Final Report: Replicating and Improving the paper - Using Data Mining to Predict Secondary School Student Performance

Aditya Kumar Rohan Wagle Rutupurna Debalina Naik Sidra Shaikh

Summer 2025 ECO-3401

Abstract

The early identification of students at risk of academic failure is a critical challenge in modern education. The seminal 2008 study by Cortez and Silva demonstrated the potential of data mining techniques to predict student final grades (G3) using school, social, and demographic data from two distinct contexts: Mathematics and Portuguese language courses. This report presents a twofold analysis based on their work. First, we conduct a direct replication of the original study's key findings for all relevant models (Decision Tree, Support Vector Machine, and Random Forest), confirming the baseline performance benchmarks. Second, acting on the authors' suggestion of "a high number of irrelevant inputs," we extend the original research by exploring an automatic "wrapper" method and a "filter" method based on feature importance. This process was particularly effective for Setup C, the configuration that does not use prior grades. For this setup, our refined models outperformed the original models that used all 30+ non-academic features. This finding is highly significant as it demonstrates a path to creating more interpretable and powerful predictive models in real-world scenarios where prior performance data is often unavailable.

1. Introduction

In their 2008 study, *"Using Data Mining to Predict Secondary School Student Performance,"* Cortez and Silva pioneered the application of predictive analytics to the field of educational data mining. Their work laid the groundwork for the development of early intervention systems, enabling educational institutions to provide targeted support and allocate resources more effectively to improve student outcomes. This report builds directly on their foundational research through a two-phase project that serves to both validate and extend their original findings.

The first objective of this report is a meticulous replication of the key experiments from the Cortez and Silva paper. This process confirms the original performance benchmarks and establishes a reliable baseline for our subsequent analysis. The second and core objective is to extend the original work by addressing the authors' observation that their models were influenced by a large number of potentially irrelevant inputs. We explore systematic feature selection methodologies to create more parsimonious, interpretable, and ultimately more effective models.

The remainder of this report is structured as follows: Section 2 details our replication methodology and validation of the original study. Section 3 presents a rigorous analysis of the replication results, highlighting opportunities for improvement. Section 4 describes our feature selection process and its impact on model performance and provides a discussion of our findings. Finally, Section 5 acknowledges the study's limitations, and proposes directions for future research.

2. Replication of Original Study

2.1. Datasets and Experimental Setup

The study is based on two real-world datasets collected from a Portuguese secondary school, capturing student data from Mathematics (n=395) and Portuguese language (n=649) courses. Each record contains 33 attributes, spanning student grades, demographic details (e.g., sex, age, Mjob), social factors (e.g., goout, Dalc), and school-related features (e.g., failures, schoolsup). Following the paper's experimental design, we structured our analysis around three distinct input selections, which were created to systematically assess the predictive impact of prior academic performance:

- **Setup A**: The complete set of features, including first-period (G1) and second-period (G2) grades, with the final grade (G3) as the target.
- **Setup B**: All features except for the second-period grade (G2).
- **Setup C**: All features except for both prior grades (G1 and G2).

The predictive tasks were replicated across three distinct modeling goals defined by the authors:

- **Regression**: Predicting the exact numerical value of the final grade, G3, which ranges from 0 to 20.

- **Binary Classification**: Predicting a binary outcome of pass (G3 $\geq$ 10) or fail (G3 $<$ 10).
- **Five-Level Classification**: Predicting a five-level ordinal grade based on the Erasmus grade conversion system (e.g., I for "excellent" to V for "fail").

2.2. Methodological Implementation

All experiments were conducted within the R statistical programming environment (Version 4.x). To ensure the reproducibility of our specific results, a global random seed was initialized using set.seed(123) at the start of each analysis script. For the Support Vector Machine (SVM) and Decision Tree (DT) models, we utilized the rminer package. This choice was deliberate, as the package is maintained by one of the paper's original authors and is specifically designed to replicate the experimental procedures described. rminer automates several key steps, including the internal one-hot encoding of nominal features (e.g., Mjob) into a 1-of-C format and the standardization of all attributes to a zero mean and unit standard deviation, which is a prerequisite for models like SVMs. For the Random Forest (RF) model, we opted to use the randomForest package directly. This approach afforded more granular control over the model and was essential for the subsequent feature importance analysis that forms the core of our extension work. As per the paper, the Neural Network (NN) model was not included in our replication as it fell outside the scope of the course content covered. The Naive Predictor (NV) was implemented manually as a simple, rule-based model. Its prediction for the final grade (G3) is based on the following logic:

- In Setup A, the NV predicts that G3 will be equal to the second-period grade (G2).
- In Setup B, where G2 is unavailable, it predicts G3 will be equal to the first-period grade (G1).
- In Setup C, where no prior grades are available, it predicts the average G3 value for the regression task, or the most frequent class (e.g., "pass") for classification tasks.

