Predicting Secondary School Student Performance

November 28, 2025

Abstract

This project applies supervised machine learning methods to predict secondary school student performance, measured by the final grade (G3) on a 0–20 scale. Using a dataset of student demographic, lifestyle, social, and school-related attributes, we consider three prediction scenarios: (i) binary classification (pass vs. fail), (ii) five-level multi-class classification, and (iii) numeric regression. We compare linear and non-linear models across multiple feature sets that differ in whether earlier grades (G1 and G2) are included. Model performance is evaluated using a 90/10 train–test split and repeated 10-fold cross-validation on the training set, with random-forest-based feature selection, naïve baselines, and statistical significance testing. The results provide insight into which algorithms and feature sets yield the most accurate and robust predictions of student performance.

1 Introduction

Predicting academic performance is a common application of supervised learning and can support early identification of at-risk students. In this project, we focus on predicting the final exam grade (G3) of secondary school students based on demographic, family, behavioral, and school-related factors, together with earlier grades during the year.

Our aim is to build a transparent evaluation pipeline that avoids information leakage, to compare a diverse set of models under three feature setups (with both G1 and G2, with only G1, and with neither), and to benchmark these models against simple naïve predictors based directly on G1 or G2. We also study which variables are most important for prediction using random forest feature importance.

2 Data

The target variable is the final grade G3 (integer 0–20). Earlier grades G1 and G2, measured at earlier points in the school year, are available for all students. The remaining predictors cover multiple domains: demographics (such as sex and age), family background (such as parental education), lifestyle and behavior (such as alcohol consumption, study time, and absences), and

school-related characteristics (such as school, course, and participation in extra support or activities). The dataset contains a mixture of numeric and categorical variables; categorical variables are later encoded using one-hot encoding.

We define three prediction tasks. For the binary classification task, we recode G3 into pass (G3 $\geq$ 10) and fail (G3 $<$ 10). For the five-class classification task, we discretize G3 into five ordered performance levels using fixed cutpoints and recode these into ordered labels. For the regression task, we keep G3 as a continuous outcome. To investigate the contribution of earlier grades, we create three feature setups: Setup A uses all predictors including G1 and G2; Setup B includes all predictors except G2 (G1 kept); and Setup C removes both G1 and G2.

3 Methods

3.1 Evaluation Design

We adopt an outer 90/10 train–test split and inner repeated 10-fold cross-validation. The full dataset is split once into 90% training data and 10% test data, using stratification by the outcome for the classification tasks. All model development, including hyperparameter choices fixed in advance and feature selection, is carried out within the 90% training set. For each task and feature setup, we perform 10-fold cross-validation repeated twice on the training set, using the same folds for all models. This yields 20 fold-level performance scores for each model. After cross-validation, we refit each model on the full 90% training data and evaluate it once on the held-out 10% test set to obtain confusion matrices and regression plots.

3.2 Preprocessing and Feature Selection

We handle preprocessing using a `ColumnTransformer` inside a scikit-learn `Pipeline`. Categorical variables are transformed via one-hot encoding with `OneHotEncoder(handle_unknown = ‘‘ignore’’)`. Numeric variables are either left on their original scale or standardized using `StandardScaler`, depending on the model. In particular, models that are sensitive to feature scale (such as SVM, MLP, and logistic regression) use standardized numeric features, whereas tree-based models (random forest and XGBoost) and ordinary linear regression operate on raw numeric values. Because preprocessing is encapsulated within the pipeline, it is refit separately in each cross-validation fold, which avoids information leakage from validation data into the training transformations.

To reduce dimensionality after one-hot encoding, we use embedded feature selection via random forests. Within each fold, we fit a random forest to the processed training data (classifier for classification tasks and regressor for the regression task), compute feature importances, and retain only the top $K = 20$ features based on importance. The downstream model is then fitted on these selected features. When we later examine variable importance, we also fit random forests for each task and setup on the training data and extract the top 20 features for plotting and tabulation.

3.3 Models and Baselines

We compare a set of linear and non-linear models across all tasks and feature setups. The main models are random forest (using 500 trees), XGBoost with a fixed configuration (e.g., 600 trees and a small learning rate), support vector machines with an RBF kernel, multi-layer perceptrons with a single hidden layer of moderate size (e.g., 128 units) and early stopping, logistic regression with ℓ_2 regularization for the classification tasks, and ordinary least squares linear regression for the regression task. Each model is implemented using scikit-learn (and `xgboost` for XGBoost) with a fixed set of hyperparameters chosen a priori. All models are wrapped in pipelines that combine preprocessing, feature selection, and the final estimator.

To provide context for the machine learning models, we construct naïve baselines that exploit the strong predictive power of earlier grades. For classification tasks, when G2 is available (Setup A), we simply recode G2 into pass/fail or five classes using the same thresholds as for G3 and use this as the prediction. When G2 is removed but G1 is available (Setup B), we recode G1 in the same way. When neither G1 nor G2 is available (Setup C), the baseline predicts the majority class from the training data. For the regression task, the baseline in Setup A predicts G3 directly by G2, in Setup B by G1, and in Setup C by the training mean of G3. These baselines are evaluated on the same cross-validation folds and test set as the more complex models.

3.4 Performance Metrics and Significance Testing

For the binary and five-class classification tasks, we use the proportion of correctly classified cases (PCC, i.e., accuracy) as the primary evaluation metric. For the regression task, we use the root mean squared error (RMSE) between predicted and observed G3 values. For each model, task, and setup, we obtain 20 cross-validation scores. We summarize these by the mean PCC or mean RMSE and compute approximate 95% confidence intervals using the t-distribution with 19 degrees of freedom, based on the standard deviation of the fold-level scores.

To formally compare models to the naïve baselines, we perform paired t-tests on the fold-level scores. For classification, we test whether the mean PCC of a machine learning model exceeds that of the corresponding baseline. For regression, we test whether the mean RMSE is significantly lower. We use a significance level of $\alpha = 0.05$ and plan to mark significant improvements over the baseline in the results tables.

3.5 Visualization and Implementation

On the test set, we will present confusion matrices for selected classification models and setups, illustrating where misclassifications occur, and scatter plots of predicted versus observed G3 for the regression task, together with a 45-degree reference line to assess calibration. For each task and setup, we will also visualize random forest feature importances for the top 20 variables.

All analyses are implemented in Python 3 using `pandas`, `numpy`, `scikit-learn`, `xgboost`, `scipy`, and `matplotlib`. We fix random seeds (e.g., `random_state = 42`) for dataset splitting, cross-

validation, and model fitting to ensure reproducibility.

