40% of success is showing up: Random Forest Classification of Academic Success using non-grade features

Jackson McBride, Michal Hajlasz

Department of Data Science, Department of Computer Science | Columbia University

Introduction

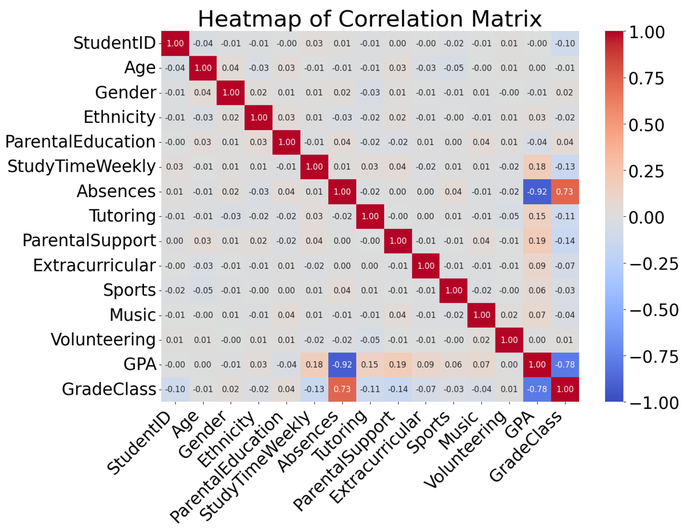
Our project aims to understand and predict the grade class of high school students as a function of 13 other features related to demographics, study habits, parental involvement, and extracurricular activities. A better understanding of indicators for academic performance may allow schools to better allocate resources to help their students succeed.

We obtained our dataset from Kaggle, uploaded by Rabie El Kharoua. The dataset contains detailed information on 2,392 high school students. The target variable, grade class, classifies student GPA into 5 discrete ranges (A, B, C, D, F). [3]

The dataset is largely unstructured with minimal regularity. Most variables exhibit low correlation with each other, and t-SNE analysis reveals no distinct clusters. Nonetheless, we were able to achieve 69% accuracy in predicting grade class with a random forest model. Our analysis shows that absences are by far the most important predictor for high school academic performance.

Correlation Analysis

The figure below presents a heatmap of the correlation matrix between all features in the dataset. Only three pairs of variables exhibit a correlation greater than 0.19 in magnitude.



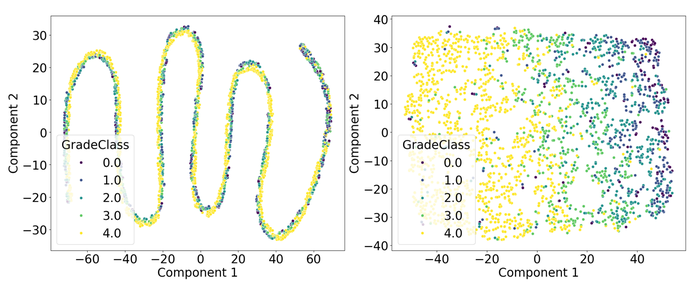
- GPA has a correlation of -0.92 with Absences.
- Grade class has a correlation of 0.73 with Absences.
- GPA has a correlation of -0.78 with Grade class.

As expected, grade class and GPA are highly correlated. The correlation is negative because a low grade class value indicates a high GPA. Excluding GPA, 'number of absences' is by far the feature by far the most correlated with grade class.

t-SNE Analysis

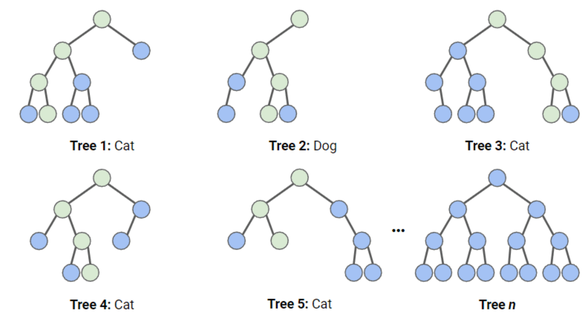
To better understand the structure of our data, we ran a t-SNE analysis, initially with all features, which resulted in the peculiar snake pattern of the figure on the left. This pattern is the result of an artificial structure induced by the variable StudentID. When the StudentID column is excluded, a t-SNE analysis produces the figure on the right, showing no distinct clusters in our data.

We color coded each student data point according to their grade class, with purple indicating highest performance and yellow indicating lowest performance. As we can see in the figure on the right, the aggregated feature Component 1 is highly correlated with grade class, with low performing students on the left and high performing students on the right. This indicates that some general notion of academic performance generates a lot of the structure in the dataset.



Model

Random Forests are powerful ensemble learning methods ideal for classification tasks. They operate by generating multiple decision trees, each independently classifying observations based on their features. A common problem with Decision Trees is their tendency to overfit the data. Random Forests mitigate this issue by randomly selecting subsets of features to build each tree during training. Once the forest of trees is constructed, the Random Forest model makes a classification decision for an observation based on the majority vote from all the trees (hence the name "ensemble method").[1]



Each decision tree in a Random Forest is constructed using a bootstrap sample from the training data. For a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, a bootstrap sample D_b is created by randomly sampling n observations with replacement.

