CITY
UNIVERSITY OF LONDON
EST 1894

Student Performance Evaluation using Naïve Bayes and Random Forests

INM431 Machine Learning coursework – Daniel Sikar - daniel.sikar@city.ac.uk

Description and motivation

Studies in Europe and the USA have shown the fiscal and social costs of early school leaving, which typically reduces lifetime earnings and leads to higher unemployment incidence and duration, while the benefits of education include gain in net earnings and wealth, improved health status / life expectancy, lower reliance on government health and welfare programmes and reduced expenditures on criminal justice [1][2][3].
One important problem is to automatically detect students that are going to do poorly in a course early enough to be able to take remedial actions [4]. A number of Machine Learning algorithms have been used in student grade prediction. We base our study on prior work [5], using Naïve Bayes and Random Forests applied to secondary school student performance data.

Initial analysis of the data set including basic statistics

The data sets being investigated, made publicly available at the UCI Machine Learning Repository[6], were collected by using school reports and questionnaires, at two Portuguese secondary schools and refers to subjects of Mathematics (student-mat.csv) and Portuguese (student-por.csv). Both data sets contain the same column headers and were combined for the purpose of our analysis (student-labelled.csv).

The combined data set has 1044 observations and 33 attributes, consisting of 4 nominal, 13 binary and 16 numeric attributes, three numeric attributes being grades (G1, G2 and G3) following a 20-point grading scale. Attribute G3 contains the final grade. A "Result" binary attribute column was added with a value of 0 for fails (G3 less than 10) and a value of 1 for passes (G3 greater or equal to 10). Based on the "Result" attribute, the dataset was found to be unbalanced, with 78% passes and 22% fails, which shall be considered when evaluating predictive model performance.

In Fig. 1, the scatter plots show that there is strong correlation between the prior (G1, G2) and final (G3) grades and also between prior grades. The plots also show that there are outliers: data points where the final grade is zero. The data source does not explain why these grades are zero - if we assume the most likely explanation is that the student did not take the exam we could consider eliminating these from the analysis or perform analyses with and without these points.

The histograms imply that Final Grades are higher for students with internet access, no romantic relationship and extra-curricular activities. Statistical means are higher for these subsets.

The box plots suggest that students with some free time (freetime) and moderate amounts of alcohol intake during weekdays (Dalc) and weekends (Walc) have higher mean grades. According to the data source[5], attribute "Health" is described as "current health status (numeric: from 1 – very bad to 5 – very good)". Given the "health" by final grade "G3" box plot shows that students in poorest health have achieved higher mean grades, the described gradient is assumed to be inverted.

Additional data analysis box plots (not shown in Fig. 1), plotting “quality of family relationships (numeric: from 1 – very bad to 5 – excellent) versus final grade, show students with better family relationships have higher mean grades, the same being the case for students with parents with higher levels of education (numeric attributes Medu and Fedu).