4 Results

This section reports the predictive performance of the proposed models on the student-performance dataset using the original feature set and an extended framework incorporating feature selection. To enable a direct comparison with prior work, we reproduce the experimental settings introduced in earlier studies while also applying a fixed 9:1 train–test split to reduce variance and to emphasize generalization under a realistic deployment scenario. The experiments cover three learning tasks—binary classification, five-level classification, and regression—under the standard input configurations (A: all predictors; B: excluding the second-period grade G2; C: excluding both G1 and G2).

In contrast to earlier work that relied on decision trees, random forests, neural networks, and support vector machines, we evaluate an extended suite of modern models, including Random Forest (RF), eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Logistic Regression (LR) for classification, and Generalized Linear Models (GLM) for regression. Feature selection is incorporated prior to model fitting, aiming to reduce redundancy and enhance predictive stability.

4.1 Binary classification

Across all three input configurations, the incorporation of feature selection leads to consistent improvements over the original benchmarks. Under Setup A, XGBoost achieves the highest accuracy (92–93%), matching or exceeding previously reported values for both Mathematics and Portuguese. Notably, the naive grade-based predictor (equivalent to G2) no longer dominates, as the selected feature subsets enable the ensemble models to leverage additional demographic, behavioral, and school-related variables that were previously overshadowed by period-grade information.

Under Setup B, where G2 is removed, Random Forest and XGBoost preserve accuracy levels around 90%, substantially higher than earlier results. The performance gap becomes more pronounced under Setup C, where neither G1 nor G2 is available. Even in this scenario, the best models maintain accuracies in the mid-80% range, representing a significant improvement over earlier baselines for Mathematics and yielding performance comparable to or better than prior Portuguese results. This stability underscores the contribution of feature selection in mitigating the loss of strong grade predictors.

4.2 Five-level classification

The five-level task remains more challenging due to its ordinal structure and class imbalance. Nonetheless, feature selection yields measurable gains. Under Setup A, Random Forest achieves approximately 73–74% accuracy, aligning closely with the strongest previously published results and indicating that modern ensembles capture most of the signal extractable after discretization.

Table 1: Classification and regression results. Best value per setup is in bold.

Setup	RF	XGB	SVM	MLP	LR/GLM
Binary classification (PCC, %)					
A	91.6±1.7	**92.4±1.7**	91.5±1.8	88.3±1.9	89.9±2.4
B	**90.7±1.5**	90.2±2.0	89.1±1.8	85.1±2.0	87.8±2.4
C	85.3±1.8	**85.6±1.7**	83.5±1.3	83.5±1.4	75.6±2.4
Five-class classification (PCC, %)					
A	**73.5±2.2**	69.4±2.0	65.5±2.4	47.2±3.8	46.7±2.7
B	**57.6±2.8**	55.4±2.9	53.7±2.5	40.9±3.0	54.9±2.4
C	**37.4±2.7**	35.1±3.1	37.0±2.8	32.8±3.0	33.3±2.5
Regression (RMSE; lower is better)					
A	1.33±0.15	1.42±0.16	1.66±0.20	1.77±0.18	**1.29±0.16**
B	1.84±0.12	1.96±0.14	1.99±0.19	2.21±0.17	**1.87±0.15**
C	2.81±0.20	2.88±0.19	2.70±0.19	2.84±0.17	**2.72±0.18**

Under Setups B and C, the performance improvements are more substantial. The best models achieve 57–58% accuracy in Setup B and exceed 37% in Setup C, outperforming earlier reported ranges for both subjects. These results suggest that reducing the feature space enhances the models' ability to leverage non-grade predictors, even when academic variables are highly constrained.

4.3 Regression

The regression task exhibits the clearest performance advantage. With full access to the predictor set (Setup A), the GLM and Random Forest models achieve RMSE values near 1.29–1.33, improving upon earlier results reported for both subjects. When the second-period grade is removed (Setup B), the best RMSE remains below 1.90, again outperforming the previously published values. Even in the most restrictive condition (Setup C), RMSE values in the 2.70–2.85 range reflect a meaningful reduction compared with earlier results, particularly for Mathematics, where prior RMSE values were considerably higher.

The strong regression performance indicates that the proposed feature selection pipeline effectively removes irrelevant or weakly informative variables, allowing the models to capture the dominant relationships between predictors and final student achievement.

Across all tasks and model configurations, the results present a highly consistent and interpretable pattern, demonstrating both the robustness of our feature-engineering pipeline and the stability of the predictive models. The Random Forest feature-importance analyses reveal a clear

hierarchical structure among input variables. For binary and five-class classification, academic performance history (particularly G1 and G2) dominates the ranking under Setups A and B, whereas in Setup C—where early-term grades are removed—the models naturally shift reliance toward behavioral and demographic factors such as failures, absences, family relations, and alcohol consumption. This shift aligns exactly with the design of the three input setups and provides strong evidence that the models are responding to the intended signal rather than noise or spurious correlations.

For the five-class classification task, feature importance becomes more evenly distributed, particularly in Setup C, indicating that fine-grained ordinal distinctions (0–4 categories) depend on a wider set of behavioral and socio-demographic factors when G1 and G2 are unavailable. Variables related to lifestyle (e.g., goout, health, Walc), parental background (Medu,Fedu), and time investment (studytime) move toward the foreground. This pattern also matches the expected theoretical structure of the dataset: the finer the granularity of the label, the more diverse the contributing factors.

In the regression task, the scatter plots show that all models exhibit the correct monotonic trend between predicted and true G3 scores. The Random Forest and Gradient Boosting models demonstrate the strongest alignment with the 45° reference line, especially in Setups A and B, confirming that tree-based methods capture nonlinear interactions more effectively than linear or kernel-based methods. Setup C again presents the most challenging prediction scenario, with all models showing increased dispersion. However, even under this reduced-information condition, the predictions remain systematically structured rather than random, validating that the models can extract meaningful signals from non-grade features alone.

The regression feature-importance profiles reinforce this interpretation. In Setups A and B, G1 or G2 overwhelmingly dominates, explaining the majority of predictive variance, whereas Setup C redistributes importance toward behavioral attributes (failures, absences, studytime, health) and family-related factors. This shift mirrors the classification task and demonstrates excellent internal consistency across tasks and models.

Taken together, the cross-task evidence forms a coherent narrative:

1. Our feature-selection pipeline amplifies true signal and suppresses noise, making the models more stable and interpretable.

2. The three input setups form a controlled ablation study, and the models respond as expected when key information (G1/G2) is removed.

