Ensemble Machine Learning System for Student Academic Performance Prediction

Yinkai Wang*
George Mason University
ywang88@gmu.edu

Kaiyi Guan†
George Mason University
kguan@gmu.edu

Aowei Ding†
George Mason University
ading@gmu.edu

Yuanqi Du†
George Mason University
ydu6@gmu.edu

ABSTRACT

Educational data mining is essential to our lives, as it provides the tools for us to analyze education-related problems and achieve better education. Particularly under COVID-19, the transition from in-person classes to online learning mode leads to a surge in educational data, and increasing attention has been drawn to educational data mining. However, many long-lasting problems in educational data analysis have not been properly solved due to various challenges. We identify several challenges in the student performance prediction task: (1) Good indicators for the final grade are hard to find and the final grades are only highly correlated to the grades in the previous terms. (2) The lack of one automatic and stable system that automates the data analysis cycle and produces satisfying prediction results. To tackle these two challenges, we implement an ensemble-based machine learning system for the student academic performance prediction task which automates the data analysis cycle in the absence of students' grades from previous terms. Specifically, our model consists of two components, ensemble feature engineering module and ensemble prediction module. Extensive experiment results have shown the superiority of our model over other traditional machine learning models, both in stability, efficiency and accuracy. Specifically, our model outperforms the best traditional machine learning algorithms by up to 14.8% in prediction accuracy.

Keywords

Ensemble Machine Learning, Classification Model, Machine Learning System, Student Performance Prediction

1. INTRODUCTION

From traditional lectures to novel E-learning, education quality is a long-lasting problem that we aim to improve [7]. While developing new tools for better education, student performance analysis is an essential yet challenging problem. During COVID-19, online education has been advanced and a large number of accumulated data has drawn tremendous attention [8]. Educational data mining community is growing to improve the education quality by better algorithms and larger datasets [9]. However, there are still many challenges left unsolved. (1) The high dependency between final grades and past grades. Even though past grades are good indicators for final grades, it lowers the explainability of the features, which we like to explore whether some indirect factors (e.g. internet access) could have an implicit effect on the grades. (2) The inefficiency and repeatable process of the manually-tuned models for analysis. Manually processing the data analysis cycle is very inefficient; an automatic system could highly speed up the data analysis process and make the process more stable. To solve the aforementioned challenges, we implement an ensemble-based automatic machine learning system for student performance prediction. Our model consists of two components, feature engineering and predicting, of each is an ensemble of multiple reliable algorithms. Our model is capable of analyzing the data and making predictions on student performance automatically, with more stability, effectiveness, accuracy and explainability. Our contributions are summarized as follows:

- A new machine learning system is proposed to improve the accuracy of student performance prediction tasks. Our model consists of two components, feature engineering, data sampling and predicting, to process and analyze the data.
- An ensemble-based feature engineering algorithm is proposed to enhance the acceptability and accuracy of the feature engineering phase.
- A graph-based ensemble prediction model is proposed to make a more stable and accurate student performance prediction.
- Extensive quantitative evaluations and qualitative analysis are performed to show the superiority of our proposed model in stability, efficiency, accuracy and explainability.

*equal contribution
†corresponding author

The road map of this paper is shown as follows. First, a brief summary of related works is explained in section 2. Then, we introduce the details of the proposed system in section 3. Next, the set-up and results of the experiment are detailed and analyzed in section 4. Finally, we conclude this paper with future works in section 5.

2. RELATED WORKS

2.1 Classification Model

Classification task is critical in many areas and has numerous applications, such as spam email detection, face classification, object classification, etc [5]. A large number of classification models have also been proposed to classify different types of data, such as K-nearest-neighbor, Random Forest, XGBoost, Neural Networks, etc [1, 3]. Many works [12] have proven the effectiveness of those well-known classification models in multiple data types. In this work, we introduce the idea of ensemble learning which takes multiple reliable prediction models and make final predictions based on the combined predictions from those models. We implement a graph-based ensemble classification model which constructs a bipartite graph by the selected classification models and make predictions over the propagation of the bipartite graph.