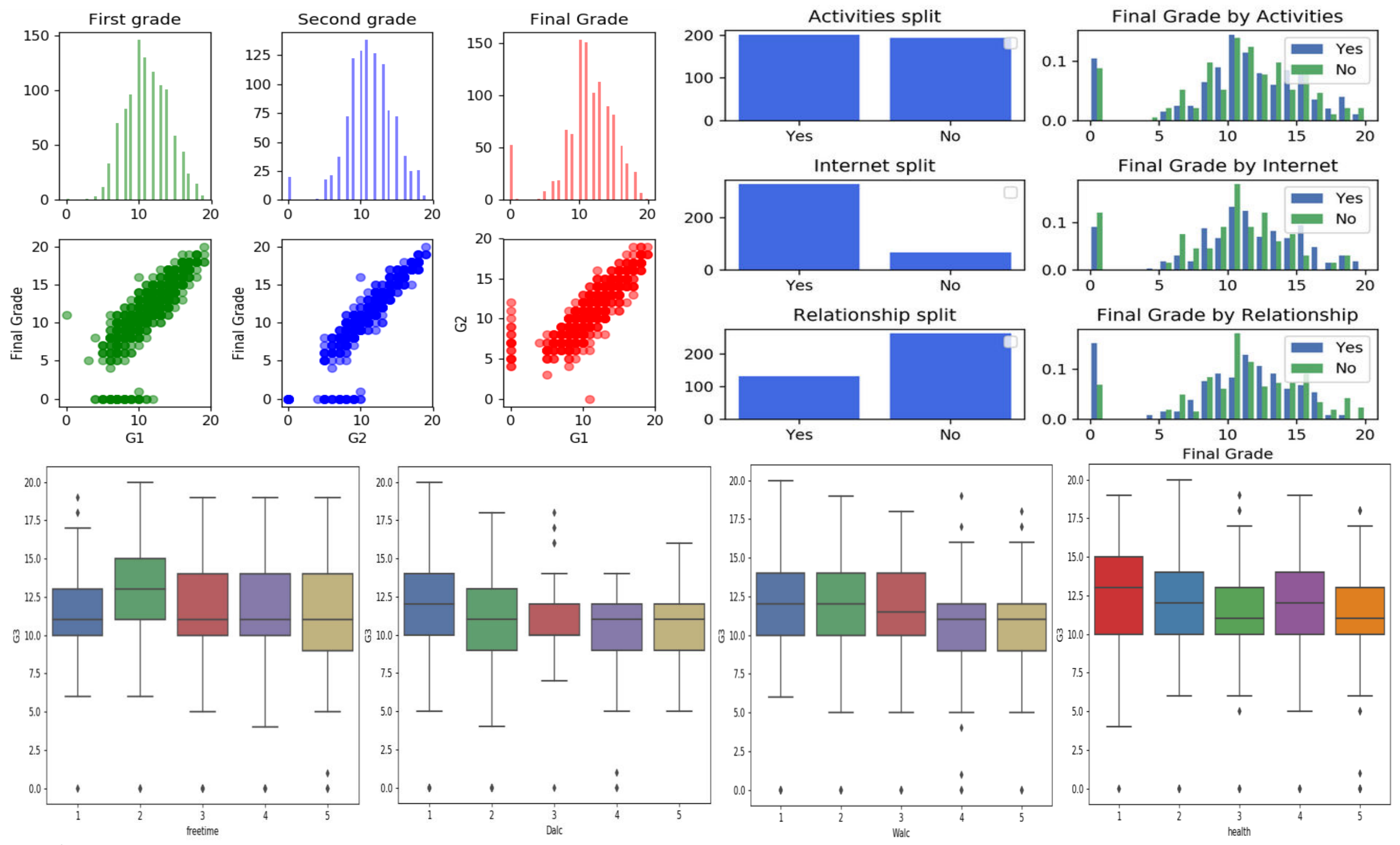

Fig. 1

Two models with their pros and cons

Naïve Bayes

- Simple technique for constructing classifiers, which for can be trained efficiently for supervised learning tasks.
- Uses a decision rule such as MAP (maximum a posteriori) to select the most probable hypothesis.
- Successful applications in text retrieval
- Assumption of independence eliminates the necessity for intractable probability tables’ computation

Pros

- Computationally lightweight - when compared to other classifiers such as Random Forests, Naïve Bayes classifiers are simple to train and simple to run
- Interpretability – models are readily explainable through the assumption of independence, with attribute probabilities accounting for outcome

Cons

- Assumption of independence – attributes are assumed to have no correlation. In the initial analysis, this has shown not to be the case. Intermediate grades G1 and G2 were found to have a strong positive correlation between each other and also to be strongly correlated to final grade G3
- Some cases, specially with smaller datasets, may required smoothing[9] of data, eliminating zero probabilities which may otherwise skew results

Random Forests

- First described by Breiman et at [8] and implemented in R package RandomForest.
- Popular method for different machine learning tasks. Robust to the inclusion of irrelevant features.
- Expands on the concept of searching over a random subset of the available decisions when splitting a node, in the context of growing a single tree.

Pros

- Decision trees that grown very deep then to overfit, by learning irregular patterns (noise). Random Forests offer a way to average out multiple decision trees, trained on different parts of the same training set [7]. Increase final performance
- Models can be trained in parallel which facilitates distributing computation.
- Less variance than single decision trees.