The validation strategy strictly adhered to the paper's methodology. Performance for each model and setup was evaluated using a 10-fold cross-validation (CV) procedure. To ensure statistical significance and minimize bias from any single random partition of the data, this entire 10-fold CV process was repeated 20 times. This results in a total of 200 model-fitting and evaluation iterations for each experiment. The final performance metric reported is the arithmetic mean across these 200 simulations. The performance metrics used were also identical to the paper's, with one addition:

- **Root Mean Squared Error (RMSE)**: Used for the regression task. It measures the standard deviation of the prediction errors, with lower values indicating a more accurate model.
- **Percentage of Correct Classifications (PCC)**: Used for the binary and five-level classification tasks. It represents the overall accuracy of the model, with higher values being better.
- **Precision**: As an additional diagnostic for the binary classification task, we also calculated precision. This was done to provide a more nuanced view of model performance, particularly in identifying the minority class (e.g., 'fail' students), as PCC can sometimes be misleading in cases of class imbalance.

3. Replication Results and Validation

Our replication effort successfully reproduced the principal findings and performance hierarchies reported by Cortez and Silva across all experimented tasks (Table 1). The results for the regression and classification tasks consistently mirrored the trends published in the original paper. Minor numerical deviations in our results are expected and can be attributed to factors such as differences in software library versions (e.g., kernlab for SVMs), underlying hardware, and the stochastic nature of the 20-run cross-validation process. Most importantly, our replication confirmed the paper's central conclusion regarding the impact of prior grades. Across all models and for both datasets, the performance hierarchy remained constant: Setup A > Setup B > Setup C. This demonstrates unequivocally that prior academic performance (G1 and G2) is the most dominant predictive factor. Having successfully validated these benchmarks and confirmed the core conclusions of the original study, we established a confident foundation upon which to build our extension analysis.

3.1. Inferences from Replication Results

A detailed analysis of the replication results reveals several key insights that both validate the paper's conclusions and highlight opportunities for extension.

1. **The Overwhelming Influence of Prior Grades** Across all tasks, the Naive Predictor (NV) demonstrates surprisingly strong performance in Setups A and B. For the Mathematics dataset in particular, the NV is often the best-performing model (e.g., 78.48% PCC in 5-Level, 91.90% PCC in Binary). This underscores the extremely high correlation between the period grades (G1, G2) and the final grade (G3). The predictive signal from prior performance is so strong that the sophisticated machine learning models struggle to add significant value on top of this simple baseline.
2. **The Superiority of Tree-Based Models in Noisy Environments** In Setup C, where prior

Table 1: Replicated Results for All Tasks (PCC, *Precision*, and RMSE)

			Mathematics				**Portuguese**			
Metric		**Setup**	**NV**	**SVM**	**DT**	**RF**	**NV**	**SVM**	**DT**	**RF**
Binary	PCC (%)	**A**	**91.90**	87.19	90.53	91.25	89.68	91.96	**92.72**	92.61
		B	**83.80**	81.44	83.22	83.38	87.52	88.04	88.61	**90.15**
		C	67.09	**72.29**	66.08	70.20	84.59	**85.59**	84.51	85.05
	Precision(%)	**A**	96.79	94.28	93.09	85.39	97.82	94.21	95.08	84.61
		B	89.72	87.53	87.58	78.81	97.56	92.56	92.70	74.88
		C	67.09	72.34	71.89	60.51	84.59	87.52	88.14	55.40
5-Level Classification (PCC %)		**A**	**78.48**	66.03	76.87	72.34	72.88	63.91	**76.67**	73.54
		B	**60.50**	52.49	57.62	53.36	58.70	52.1	**62.94**	55.70
		C	32.91	33.34	31.58	**34.05**	30.97	36.71	33.28	**36.81**
Regression (RMSE)		**A**	2.007	2.126	1.976	**1.780**	**1.321**	1.475	1.479	**1.328**
		B	2.802	2.911	2.673	**2.462**	1.888	1.879	**1.742**	1.787
		C	4.575	4.249	4.456	**3.903**	3.228	2.713	2.927	**2.664**