2.2 Ensemble Machine Learning

Ensemble machine learning is based on the idea to combine and take advantages of multiple models and achieve better and more stable performance. Two popular ensemble ideas are bagging and boosting, which bagging aggregates the results over multiple models, while boosting creates a stronger model by multiple weak models. Two well-known instances for bagging and boosting are Random Forest [1] and XGBoosting [3], respectively. We integrate the idea of bagging throughout our system, in all three modules. We aggregate the results of multiple feature engineering, data sampling and prediction models and make our final predictions.

2.3 Student Performance Prediction

Student performance prediction is a long-lasting problem in the educational data mining field [10]. Most of the recent works focus on applying the rising deep learning models to the problem [10], while rare people continue to explore the capability of traditional machine learning models. Another interesting research topic is the explainability of the models and the feedback to improve educational quality is exciting. Our model explores the capability of traditional machine learning models via an ensemble learning idea and the explainability of our approach by analyzing the features and results.

3. METHODOLOGY

3.1 Ensemble Feature Engineering Module

We implement an ensemble-based iterative feature selection algorithm which takes all the features from the beginning and iteratively search for better feature sets. The idea of this algorithm implements the greedy algorithm which follows the problem-solving heuristic to make an optimal choice every single step. Specifically, the algorithm takes measurement (i.e. importance) of each feature weighted by a list of reliable classification models. Every time, each feature is dropped from the feature set and the classification performance (accuracy) change indicates the importance of the feature. To make the feature importance more reliable, we ensemble multiple classification models and take the best one as the current best score. Then, it drops the worst negatively-affecting feature every iteration. Additionally, to leverage the side effect of the algorithm being too greedy, we set a stop mechanism, parameterized by k, which stops after a continuous lower performance in k iterations and rollback to the previous best feature set. After experiments with different values of k, the result showed that if drop positively-affecting feature more than 3 times, the accuracy will keep decreasing. In practice, we set this parameter to 3 and the list of classification models selected are [Random Forest, SVM, XGBoost]. The pseudo-code for the algorithm is shown in code block 1.

3.2 Ensemble Prediction Module

We implement a graph-based ensemble classification model, which takes several reliable classifications and clustering models and aggregates the results by the message propagation over a bipartite graph to achieve better performance. The idea is to ensemble over several reliable algorithms to achieve more accurate and stable results. It is very common (e.g. in Kaggle competitions) to ensemble the results of supervised learning models, while it is not that common to study how unsupervised learning models, such as clustering algorithms, can help in the ensemble process. Here, we aim to assist the prediction from the supervised classification learning models with certainty from the unsupervised clustering models. We first initialize all our models (classification, clustering) with Grid-searched parameters. We construct the graph $G = (V, E)$, where nodes V represent objects (either a data point X or a group Y), and edges E represent the connection between a data point and a group (class/cluster) determined by the classification/clustering algorithms. Empirically, we utilize three classification models and two clustering models. The total groups are 20 for classification models. We leave all of the groups on the right side of the bipartite graphs (i.e. 40 groups) because there is no consistent correspondence between different clustering models. Specifically, each edge is built if one classification model predicts the class of the sample or clustering algorithms cluster the samples together to one group. Because there is no supervision for the clustering algorithm, we utilize it to enhance the confidence of the classification models. We do so by calculating a confidence matrix $C \in X^{I,J}$, where I is the total number of the samples and J is the total number of the groups. Each entry $X_{i,j}$ represents the confidence score of a sample being in a group, which is initially measured by the predicting accuracy of the classification models. The confidence score is calculated via propagation over the graph:

$$X_{i,j} = X_{i,j} + \sum_{i=1}^{I} \frac{\frac{\sum_{j=1}^{J} X_{i,j}}{J}}{I} \tag{1}$$

Finally, the prediction class is determined by taking the label with the maximum probability, or certainty.

$$L_i = \underset{j}{\operatorname{argmax}}\, X_{i,j} \tag{2}$$

The groups are shown on the left and the bipartite graph is shown is the right of Fig. 1

