attribute, with an accuracy of 0.8. Noteworthy is the fact that alcohol consumption, as well as number of absences and the will to continue studies, result being the most useful variables in predicting the pass-fail class.

3.5. Variable Importance

A further study of variable importance can be variable importance. This highlights the most important variables with respect to each other. An example is given by the random forest implementation of this technique, and it's explained below. With the Random Forest model it is possible to know the contribute of each variable to the development of the model by the statistics provided by the random forest learner. The importance of a variable is determined by the times a variable is selected as a splitting node divided by the times it's candidate for the split for the first three levels of the tree; these values are evaluated on the training set. As we can see the most used variable is "higher" (51%) followed by "Dalc"(40,2%),"absences"(40,1%) and "Fedu"(38,2%). Unfortunately, on Knime it's not been possible to obtain these statistics on the other models, but we can confirm that the alcohol consumption is important to determine if a student pass or not the exam as well his/her will to continue the study.

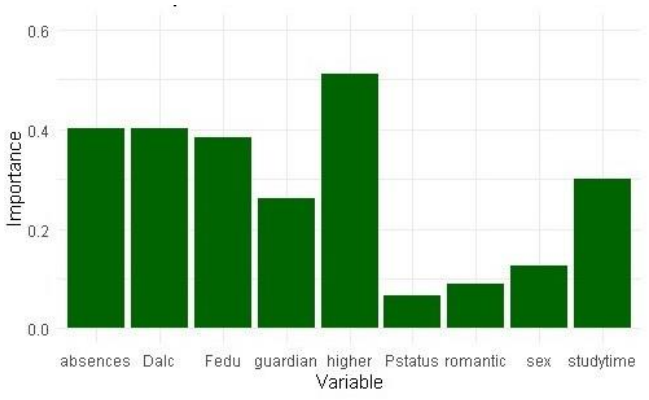

Fig. 7. Variable importance according to the random forest

4. Models used

4.1. Decision Tree

Decision tree is a supervised machine learning algorithm. It is tree-structured: the internal node represents the feature, the branches represents the splitting rule and the leaf nodes are the outcomes. A decision tree asks if a condition is satisfied or not and then splits the tree into two sub-trees; the best attribute for the node can be chosen in two ways: the information gain and the Gini index, they are based on impurity measures and the algorithm selects the variable and the condition that minimizes the impurity for the 2 subtrees. This algorithm has the advantage that it's easily interpretable and requires less data cleaning compared to other methods; however, it contains lot of layers that make it complex and it may have overfitting issues.

4.2. Naïve Bayes

Naïve Bayes classification is part of probabilistic methods. This method is based on Bayes theorem, with an important assumption of independence among variables that, in this dataset, is completely satisfied. Naïve Bayes classification starts with the computation of posterior probabilities of the class attribute given the significant explanatory variables. Then, the class with highest posterior probability is the outcome of the prediction. [2]

4.3. Logistic Regression

The third method used in the workflow is the logistic classification, that is a regression-based approach. The idea beyond logistic regression is to find the best fitting model to describe if a student passes of fails the final exam. The Logistic regression classification uses the Logistic function to determine the probability to pass or fail the exam, given the independent variables. The difference with Naïve bayes lies in the model used: while in Naïve Bayes a posterior probability is calculated, in Logistic regression, a parametric conditional probability is used to classify. [3]

4.4. Support Vector Machines (SVM)

Support vector machine is a separation approach. It searches for the hyperplane that maximizes the margin distance between points. Suppose to classify data in many dimensions, there could be many hyperplanes that equally divide well points, but by SVM chooses the one that maximizes the margin. Then, points that lies in on the right or on the left are classified as pass or fail.