Cons

- The trade-off of better performing models is the loss of interpretability. As ensemble models are inherently less interpretable than an individual decision tree.
- Computationally costly, training large ensembles is more time consuming and has higher demands on memory

Hypothesis statement

- We aim to train and run predictive models based on prior work[5], showing that good predictive accuracy can be achieved with machine learning models, provided that the first (G1) and/or second (G2) school period grades are available.
- We expect the trained models not to be as accurate when G1 and/or G2 attributes are not used.
- If our models prove accurate, similar models could be used in providing timely support for students more likely to fail.
- We expect Random Forest to outperform Naïve Bayes models based on previous studies[8] and expect Naïve Bayes models be quicker and more computationally efficient to train compared to Random Forests.
- Some overfitting is expected to be observed due to the small size of the original dataset, given it will decrease even further for the purpose of training once subsets have been removed for testing and validation.
- Some models are expected to present worse performance than randomly choosing an output, given the unbalanced data and the majority class (pass) percentage (78%).

Description of choice of training and evaluation methodology

- Dataset was randomly split into a training set (90%) and a test set (10%).
- Attributes were omitted for some models, to replicate findings in original paper [5].
- Different hyperparameters where tried during a grid search.
- Training evaluation was performed using Kfold, Holdout and n-fold cross validation.
- Test data was classified by best model, then training accuracy was compared to testing accuracy to gauge model optimality and to account for overfitting
- Percentage of Correct Classifications (PCC) was used. A high PCC (i.e. near 100\%) suggests a good classifier[5]. Additional rates of true positives and negatives, and false positives and negatives, we also used to evaluate models as well as to check if models were performing well just by predicting majority classes.

Choice of parameters and experimental results

Naïve Bayes

- Three scenarios were chosen; including attributes G1 and G2, including only G1, and excluding both. As hyperparameters, we used *PredScheme* (Table 2), two data type schemes with the original data types (1), and mostly categorical (2) where a number of categorical attributes with numeric data types were implicitly declared as a categorical. Categorical attributes were assumed to have a multivariate multinomial distribution, and numeric attributes a normal distribution. *Validation* was also used as a hyperparameter with relevant *Param* number of folds or holdout percentage.

Results:

- Presence or absence of attributes G1 and G2 influence training and testing accuracies (trainAcc, testAcc, trainAcc_G1, testAcc_G1, trainAcc_G1,_G2, testAcc_G1_G2).
- Better performance achieved by using original data types
- Much faster to train to Random Forests
- Models being trained with less data (larger holdout values) were most prone to overfitting – increased difference between training and testing accuracies
- Better results obtained by using attributes not explicitly converted to categorical data types
- Test results better than training results for best model, that included attributes G1 and G2

Random Forests

A search for best hyperparameters was attempted:

- Using adaptive logistic regression and bootstrap aggregation methods in turn
- Varying number of objective function evaluations, which as far as we could establish, changed the size of the forest, each decision tree being a function
- Adding G1 and G2 attributes
- Changing number of learning cycles, learning rate and maximum number of splits

Results:

- When previous grades G1 and G2 (Table 1 High Correlation Attributes) were not used, there was a drop in performance with increased differences in training and testing accuracies, suggesting the model was overfitting.
- The best model was also found to be the optimal, with the smallest proportional difference between training and test data set accuracies.
- Testing accuracies were consistently higher than training accuracies for best overall models.
- Very slow to train, especially with a high number (e.g. 100) of objective function evaluations
- Adaptive logistic regression (LogitBoost) found to be best ensemble aggregation method, over bootstrap aggregation[8]