Data: Student Performance Data Set
Initialize a list of all classification models M, model scores PM, all features G, scores for all features P as 0s, a stop counter K, parameterized by k, an empty list of temp features T, a best feature set B, a best accuracy A;
for *model m in the model list M* **do**
    Run the model m;
    Record the prediction accuracy to PM;
end
Record best accuracy A with highest score in PM;
Record best feature set B with highest score in PM;
Initialize stop counter $K = 3$;
while *condition* $K == 0$ *is not satisfied* **do**
    for *feature f in the feature list F* **do**
        Drop feature f;
        for *model m in the model list M* **do**
            Run the model m;
            Record $max(PM, prediction\ accuracy)$ to PM;
            Record $max(PM, prediction\ accuracy)$ to P;
        end
        Add feature f;
    end
    if $K == 0$ **then**
        Resume the best feature set B;
        Resume the best accuracy A;
        return the best feature set B, best accuracy A;
    else
        if *all scores in* $P > 0$ **then**
            Drop the feature with the highest score P drop;
            if *highest score > best accuracy A* **then**
                Record best accuracy A with highest score PM;
                Record best feature set B with highest score PM;
            end
        else
            Drop the feature with the lowest score P drop;
            K $-= 1$;
            Continue;
        end
    end
end

Algorithm 1: Ensemble Feature Engineering Module

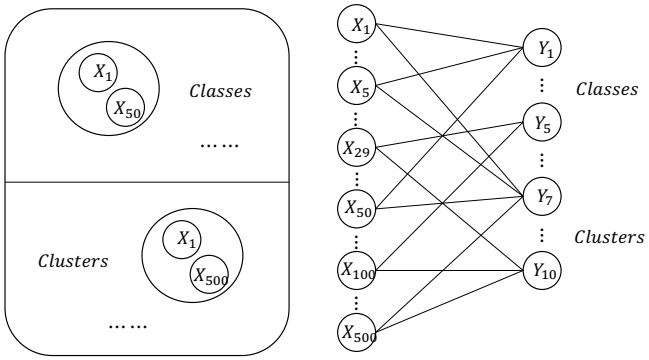

Figure 1: The groups of the graph-based ensemble classification model and the bipartite graph constructed by the model.

4. EXPERIMENTS

4.1 Experiment Set-up

Dataset Description. We take one commonly used benchmark dataset, Student Performance Data Set [4], from the UCI Machine Learning Repository[1]. The dataset consists of student achievements in secondary education of two Portuguese schools. It contains many important factors, including student grades, demographic, social and other school-related features. The detailed features are shown in 5. We analyze the dataset by visualizing the correlation among all features, in 3. Notably, first-period grade and second-period grade are highly correlated to the final grade. Interestingly, we find father education level is highly correlated with mother's education level. Mother's education level is highly correlated with mother's job, too. However, father's education level is not correlated with father's job. This provides some statistical insights about the correlations between father/mother education levels and their jobs. Additionally, workday alcohol consumption is highly correlated with weekend alcohol consumption. It is worth noting that the correlations among different term grades are extremely high and the distributions are close to a Gaussian distribution (Fig. 2. We aim to dig into other "hidden" features rather than the grade-related features in our setting.

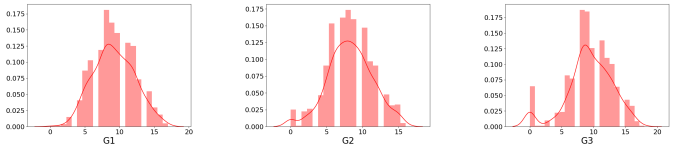

Figure 2: Score feature distributions of the dataset [4].

Comparison Methods Details. We take five reliable classification models, K-nearest-neighbor, SVM, Random Forest, XGBoost and Multi-layer Perceptron (MLP) as our baselines. We take the implementation of them from the scikit-learn library[2]. In terms of evaluation metrics, we take accuracy scores to evaluate the prediction performance of our