4.5. Random Forest

The random forest is an ensemble learning method that unifies more decision trees with the aim of stabilizing the prevision error and reducing overfitting. The main difference between decision tree and random forest is that the second one employs the bagging method, so the samples are different for each decision tree; in addition, in the random forest the split variables are selected randomly and we obtain different outputs by each tree. In case of classification, for every observation the final prediction will be the most common prevision made by the trees of the forest. This model is very flexible and adaptable to datasets with different variables

and/or missing values, nevertheless it's less interpretable than other classification models.

5. Results

After the implementation of various models, the best one must be chosen. To validate the models, we first need to remark the measures used to evaluate a classification model. Firstly, accuracy can be studied; this is the ratio between number of correct classified elements on number of total assessments. The boxplots show the results for the 5 folds, while the line next to it is the value for the validation test.

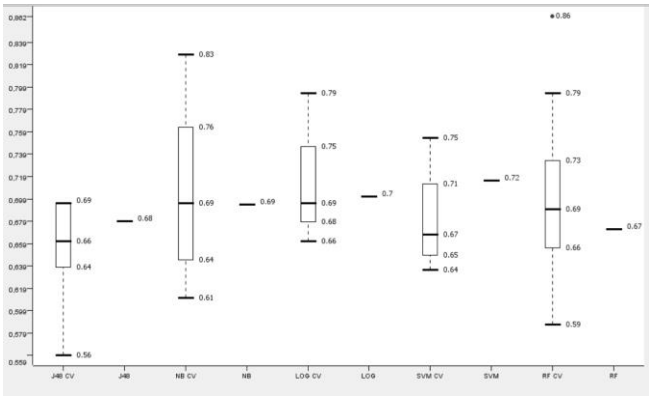

Fig. 8. Model accuracy

The best choice is the one which doesn't overfit the model, doesn't overspecialize on the training set but still reaches high accuracies both in train and validation sets. Support vector machine and logistic regression are the best in achieving these goals. It is important to say that also Naïve Bayes worked well on classification, but it has two main problems: the first is that it has a higher range between the cross validations, so it is less confident, the second is that it has less reliability on validation set, even if of only 0.01 points, with respect to Logistic and SVM.

Other important measures have been used to evaluate the results and to choose the best model, these are: recall, precision, sensitivity, specificity and F-measure, that is the harmonic mean of recall and precision. The graph below shows the metrics mentioned above, respectively for class 1 and class 0 predictions. The best model is the one that has highest values for all; this is not always feasible in real-world data, so the best option is the one that has on average the highest number of peaks in the graphs. Support vector machine and logistic regression are the best models in predicting both class 0 and 1. Nevertheless, the are some differences in the two classes in terms of values: higher F-measures are achieved when predicting class 1 than in class 0 (for class 1 0.7 on average, for class 0 0.62).