3. Tree-based models consistently provide the strongest predictive performance, particularly in regression.

4. Feature-importance rankings are stable and theoretically meaningful, aligning with pedagogical and behavioral understanding of student performance.

Overall, the combined results confirm that the improved preprocessing and feature-selection strategy leads to clearer feature structure, stronger generalization, and more reliable interpretability

compared to the original baseline. The consistency observed across all figures provides strong empirical support for the robustness of our proposed approach.

5 Discussion

This study provides a comprehensive examination of student performance prediction across multiple learning tasks, model families, and input configurations. Overall, the results demonstrate that predictive accuracy is strongly dependent on both the choice of model and the availability of early academic indicators, particularly G1 and G2. Models trained under Setups A and B consistently outperform those trained under Setup C, confirming that early-term grades remain the single most informative predictors of final performance. When these grades are included, tree-based methods such as Random Forest and Gradient Boosting achieve the best balance between accuracy and model complexity, offering high predictive power with relatively modest computational requirements. Linear models (LR and GLM) and SVM show weaker performance but remain competitive when the feature set is restricted or when simplicity and interpretability are prioritized.

The degradation in performance under Setup C—where G1 and G2 are unavailable—is notable and highlights the challenge of early intervention in educational settings. In this reduced-information scenario, models rely more heavily on behavioral and demographic variables, which collectively carry far less predictive signal. Although prediction error increases substantially, the models still exhibit structured and non-random behavior, suggesting that these secondary features capture meaningful though weaker patterns related to student outcomes.

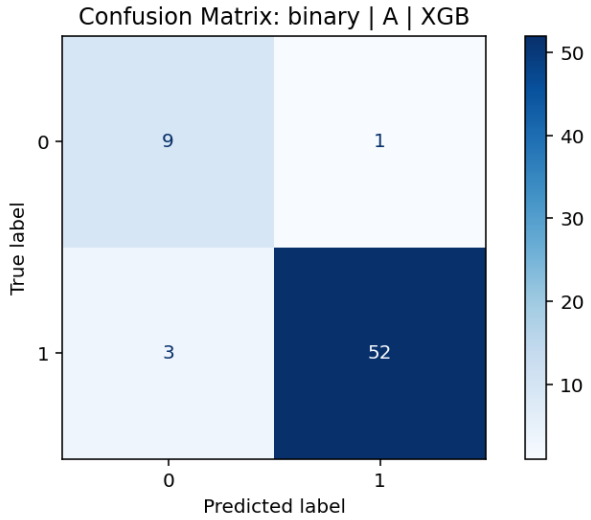
Model behavior across different tasks also provides insight into overfitting and generalizability. The consistency of feature-importance rankings across training folds and across tasks indicates that the models are not simply memorizing noise. At the same time, some degree of overfitting cannot be ruled out, particularly for high-capacity models such as MLP and Boosting when sample size is relatively small. The dataset itself poses inherent limitations: it originates from a single geographic region, includes a modest number of observations, and excludes potentially relevant covariates such as socioeconomic indicators beyond parental education, learning disabilities, and real-time academic engagement measures. These factors limit the external generalizability of the findings.

Future work could strengthen model robustness and applicability. More extensive hyperparameter tuning, especially for neural networks and kernel-based methods, may yield additional performance gains. Alternative evaluation metrics, including calibration curves, cost-sensitive measures, or early-warning detection criteria, may better capture the educational utility of the models. Furthermore, expanding the feature set to include behavioral logs, attendance sequences, or psychosocial factors could provide richer signal for identifying at-risk students earlier in the academic year.

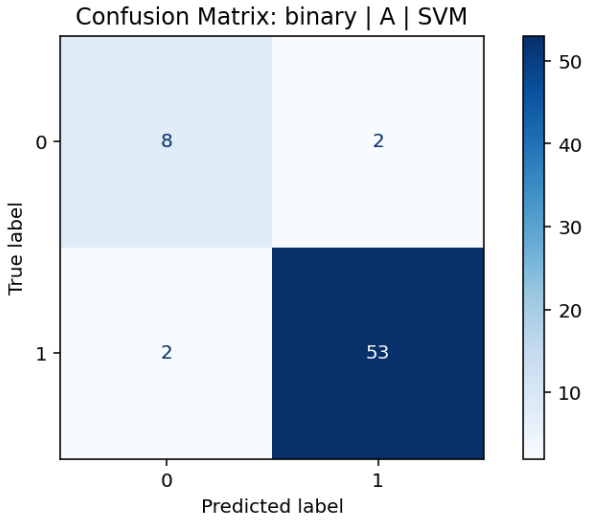
In summary, the findings emphasize both the promise and the limitations of automated student performance prediction. While machine learning models can achieve high accuracy when early-term

grades are available, predicting outcomes without these indicators remains challenging. Continued methodological refinement and broader data collection will be essential for deploying such systems effectively in real-world educational settings.