[1]https://archive.ics.uci.edu/ml/index.php
[2]https://scikit-learn.org/

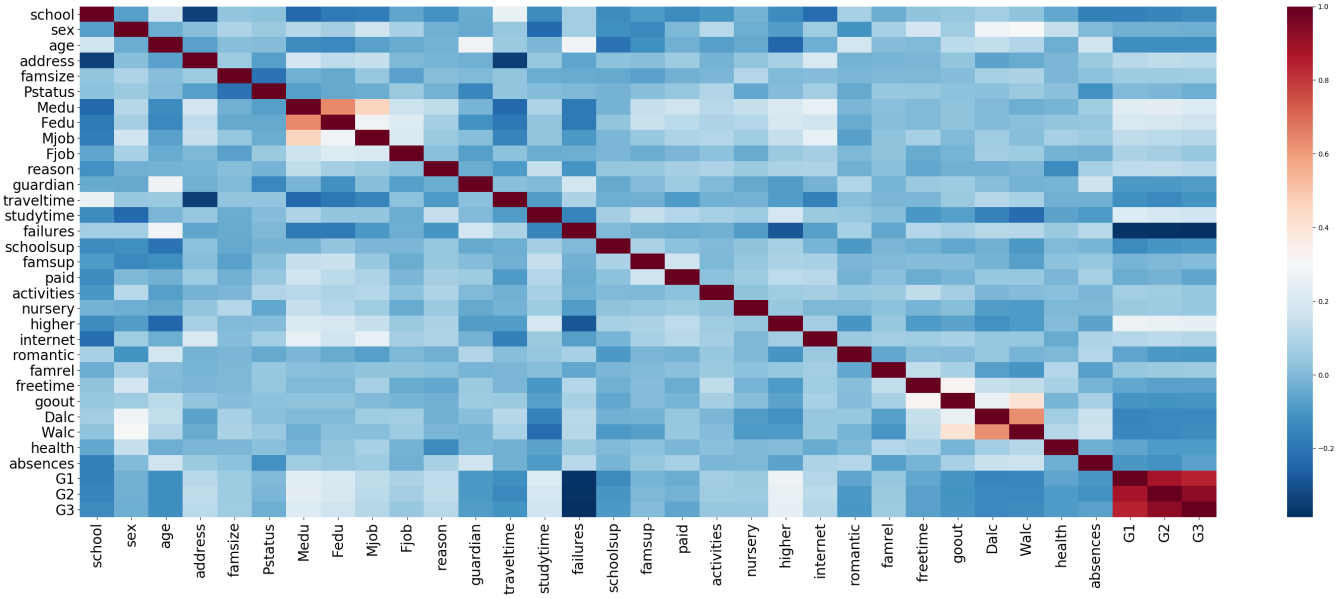

Figure 3: The heatmap shows the correlation among all the features.

model. For a fair comparison, the models that we use in our ensemble system are exactly the same as the baselines, i.e. with the same parameters. We run a grid-search to find the best setting of each model's parameter setting. We divide the train/validation/test set by 6/2/2.

4.2 Experiment Results

4.2.1 Ensemble Feature Engineering Evaluation

The final selected features by our model are school, sex, age, address, famsize, Pstatus, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, Dalc, Walc and health. The definition and details of each feature are shown in Table 5. We show the prediction accuracy of all the baseline models on data provided by different feature selection algorithms in table 1. Clearly, our ensemble feature selection model outperforms the raw data in all baseline models. To compare with other feature selection techniques, such as Principle Component Analysis (PCA) [11], and tree-based feature selection algorithm. For most of the cases, our model achieves the best performance. While the tree-based approach is also competitive while pairing with KNN or SVM. We suspect the reason is that our model is a little dominated by the best classification models during the feature selection phase. KNN and SVM are normally the worst models across all the models. Therefore, our selected features are not perfectly aligned with these two models. Besides higher accuracy, our model provides more explainability for researchers with a stronger feature selection module.

4.2.2 Data Sampling Evaluation

In order to see whether popular data re-sampling techniques are suitable for our problem, we select three well-known data re-sampling algorithms and take the implementation from the imblearn library [3]. To be specific, we take one oversampling, one undersampling and one combined-sampling algorithm, respectively. Specifically, we utilize SMOTE [2] for oversampling, TomekLinks for undersampling and SMOTE-Tomek for combined sampling [6]. The results are shown in table 2. Even though the feature re-sampling techniques show some promises in some cases, it is still not statistically reliable for us to incorporate into our system.

[3]https://imbalanced-learn.org/stable/

4.2.3 Ensemble Prediction Evaluation

We study the effectiveness of our ensemble prediction model by comparing it with other baseline models mentioned above. The results in table 3 suggest that our model is the best model, either with the presence of the grade features, or with the absence of the grade features. It is also worth noting that XGBoost ranks second in both settings which shows the great power of the algorithm. Random forest also achieves stable performance compared to KNN, SVM and MLP. Finally, KNN performs the worst in our testing. Additionally, table 3 suggests that our model is more stable as it achieves the best performance in both settings.