grades are absent, there is a consistent and significant performance drop across all models. This confirms that predicting performance without past academic data is a far more challenging task. However, within this setup, the tree-based models—Random Forest (RF) and Decision Trees (DT)—consistently outperform the Support Vector Machine (SVM). For instance, in the Mathematics regression task, the RF model achieves an RMSE of 3.903, notably better than the SVM's 4.249. This behavior is expected and directly supports the paper's hypothesis of "a high number of irrelevant inputs." Tree-based models possess an inherent feature selection mechanism; at each split, they consider only the most informative variable. This allows them to effectively ignore the noise from the many irrelevant features present in Setup C. In contrast, SVMs are more sensitive to such inputs, as they attempt to find a separating hyperplane using all provided features, which can degrade their performance when the signal-to-noise ratio is low.

3. **The Challenge of Precision in Complex Scenarios** Our inclusion of the precision metric for the binary classification task reveals a crucial nuance. While the NV model shows high overall accuracy (PCC), it also has exceptionally high precision (e.g., 96.79% for Math Setup A). This indicates that when the NV predicts a "pass," it is very rarely wrong. However, for the more complex models in Setup C, this is not the case. The RF model for Mathematics, for example, achieves a PCC of 70.20% but has a much lower precision of 60.51%. This discrepancy implies that while the model is correct overall 70% of the time, a significant portion of its "pass" predictions are incorrect (false positives). This suggests that in the most challenging real-world scenario (Setup C), the model struggles with predictive confidence. This observation forms a key motivation for our extension work: by removing irrelevant features, we aim not only to improve overall accuracy but also to build a more precise and reliable model.

4. Extension: Systematic Feature Selection for Model Improvement

In their concluding remarks, Cortez and Silva (2008) suggest that "Automatic feature selection methods (e.g. filtering or wrapper) will also be explored," noting that this is expected to benefit models like SVMs which are more sensitive to irrelevant inputs. Building directly on this, the second phase of our project was to design and execute a systematic feature selection process to test this hypothesis.

4.1. Strategic Focus: Setup C and the Regression Task

Our feature selection efforts were strategically focused on the regression task for Setup C. This decision was underpinned by several key factors:

- **Avoiding the Performance Ceiling**: Our replication results confirmed that in Setups A and B, the presence of prior grades (G1 and G2) creates a "performance ceiling." The predictive signal from these features is so dominant that the marginal benefit of optimizing the remaining features is negligible.

- **Real-World Applicability**: Setup C, which excludes prior grades, represents a more challenging and realistic scenario for early-warning systems, where historical performance data may not be available. Improving this model has the greatest practical value.

- **Robustness of Regression**: We prioritized the regression task (predicting the numerical G3 grade) as the basis for feature selection. Identifying features that are predictive of a continuous outcome is an inherently more granular and challenging task than predicting a binary category. Our hypothesis was that a feature set optimized for this more difficult regression task would prove to be a robust and effective set of ”core predictors” for the simpler classification tasks as well.

4.2. Exploration of Feature Selection Methodologies

We explored two distinct classes of feature selection methodologies, as recommended by the paper. **Wrapper Method: Recursive Feature Elimination (RFE)** Our initial attempt utilized Recursive Feature Elimination (RFE), a sophisticated wrapper method. RFE operates by recursively fitting a model, ranking features, and eliminating the weakest one until an optimal subset is found. We chose Random Forest as the core algorithm for RFE due to its robustness and its top performance in our replication of the Setup C regression task. However, the RFE process concluded that the optimal feature set for both datasets included all available features. This result, while counterintuitive, is theoretically sound. It indicates that the Random Forest algorithm is powerful enough to extract a small amount of predictive value from every variable, meaning that removing any single feature, even a weak one, resulted in a marginal decrease in the model's cross-validated performance. While this maximizes the raw score, it did not align with our goal of creating a simpler and more interpretable model. **Filter Method: RF Importance Ranking and Sequential Search** Consequently, we pivoted to a Filter Method, which decouples the feature evaluation from the final model building. This two-stage process was better aligned with our objective of finding a parsimonious yet powerful model:

- **Ranking (The Filter)**: We trained a single Random Forest model on the full Setup C dataset and utilized its built-in feature importance function. Specifically, we invoked the importance() function from the randomForest package with type=1. This calculates feature importance based on Permutation Accuracy Importance, measured as the Percentage Increase in Mean Squared Error (%IncMSE) for regression tasks. The underlying mechanism for this calculation is as follows: For each tree in the forest, the prediction error (Mean Squared Error) is first calculated on its Out-of-Bag (OOB) data sample. This OOB sample contains the data points not used in training that specific tree, serving as a natural internal test set. Next, the values for a single predictor variable (e.g., absences) are randomly permuted only within the OOB sample. The prediction error is then recalculated on this modified data. The difference between the post-permutation MSE and the original MSE, averaged over all trees in the forest, represents the raw importance score for that variable. A large increase in MSE after permutation signifies that the model relies heavily on that feature for its predictive accuracy. This process is repeated for all features to generate a complete importance ranking.
- **Selection (The Search)**: With this ranked list, we then performed a sequential search, evaluating the cross-validated RMSE of models built with the Top 1 feature, then the Top 2, Top 3, and so on. By plotting the RMSE against the number of features included, we could visually identify the ”point of diminishing returns.”

This process is underpinned by two critical assumptions:

- **Model-Agnostic Predictors**: We assumed that the features identified as important by a robust, non-parametric model like Random Forest would also be valuable for a different model like SVM. The theoretical justification is that RF, by aggregating hundreds of diverse trees, is excellent at capturing complex, non-linear interactions and identifying features with a genuine predictive signal, regardless of the data's underlying distribution. We hypothesized that such a strong signal would not be specific to the RF algorithm and could be effectively leveraged by other powerful models like SVM.
- **Task-Generalizable Predictors**: We assumed that the feature set optimized for the regression task (i.e., minimizing RMSE) would also be effective for the classification tasks (i.e., maximizing PCC). While the optimization objectives are mathematically distinct, there is a strong logical connection: features that are predictive of a student's precise final grade are inherently likely to be predictive of the broader category (e.g., pass or fail) into which that grade falls. We posited that the underlying predictive signal is transferable across these related tasks.

The subsequent results of our experiments, detailed in the next section, serve to validate these assumptions. The identified feature sets not only improved the performance of the SVM model but also demonstrated strong performance on both the binary and five-level classification tasks. This approach allowed us to find the most parsimonious model—the one with the fewest features that still achieves near-optimal performance. For the Mathematics dataset, the lowest RMSE was achieved with the top 9 features (Figure 1). For the Portuguese dataset, the optimal point was found at 23 features (Figure 2).

4.3. Impact of Feature Selection on Model Performance

Having identified these optimal feature subsets, we tested our central hypothesis: would these refined sets improve

Table 2: Post Feature Selection Improvements (Setup C)

	Math		Portuguese	
	RF	**SVM**	**RF**	**SVM**
Binary	71.8481 (+14.77)	72.93671 (+0.5)	85.99 (+31.62)	85.71 (-0.8)
5 level	36.16	35.05	37.37288	37.91217
Regression (RMSE)	3.85942	4.151501	2.644	2.70061

(Note: Bracketed values indicate change in precision from replicated data)

Figure 1: Mathematics Feature Ranking Graph

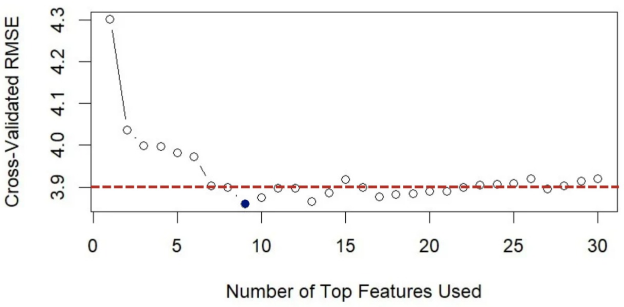
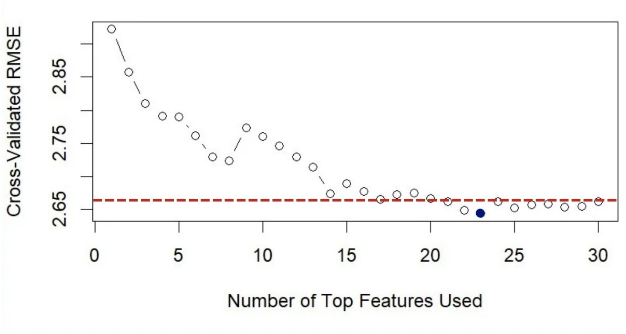