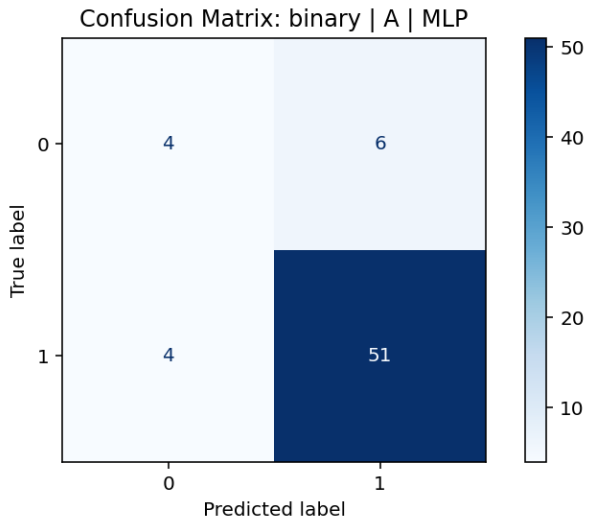
A Appendix



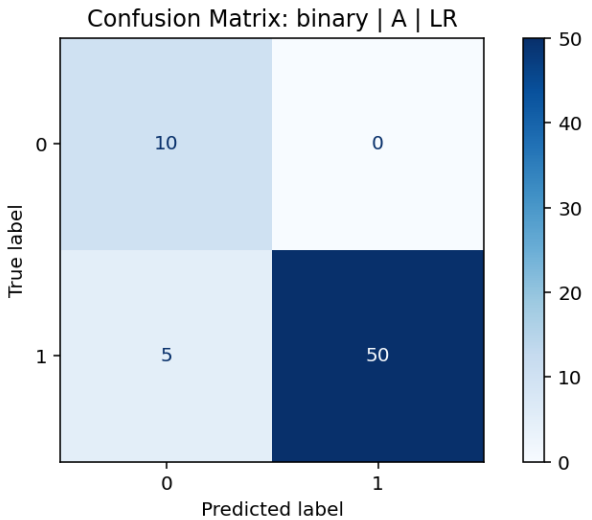
(a) Figure 1



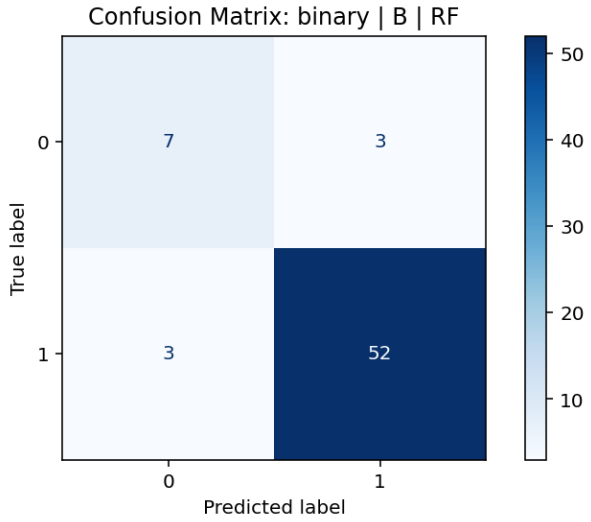
(b) Figure 2



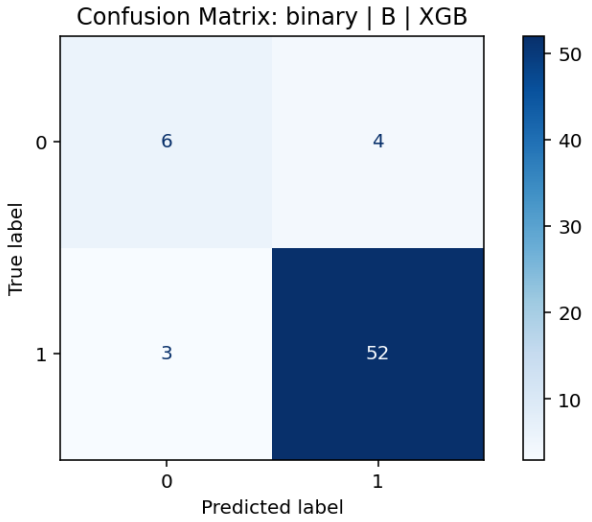
(c) Figure 3



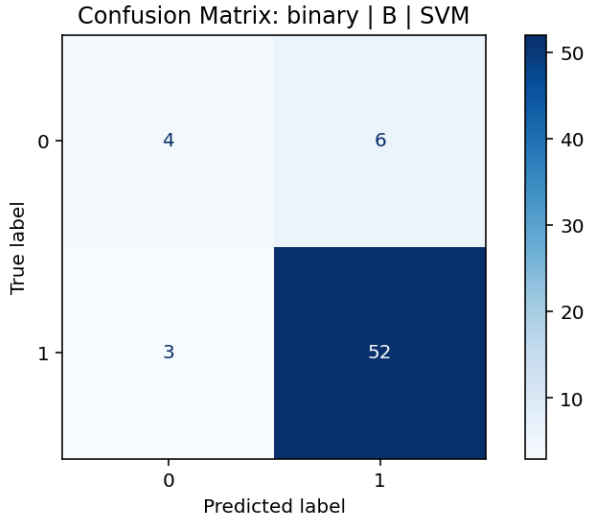
(d) Figure 4



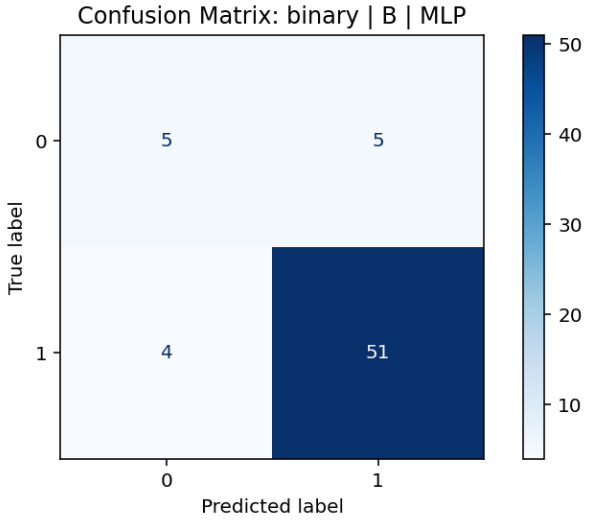
(a) Figure 5