4.2.4 Ablation Study

In table 4, we show the ablation study of our system, in which we take out one component and see whether the other component improves the performance, compared to the baseline models. Clearly, both the feature engineering module and ensemble prediction module improve the result of our system.

5. CONCLUSION

In this paper, we implement an automatic ensemble machine learning system for student academic performance prediction, which consists of ensemble feature engineering and ensemble prediction modules. The extensive experiments have shown that each component of our system outperforms the traditional machine learning methods. Overall, the system further improves the prediction accuracy, up to 14.8%. In

Table 1: Comparison of the prediction accuracy w/ and w/o feature engineering.

Models	Raw-data	PCA-90	PCA-95	Tree-based	Ensemble-based
KNN	0.1100	0.1244	0.1388	**0.1579**	0.1483
SVM	0.1292	0.1388	0.1388	**0.1388**	0.1292
Random Forest	0.1435	0.1053	0.1244	0.1627	**0.1675**
XGBoost	0.1770	0.0766	0.1100	0.1483	**0.1771**
MLP	0.1435	0.0901	0.1069	0.1738	**0.1866**

Table 2: Comparison of the prediction accuracy w/ and w/o resampled data.

Models	Raw-data	Over-s	Under-s	Combined-s
KNN	0.1100	0.0957	**0.1292**	0.0861
SVM	**0.1292**	0.0287	0.1196	0.06670
Random Forest	0.1435	**0.1531**	0.1388	0.1483
XGBoost	**0.1770**	0.1579	0.1340	0.1483
MLP	0.1435	0.0718	**0.1627**	0.1005

Table 3: Comparison of the prediction accuracy w/ and w/o feature G1,G2.

Models	w/ Grade	w/o Grade
KNN	0.2679	0.1100
SVM	0.3014	0.1292
Random Forest	0.4019	0.1435
XGBoost	0.4498	0.1770
MLP	0.2679	0.1435
Proposed Model	**0.4641**	**0.1866**

the future, we plan to investigate how data augmentation could help in the task. We also hope that our proposed system can inspire others to reinvent traditional machine learning models which can work more efficiently, stably and accurately.

6. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[4] D. Dua and C. Graff. UCI machine learning repository, 2017.

[5] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab. A comprehensive survey for intelligent spam email detection. *IEEE Access*, 7:168261–168295, 2019.

[6] A. More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*, 2016.

[7] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[8] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355, 2020.

[9] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan. Mining in educational data: review and future directions. In *Joint European-US Workshop on Applications of Invariance in Computer Vision*, pages 92–102. Springer, 2020.

[10] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. Ding, S. Wei, and G. Hu. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[11] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[12] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Table 4: Comparison of the prediction accuracy between the proposed model and other models.

Models	Raw-data	Feature-selected-data
SVM	0.1292	0.1292
Random Forest	0.1435	0.1675
XGBoost	0.1770	0.1627
MLP	0.1435	0.1483
Proposed Model	**0.1866**	**0.1922**