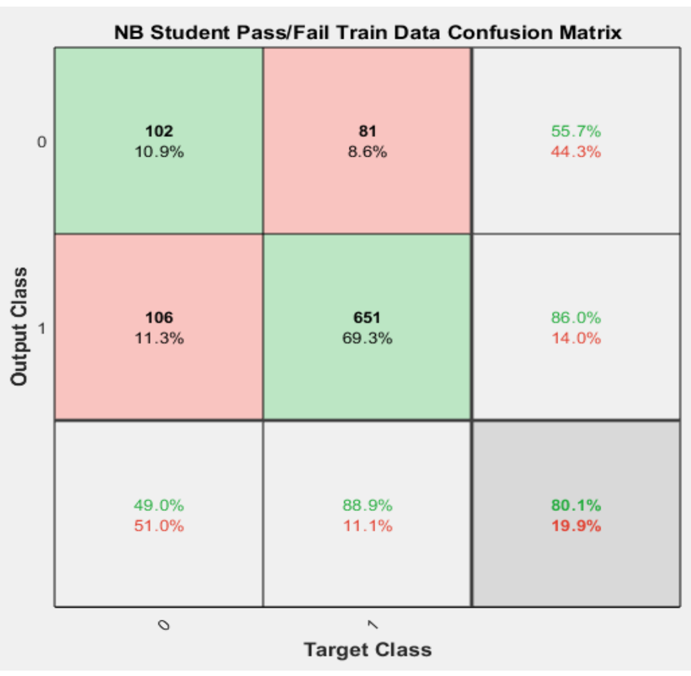

Fig. 2

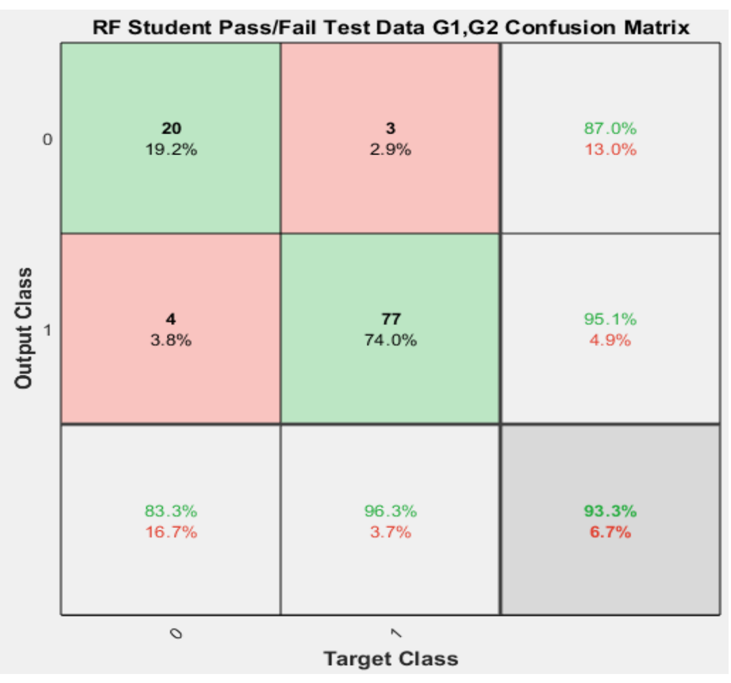

Fig. 3

Model	High Correlation Attributes	Date set	True Pos.	True Neg.	False Pos.	False Neg.	PCC (Accuracy)
Random Forests	G1, G2	Training	706	158	28	48	91.90%
Random Forests	G1, G2	Test	77	20	3	4	93.30%
Random Forests	G1	Training	686	149	47	58	88.80%
Random Forests	G1	Test	78	12	3	11	86.50%
Random Forests	None	Training	725	107	2	106	88.50%
Random Forests	None	Test	84	2	3	15	82.70%