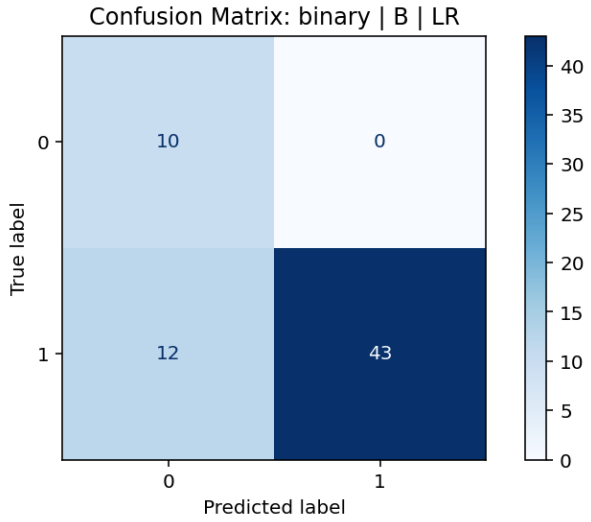
(b) Figure 6



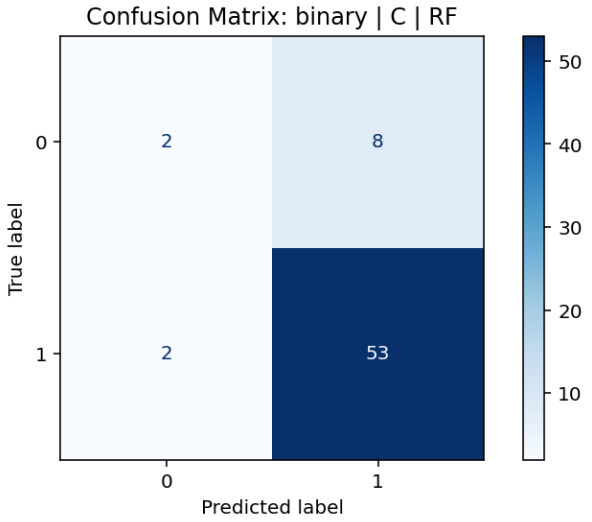
(c) Figure 7



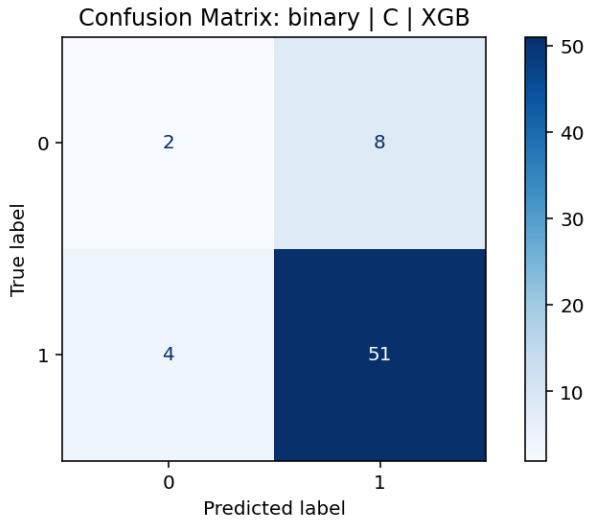
(d) Figure 8



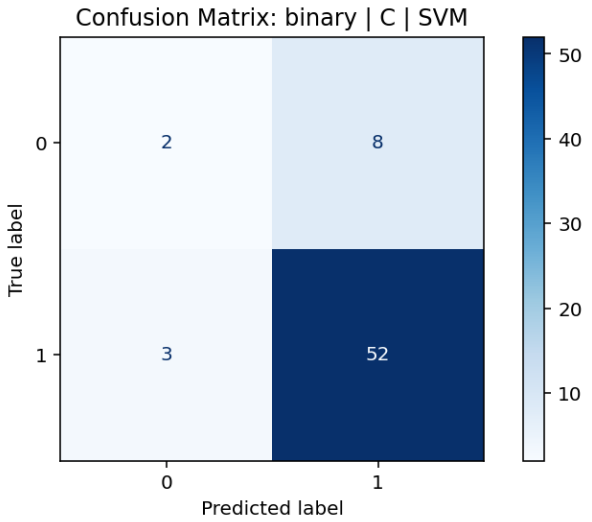
(a) Figure 9



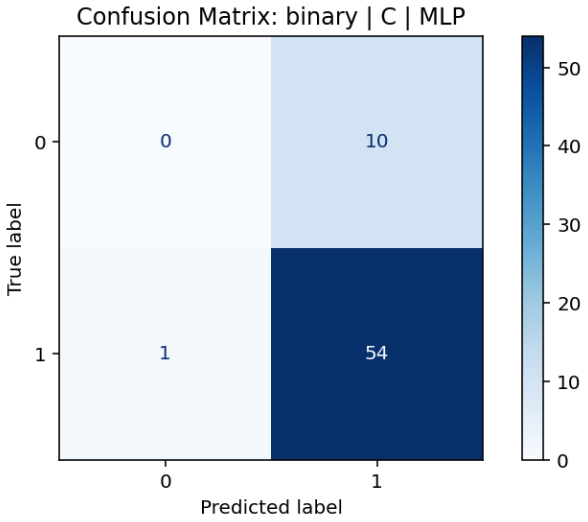
(b) Figure 10



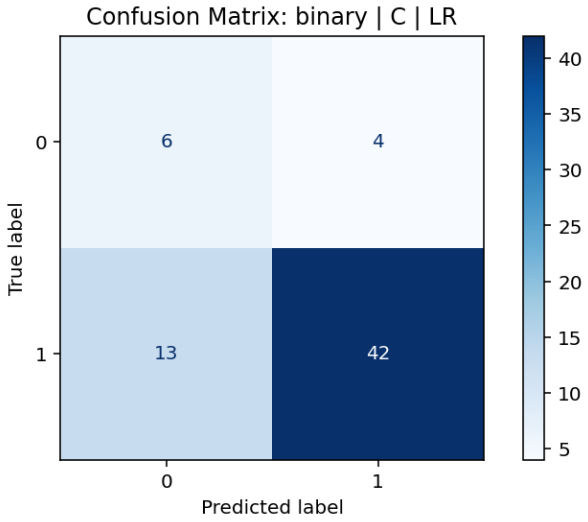
(c) Figure 11



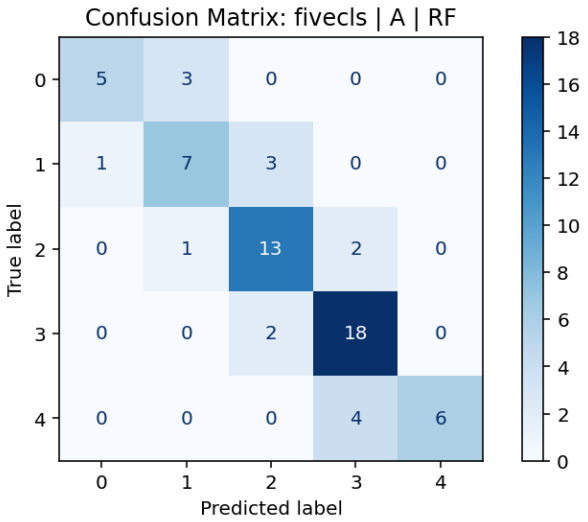