At each node, a random subset of k features is selected from the m available features. The best feature from this subset is chosen to split the node based on a criterion like Gini impurity. This process continues until a stopping condition is met.

Mathematically, given S as a subset of k features, the optimal split is:

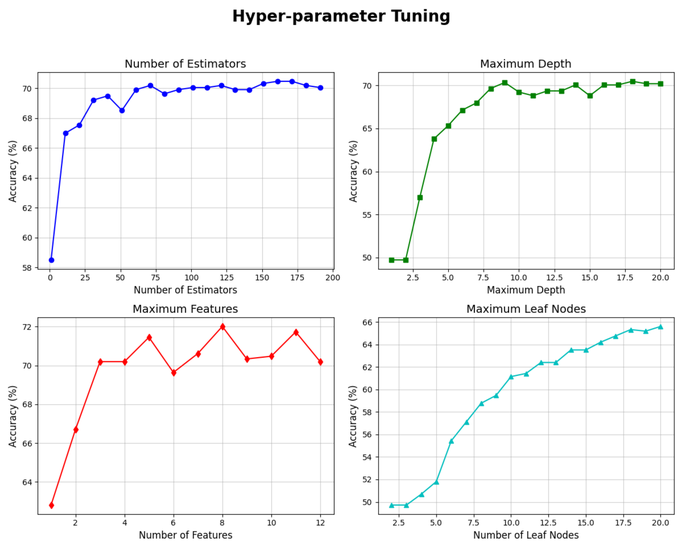
$$\text{argmin}_{f \in S} \sum_{i=1}^{|D_b|} L(y_i, f(x_i))$$

where L measures the impurity of the split. There are a number of ways to measure impurity, such as Gini Impurity, and Enropy.[2] We pick our impurity metric using cross validation.

Hyperparameter Tuning

We optimized our Random Forest classifier using iterative testing and GridSearchCV for cross-validation. The plots below illustrate the effects of different hyperparameters on model accuracy[4]:

- **Number of Estimators:** Accuracy improves with more trees, plateauing eventually.
- **Maximum Depth:** Deeper trees capture complexity but may overfit; optimal depth was found.
- **Minimum Samples Split:** Higher values reduce overfitting by requiring more samples to split a node.
- **Minimum Samples Leaf:** Larger values prevent overfitting by smoothing the model.
- **Maximum Features:** Balancing the number of features considered at each split enhances performance.
- **Maximum Leaf Nodes:** Limiting leaf nodes helps reduce overfitting.



Note: Each subplot shows the accuracy (%) for different values of the hyperparameters.

The Best Model

After conducting GridSearchCV with the following parameter grid:

- **n_estimators:** [50, 100, 200]
- **max_features:** ['auto', 'sqrt', 'log2']
- **max_depth:** [4, 6, 8, 10]
- **criterion:** ['gini', 'entropy']

We identified the optimal hyperparameters for our Random Forest classifier. The best model configuration is:

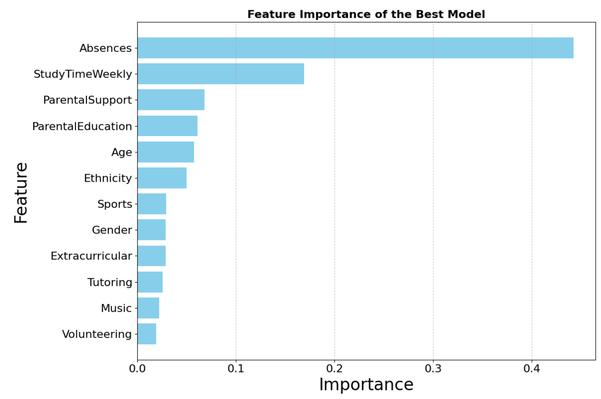
- **Criterion:** Gini
- **Max Depth:** 10
- **Max Features:** Auto
- **Number of Estimators:** 200

This configuration provided the highest cross-validated accuracy, indicating it effectively balances model complexity and performance.

Feature Significance

The bar plot below shows the importance of each feature for predicting student performance in our best model.

The most significant features include absences, weekly study time, parental support, and parental education, with absences contributing to over 40% to the model's overall feature importance.



Conclusions

Our classifier shows that using only non-academic metrics, we can effectively predict which academic bucket a student will fall into with nearly 70% accuracy. This is a non-trivial task, as illustrated by the indiscernible patterns observes in the t-SNE analysis. Furthermore, we find that the most important feature in predicting academic success is absences, whereas other features, like gender, do not provide the same predictive power.

Acknowledgements

We would like to acknowledge Emily Rothenberg for introducing us to the Data Science Symposium, as well as guiding and consulting us throughout the project. We would also like to acknowledge Isaac Chang for his creative contributions and emotional support.

References

[1] Leo Breiman. English. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A%3A1010933404324.

[2] Frank Farris. "The Gini Index and Measures of Inequality". In: *American Mathematical Monthly* 117 (Dec. 2010), pp. 851–864. DOI: 10.4169/000298910X523344.

[3] Rabie El Kharoua. *Students Performance Dataset*. 2024. DOI: 10.34740/KAGGLE/DS/5195702. URL: https://www.kaggle.com/ds/5195702.

[4] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.