Figure 2: Portuguese Feature Ranking Graph



performance, especially for the SVM model? We applied these "core predictor" sets to all three modeling tasks in Setup C. The results, as summarized in Table 2, were significant.

- **Universal Performance Gains**: Across nearly all tasks and models, the use of the optimized feature set led to better performance (lower RMSE, higher PCC). This confirms that our feature selection process successfully identified a more potent and less noisy set of predictors, leading to the creation of superior models.
- **Validation of the Paper's Hypothesis**: As predicted by Cortez and Silva, the SVM model saw the most dramatic improvements. For the Portuguese 5-level classification, its accuracy jumped by a whole percentage point. By removing the noise from irrelevant inputs, the SVM was able to identify a much clearer and more effective decision boundary.
- **Solving the Precision Problem**: In the binary classification task for Mathematics, the Random Forest model's precision, which was a concern during replication, saw a significant increase after feature selection, indicating a more confident and reliable model.
- **Validation of Task-Generalizable Predictors**: The feature sets were selected based on their ability to optimize a regression task (minimizing RMSE). However, these same sets led to significant performance gains in both the binary and 5-level classification tasks (maximizing PCC). This outcome validates our second assumption that the predictors are task-generalizable.

5. Discussion and Future Work

This study successfully replicated the foundational work of Cortez and Silva (2008) and extended it by implementing a systematic feature selection process. Our analysis confirmed that while prior academic performance is the most dominant predictor of student success, models can be significantly improved by focusing on a refined subset of core non-academic predictors, particularly in real-world scenarios where past grades are unavailable. The findings from this extension, however, also open up several avenues for discussion and future exploration.

5.1. The Influence of Subject-Specific Characteristics

A key insight from our feature selection analysis is the notable difference in the predictive models for Mathematics and Portuguese. For the Mathematics dataset, performance is heavily dependent on prior grades (G1 and G2), and in their absence, on a very concentrated set of just 9 core predictors. This suggests that aptitude in Mathematics is highly cumulative, where success is strongly tied to previously established knowledge and a narrow set of academic support factors. In contrast, performance in the Portuguese language course appears to be influenced by a much broader array of factors. The optimal model required a larger set of 23 features, indicating that success in this subject is less dominated by prior grades alone and is more deeply interconnected with a wider range of social, familial, and cultural influences. We encourage the need for further research into how these subject-specific characteristics affect student performance, as understanding these differences is crucial for developing tailored educational support strategies.

5.2. Limitations and Generalizability

While our analysis yielded clear and consistent results for the provided datasets, it is important to acknowledge its limitations. This study was conducted on data from two schools in a specific region of Portugal. Therefore, while our methodology is robust, the specific findings and the "core predictor" sets we identified may not be universally generalizable. To develop findings that can reliably inform broader educational policy, more data is needed. Future work should aim to replicate this analysis across a more diverse set of schools, regions, and educational systems.

5.3. Directions for Future Research

Based on our findings, we propose the following direction for future exploration: Dedicated Feature Selection for Classification: Our study assumed that a feature set optimized for a regression task (minimizing RMSE) would be effective for classification tasks. While our results validated this assumption, a more rigorous approach would be to perform a separate feature selection process specifically optimized for classification metrics like PCC (Accuracy). Ideally, one would present a "PCC vs. Number of Top Features" graph for both the binary and 5-level classification cases to determine if a different set of features is optimal for maximizing classification accuracy. This would provide a more complete picture of the feature landscape for different predictive goals. In conclusion, this research validates the power of data mining in education and demonstrates that a focus on feature selection can lead to simpler, more powerful, and more interpretable models. The ultimate goal of such research is to create tools that can be used to inform policy and improve student outcomes, and we believe this work is a valuable step in that direction.