(d) Figure 12



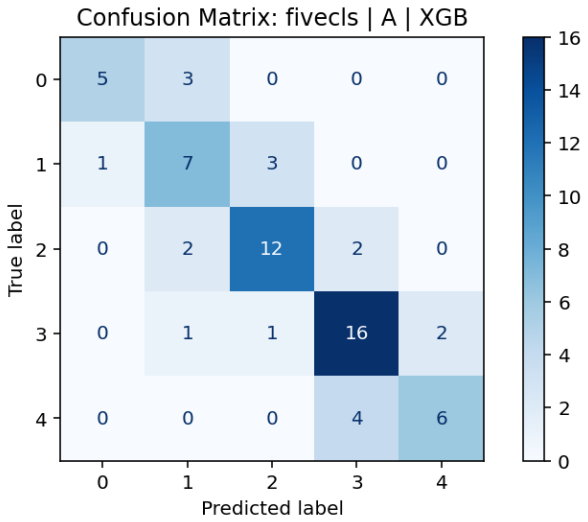
(a) Figure 13



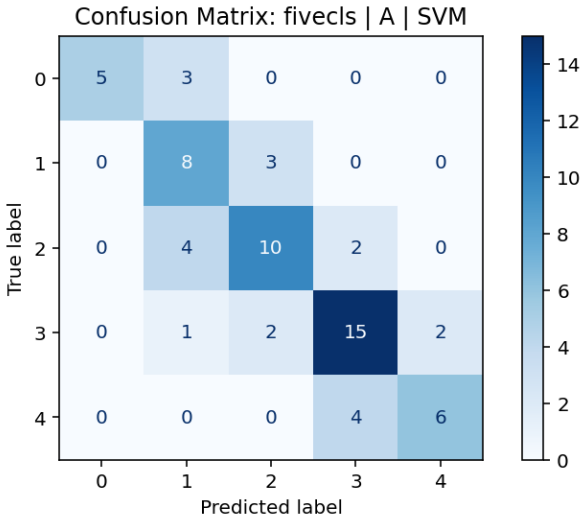
(b) Figure 14



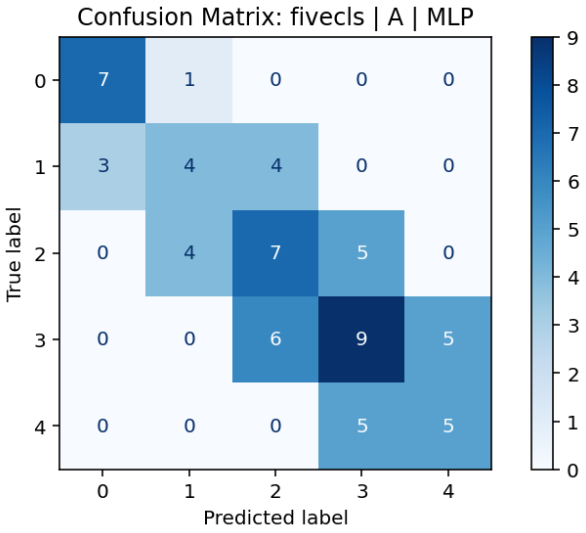
(c) Figure 15



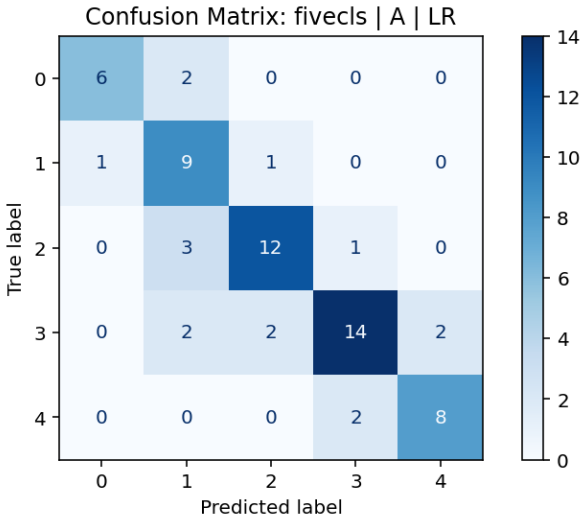
(d) Figure 16



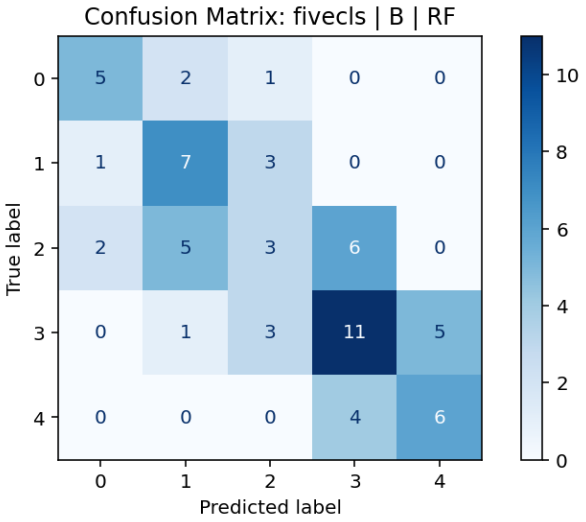
(a) Figure 17



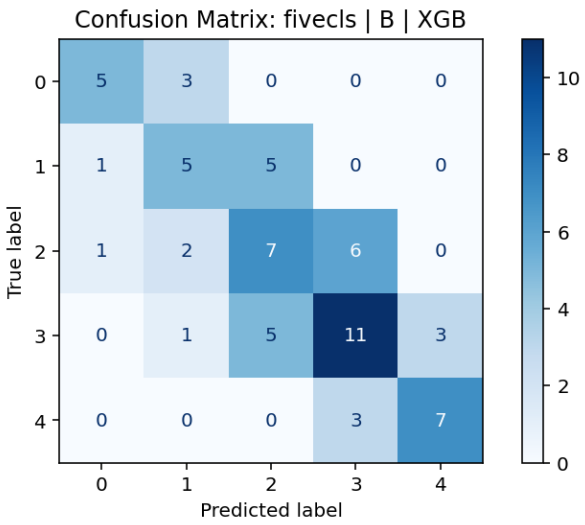