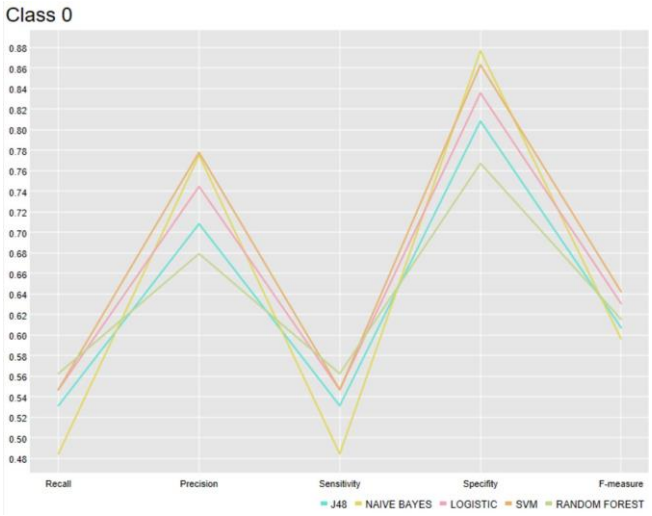

Fig. 9. Class 0 Metrics

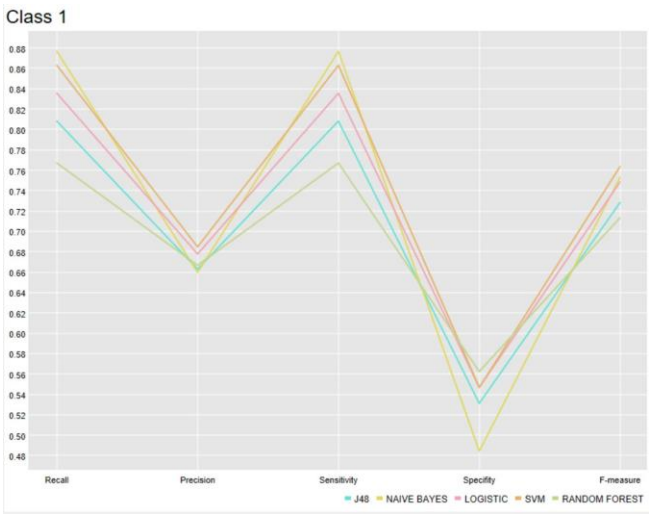

Fig. 10. Class 1 Metrics

As last operation, we can see results of our best model (SVM) on the test set, in order to understand if the work done is fine. The accuracy on the test set is of 68% so, even with a little of overfitting, the model is successful. Below, the confusion matrix relative to the test set is shown.

		PREDICTED	
		0	1
TRUE LABEL	0	52	38
	1	25	80

Fig. 11. Test Set confusion Matrix

6. Rejection concepts

Results of this classification are surely positive, in fact the classifier has learnt how to classify observations with a reasonably good certainty; nevertheless, there are observations where it is more sure and some others where it has more doubts. In these cases, a methodology where not all observations are classified can be used. It is possible, for example, to adopt a technique only for observations where the classifier has a higher security in saying whether the student will pass or not the final exam. In our case this technique can't be adopted using SVM classifier, being that it doesn't say the confidence for each observation. We decided to use the Logistic model (that is still the second best), using as threshold of confidence in predicting 65%. In doing this and seeing results in the train set, accuracy goes from 70% to 75%, even if 35% of observations are not classified. In the test set, the improvement is even more remarkable, going from an accuracy of 66,7% to 75%.

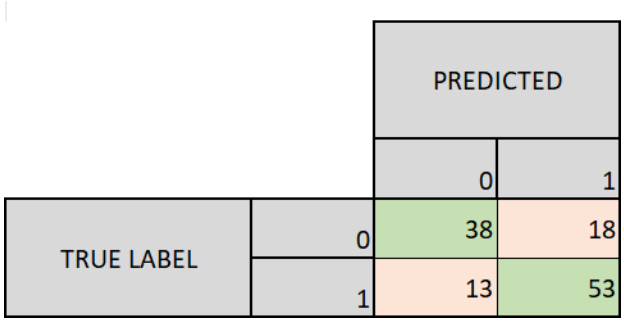

Fig. 12. Test Set confusion matrix with Rejection concepts (Logistic)

7. Future Developements

This study chose to work only with the dataset regarding Portuguese dataset. A possible future development can be a study to highlight the differences between this dataset and the Math's one, that in this study was not analyzed. It would be interesting to see whether variables change the classification pass-fail, varying from a humanistic to a scientific subject. Another implementation would be to collect a higher number of data, also from other schools, in order to obtain a dataset that has more variety and that is numerically bigger. If further studies could do that, the predictive model could be more useful as support in the decisions of teachers. For example, a student that has high probabilities of failing, could have focused assistance in order to prevent this event; this can be done through remedial courses. A limit of this study can be the fact that data of this type are extremely sensible, and so it could be disrespectful asking this information to students.

REFERENCES

[1] : https://www.kaggle.com/uciml/student-alcohol-consumption?select=student-por.csv
[2] : https://kambria.io/blog/logistic-regression-for-machine-learning/
[3] : https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/