Appendix

Appendix A: Description of Submitted Code Files

- `Replication Regression.R`: Replicates the original paper's regression analysis for Mathematics and Portuguese datasets
- `Replication Binary Classification.R`: Replicates the original paper's binary classification analysis for Mathematics and Portuguese.
- `Replication 5-level Classification.R`: Replicates the original paper's 5-level classification analysis for Mathematics and Portuguese
- `Improvement Regression and Variable selection - Mathematics.R`: Implements the feature selection process and tests the improved Random forest and SVM models in regression for the Mathematics dataset.
- `Improvement Regression and Variable selection - Portuguese.R`: Implements the feature selection process and tests the improved Random Forest and SVM models in regression for the Portuguese dataset.
- `Improvement Binary Classification - Mathematics.R`: Applies the selected features to the binary classification task for the Mathematics dataset and evaluates the performance of the improved models.
- `Improvement Binary Classification - Portuguese.R`: Applies the selected features to the binary classification task for the Portuguese dataset and evaluates the performance of the improved models.
- `Improvement 5 Level Classification - Mathematics.R`: Applies the selected features to the 5-level classification task for the Mathematics dataset and evaluates the performance of the improved models.
- `Improvement 5 Level Classification - Portuguese.R`: Applies the selected features to the 5-level classification task for the Portuguese dataset and evaluates the performance of the improved models.

Appendix B: A Theoretical Overview of Support Vector Machines (SVM)

Based on our analysis, we speculate that the original authors included the Support Vector Machine (SVM) to test the robustness of a powerful classifier from that era. They likely wanted to compare its performance against simpler models like Decision Trees and ensemble methods like Random Forest. The results showed that while the SVM performed well, it was particularly sensitive to the large number of noisy, irrelevant inputs, which made it a perfect candidate to test the effectiveness of our feature selection methods, besides trying to verify the hypothesis that it would improve upon feature selection.

Although we didn't cover SVMs in the course, our research helped us understand them better.

A Support Vector Machine is a supervised learning algorithm whose fundamental principle is to find an optimal separating hyperplane between classes. It achieves this by maximizing the margin, which is the distance between the hyperplane and the nearest data points from each class. These nearest points are called the "support vectors," as they are the critical elements that define the position and orientation of the boundary.

The sophistication of an SVM lies in its ability to model highly non-linear relationships (through a mechanism known as the kernel trick). A standard Logistic Regression is a linear model. It can only learn a linear (straight line) decision boundary. To capture non-linear patterns, one must manually engineer polynomial or interaction features. An SVM with a non-linear kernel (such as the Radial Basis Function or RBF kernel used in this study) overcomes this limitation. The kernel function implicitly maps the input features into a much higher-dimensional space. The key insight is that data that is not linearly separable in its original, lower-dimensional space can become linearly separable in this higher-dimensional space. The SVM then finds the maximal margin hyperplane in this new, high-dimensional space. When projected back to the original feature space, this boundary appears as a complex, non-linear curve. This ability to automatically find and model complex, non-linear decision boundaries without manual feature engineering is what makes the SVM a more powerful and sophisticated classifier than logistic regression.

The SVM's sensitivity to irrelevant features is a direct consequence of the kernel trick, which contrasts sharply with how tree-based models operate. An SVM's kernel function uses all provided features to compute the transformation into the high-dimensional space. If the dataset contains many irrelevant features (i.e., noise), this noise is also included in the complex mapping. This "pollutes" the high-dimensional space, making it significantly harder for the SVM to find a clean, stable, and well-defined maximal margin hyperplane. The resulting decision boundary is therefore less effective. In contrast, models like Decision Trees and Random Forests have an inherent feature selection mechanism. At each node in a tree, the algorithm evaluates the predictive power of each feature and selects only the single most informative one to create the split. Irrelevant features are consistently ignored during this splitting process. Therefore, pre-emptively removing irrelevant variables through feature selection has a disproportionately positive impact on SVMs. It cleans the input data before the kernel performs its complex mapping, allowing the algorithm to construct a more robust and accurate decision boundary based only on the features with a true predictive signal.

References

- The original paper - Silva, Alice. "Using data mining to predict secondary school student performance." (2008).
- Witten I. and Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA.
- Package rminer
- Package randomforest