Optimal	🟩
Overfitting	🟥

Table 1 – Evaluation of Random Forests

Analysis and critical evaluation of results

- Fig. 2 shows a confusion matrix for the best performing Naive Bayes model with no G1 and G2 attributes included in training (and consequently not used in predictions), this is the "worst" of the best models, with original data set data types and Holdout validation (0.5), the data set is therefore as small as it could possibly get under training conditions. We see that false positives and false negatives are well into the double figures, with the model performing marginally better than "guessing" the majority class for the training data set, and about the same as "guessing" with the test data set. Compared to other models obtained in the grid search, the increased proportional difference between training and testing accuracies suggests overfitting, and problems with lack of sufficient data.
- Fig. 3 shows a confusion matrix for the best overall model, a Random Forest which includes attributes G1 and G2, trained with adaptive logistic regression as an ensemble aggregation method and 100 Decision Trees. The model makes good predictions of majority as well as minority classes. With False Positives and False Negatives in the lower single figures. We found that, for the size of the data set, 100 decision trees was excessive in terms of increased accuracy, as accuracy tended to stabilize from 30 decision trees upwards. This may not be the case with a higher number of attributes, such as the 33 used in this study. Likewise, with fewer attributes, it may be the a Random Forest model can be trained just as efficiently with a smaller number of decision trees.
- Some attributes, although originally expressed as numeric, are in reality categorical, such as quality of family relationships (famrel) ranging from 1 (very bad) to 5 (excellent). We attempted to study the impact of data types in our models, as part of the grid searches, implicitly declaring categorical attributes as such. Overall, having categorical models declared as such (PredScheme = 2) did not make much impact with Naive Bayes models, the optimal models (less overfitting, less proportional difference between training and testing accuracies) being generated with the original implicit data types. With Random Forests, a majority of categorical data types generated extremely poor models to the point they should have been documented, as an artefact for further work.
- Naive Bayes best model (training accuracy 89.94%, testing accuracy 91.35%) performed better than expected and not too far behind the best Random Forest model (91.90%, 93.30%). This result suggests than in some scenarios and for specific cases, Naive Bayes classifiers might offer a viable alternative to Random Forests once the computational trade-off is accounted for. NB models were normally trained in tens of seconds, while RF took minutes, which, considering the small size of the dataset (1044 observations, 33 attributes, 1 added attribute) suggests the RF to be inefficient, though it could be optimized once optimal number of trees in forest had been established. Reducing the number of features remained to be investigated, although behind the scenes, RF would have sampled attributes, though we cannot interpret that effect in our models.
- Both models, when less (and knowingly important G1 and G2) attributes were used, tended to overfit. This highlighted two facts. One is the issue of insufficient data, the other and perhaps most important is that, as the motivation of this study is to automatically detect students that are going to do poorly early enough such that remedial action is still possible, the "earlier" models, that is to say, when intermediate grades G1 and G2 would not have be available from a chronological point of view, are the worst performing. Some questions we ask ourselves but were not able to answer; can accurate enough predictions be made before attributes G1 and G2 are available? Can the overfitting models be relied upon in any way? Would it be too late to take remedial action by the time attributes G1 and G2 became available? These could probably do with some domain knowledge.
- Both best models presented higher accuracy with test data prediction (NB 91.35%, RF 93.30%). This was consistently the case after some initial tuning and could be accounted for by small size of testing data. Since increasing the size of the testing data would as a result decrease the size of training data, increasing as a result the likelihood of overfitting, increasing the randomly sampled testing (10% of total number of observations) soon proved it was not an option.
- Given the domain, it is possible that even though the data set is relatively small, it is representative, and larger data sets would present the same pass/fail ratios we would be attempting to rebalance in the smaller data set. Such questions must be considered when artificially changing this ratio. There are differing views in the literature as to performance gains with respect to rebalancing data, specifically with smaller datasets[10].
- Both models NB and FR showed that with a smaller number of attributes (31), training accuracy surpassed testing accuracy, (NB 81.06% - 77.89%, RF 88.50%, 82.70%) once attributes were added (up to 33), testing accuracy begins to approach and eventually overtake training accuracy (NB 89.94% - 91.35%, RF (91.90%, 93.30%). This aspect could not be interpreted by our analysis.
- Overall the results are encouraging and suggest Naïve Bayes and Random Forest classifiers could have a role in student performance evaluation and decreasing the social cost of early school leaving.