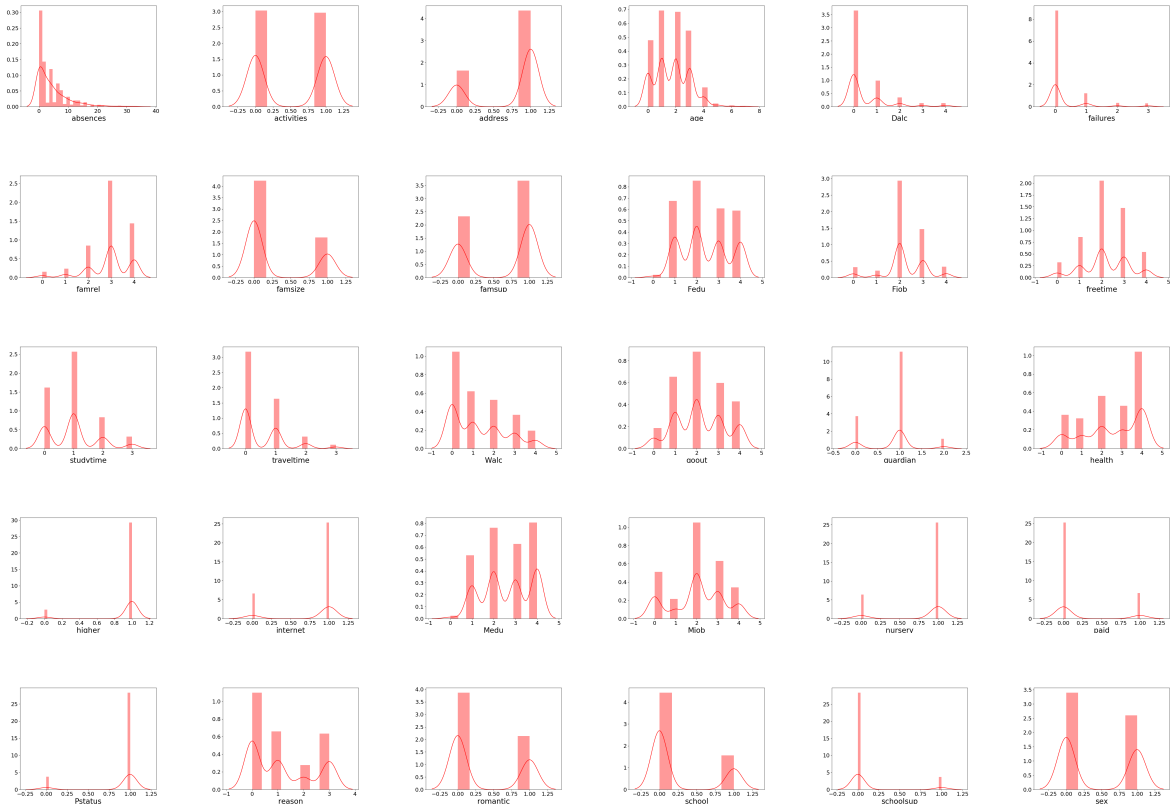

Figure 4: Feature distributions of the dataset [4].

Feature	Meaning	Value
school	School	Gabriel Pereira - 'GP' or Mousinho da Silveira - 'MS'
sex	gender	Female(F) or Male(M)
age	Age	from 15 to 22
address	Address	Urban - 'U' or Rural - 'R'
famsize	Family Size	less or equal to 3 - 'LE3' or greater than 3 - 'GT3'
Pstatus	Parent's Cohabitation Status	living together - 'T' or apart - 'A'
Medu	Mother's Education	none - 0, primary - 1, 5th-9th grade - 2, secondary - 3, higher - 4
Fedu	Father's Education	none - 0, primary - 1, 5th-9th grade - 2, secondary - 3, higher - 4
Mjob	Mother's Job	'teacher', 'health' care related, civil 'services', 'at_home', 'other
Fjob	Father's Job	'teacher', 'health' care related, civil 'services', 'at_home', 'other
reason	Reason to Choose this School	close to 'home', school 'reputation', 'course' preference, 'other'
guardian	Student's Guardian	'mother', 'father', 'other'
traveltime	Home-to-School Travel Time	<15min - 1, 15-30min - 2, 30-60min - 3, >60min - 4)
studytime	Weekly Study Time	<2hrs - 1, 2-5hrs - 2, 5-10hrs - 3, >10hrs - 4
failures	# of Past Class Failures	n if n <= 3 else 4
schoolsup	Extra Education Support	yes or no
famsup	Family Education Support	yes or no
paid	Paid Classes (Math or Portuguese)	yes or no
activities	Extra-curricular Activities	yes or no
nursery	Attended Nursery School	yes or no
higher	Wants to Take Higher Education	yes or no
internet	Internet Access at Home	yes or no
romantic	with a Romantic Relationship	yes or no
famrel	Quality of Family Relationship	from 1 - very bad to 5 - excellent
freetime	Freetime after School	from 1 - very low to 5 - very high
goout	Going out with Friends	from 1 - very low to 5 - very high
Dalc	Workday Alcohol Consumption	from 1 - very low to 5 - very high
Walc	Weekend Alcohol Consumption	from 1 - very low to 5 - very high
health	Current Health Status	from 1 - very bad to 5 - very good
absences	# of Absences	from 0 to 93
G1	First Period Grade	from 0 to 20
G2	Second Period Grade	from 0 to 20
G3	Final Grade	from 0 to 20

Table 5: Detailed collected features in the dataset [4].