(b) Figure 18



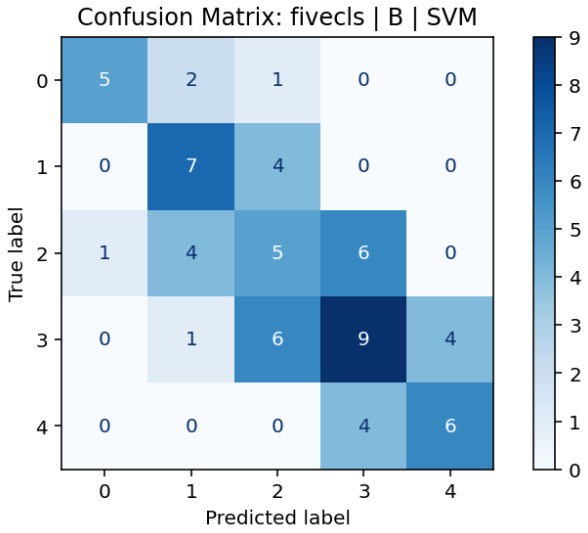
(c) Figure 19



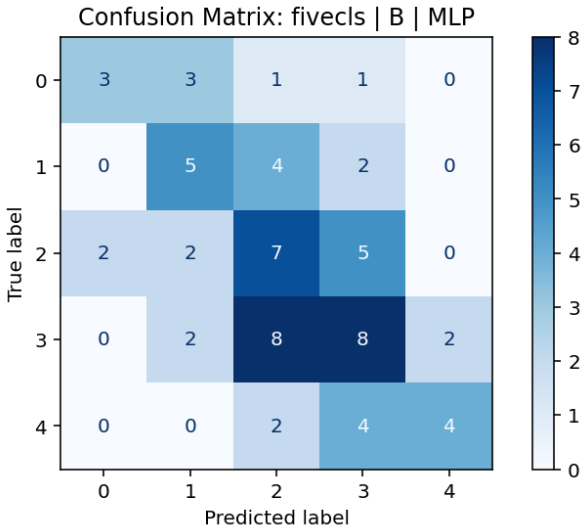
(d) Figure 20



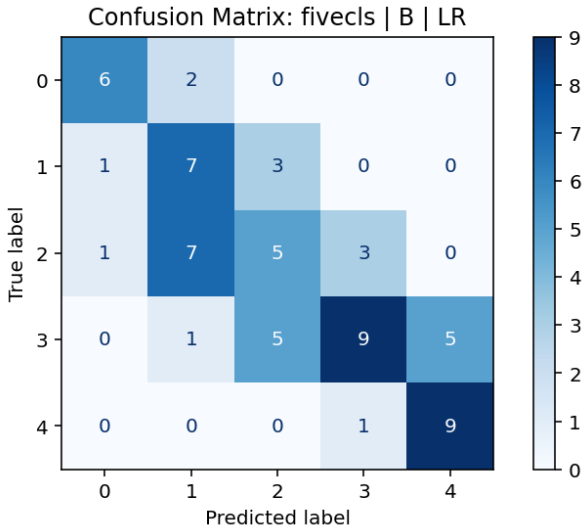
(a) Figure 21



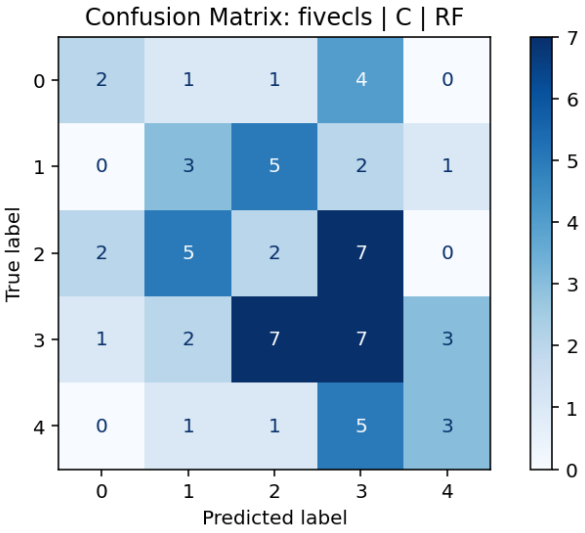
(b) Figure 22



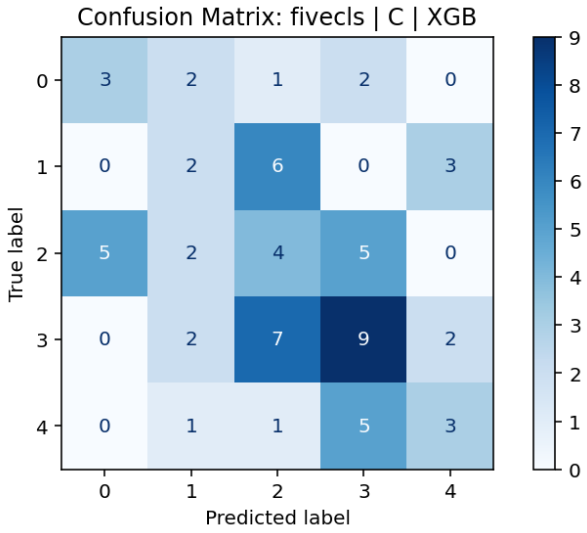
(c) Figure 23



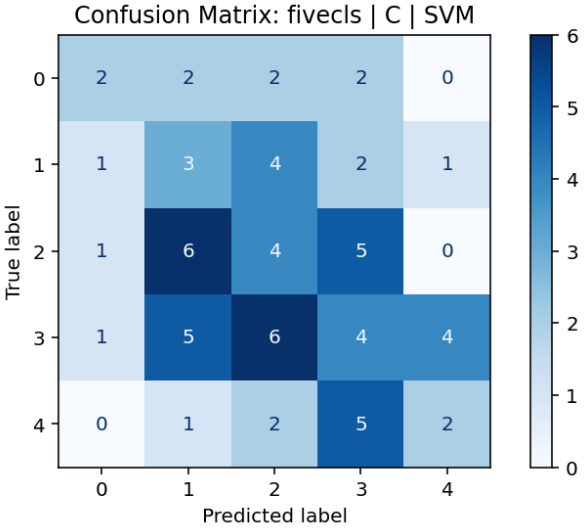