PredScheme	Validation	Param	trainAcc	testAcc	trainAcc_G1	testAcc_G1	trainAcc_G1_G2	testAcc_G1_G2
1	KFold	2	78.30%	82.69%	82.98%	81.73%	86.70%	89.42%
1	KFold	3	77.98%	82.69%	82.87%	81.73%	87.02%	89.42%
1	KFold	4	78.19%	82.69%	83.83%	81.73%	87.66%	89.42%
1	KFold	5	78.30%	82.69%	83.51%	81.73%	87.13%	89.42%
1	Holdout	0.1	79.31%	81.73%	83.69%	81.73%	88.53%	89.42%
1	Holdout	0.15	78.85%	82.69%	84.48%	81.73%	87.86%	89.42%
1	Holdout	0.2	80.05%	80.77%	85.11%	82.69%	87.50%	88.46%
1	Holdout	0.25	78.01%	81.73%	84.40%	81.73%	88.37%	91.35%
1	Holdout	0.3	79.94%	78.85%	85.41%	82.69%	87.84%	89.42%
1	Holdout	0.35	81.02%	84.62%	83.96%	80.77%	88.05%	88.46%
1	Holdout	0.4	78.72%	80.77%	84.93%	81.73%	87.94%	89.42%
1	Holdout	0.45	79.88%	80.77%	84.72%	80.77%	89.94%	91.35%
1	Holdout	0.5	81.06%	77.89%	85.53%	80.77%	85.75%	89.42%
2	KFold	2	77.23%	82.69%	83.40%	81.73%	87.34%	88.46%
2	KFold	3	76.81%	82.69%	82.98%	81.73%	86.70%	88.46%
2	KFold	4	76.49%	82.69%	82.87%	81.73%	86.81%	88.46%
2	KFold	5	77.02%	82.69%	83.30%	81.73%	87.02%	88.46%
2	Holdout	0.1	79.31%	79.81%	83.33%	81.73%	87.23%	88.46%
2	Holdout	0.15	77.97%	79.81%	84.98%	80.77%	88.11%	89.42%
2	Holdout	0.2	78.72%	82.69%	84.84%	81.73%	87.77%	89.42%
2	Holdout	0.25	79.01%	80.77%	85.39%	80.77%	87.94%	88.46%
2	Holdout	0.3	78.27%	78.85%	83.28%	81.73%	88.15%	89.42%
2	Holdout	0.35	80.36%	79.81%	86.42%	82.69%	88.54%	90.39%
2	Holdout	0.4	76.95%	82.69%	85.64%	76.92%	88.30%	88.46%
2	Holdout	0.45	79.69%	78.85%	86.85%	79.81%	89.56%	90.39%
2	Holdout	0.5	80.64%	79.81%	85.75%	78.85%	89.15%	87.50%

Table 2 – Naïve Bayes Grid Search

Optimal	🟩
Overfitting	🟥

Lessons learned and future work

- Our study confirmed our hypothesis. Based on dataset provided, student performance can be predicted using models such as NB and RF. RF did outperform NB though the study could have benefited from a larger dataset, where overfitting was observed with smaller training sets
- Although student final grade G3 has a high positive correlation with past grades G1 and G2, and our models have shown to be more accurate when all available data set attributes are used, we have found in our exploratory data analysis there are other relevant features (e.g. lifestyle habits, parent’s job and education, alcohol consumption). Future work would benefit from establishing how other relevant features affect model accuracy. Random Forests, may have partially reached this end, though our models were not interpretable at this stage.
- The data sets were found to be unbalanced, as a result, a hypothetical model which only predicts the majority class could have matched the performance of the majority of Naïve Bayes models trained without attributes G1 and G2. To address this issue, a potential improvement would be to artificially boost the minority class.

References

[1] Belfield, C. (2008) Cost of Early School-Leaving and School Failure (p. 48). New York: Economics department, City University of New York
[2] Brunello, G. & Paola, M.D. 2014, "The costs of early school leaving in Europe", IZA Journal of Labor Policy, vol. 3, no. 1, pp. 1-31.
[3] Fernandez-Gutierrez, M. & Martinez, J. 2014, "THE NON-MONETARY COSTS OF EARLY SCHOOL LEAVING: ESTIMATION IN TERMS OF YEARS OF GOOD HEALTH", EDUCACION XX1, vol. 17, no. 2, pp. 241-263.
[4] Meier, Y., Xu, J., Atan, O. & van der Schaar, M. 2016;2015;, "Predicting Grades", IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959-972.
[5] 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.
[6] https://archive.ics.uci.edu/ml/datasets/Student+Performances Retrieved 23.11.2019
[7] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.
[8] Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324
[9] Manning, C. D., P. Raghavan, and M. Schütze. Introduction to Information Retrieval, NY: Cambridge University Press, 2008.
[10] Olson D.L. (2005) Data Set Balancing. In: Shi Y., Xu W., Chen Z. (eds) Data Mining and Knowledge Management. CASDMKM 2004. Lecture Notes in Computer Science, vol 3327. Springer, Berlin, Heidelberg