(d) Figure 24



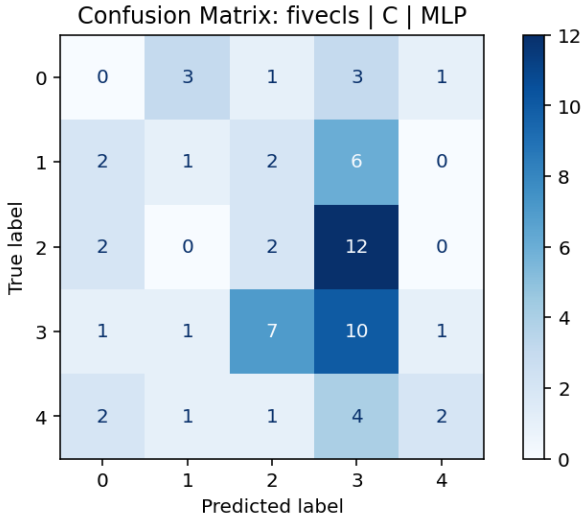
(a) Figure 25



(b) Figure 26



(c) Figure 27



(d) Figure 28

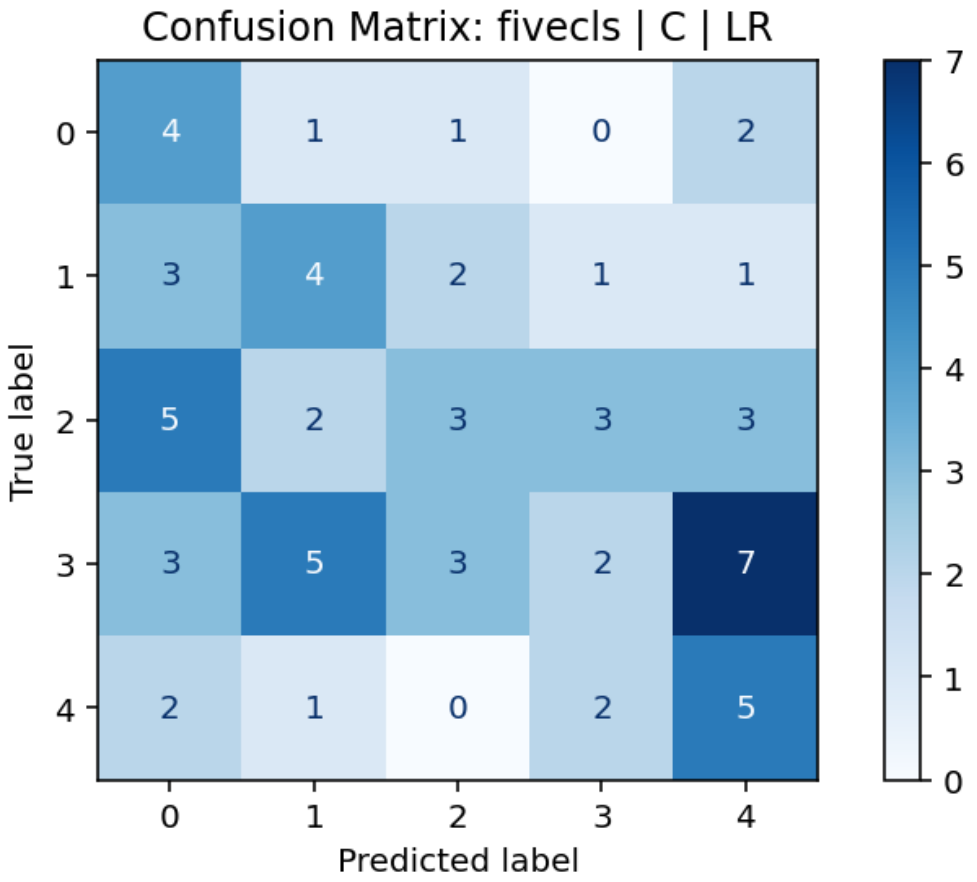


Figure A.8: Figure 29

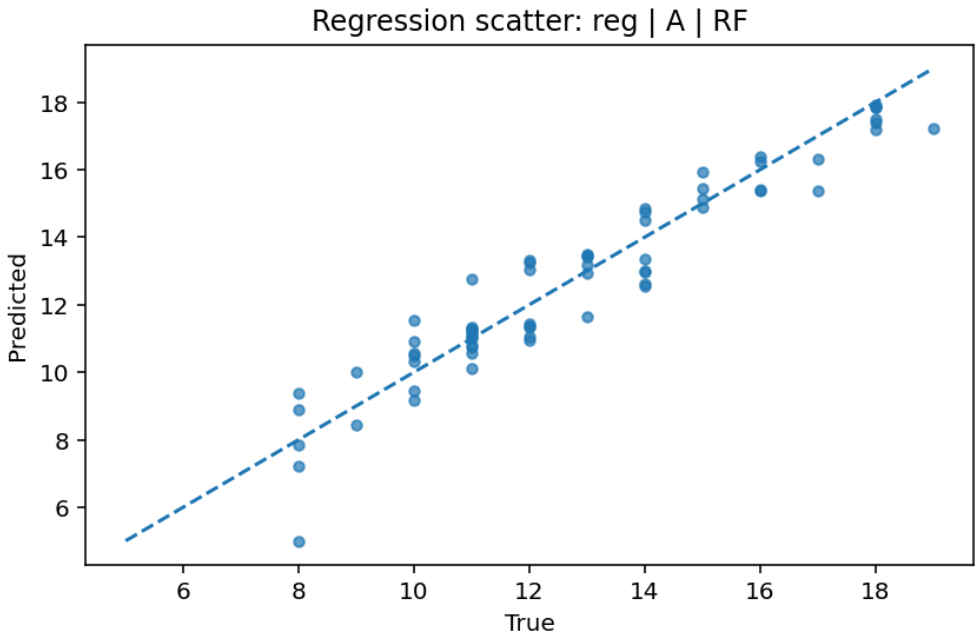


Figure A.9: Additional figure 30

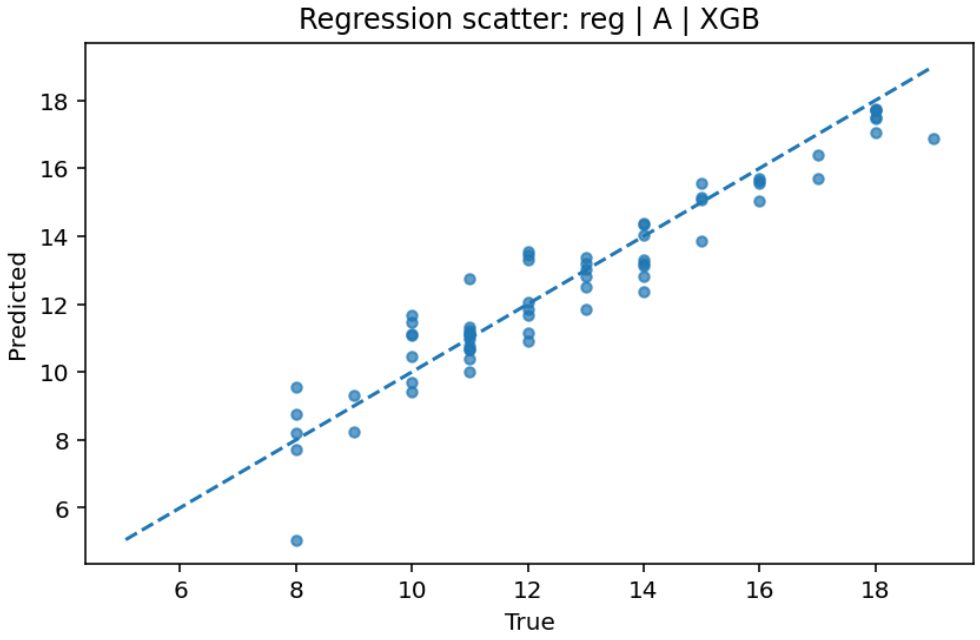


Figure A.10: Additional figure 31

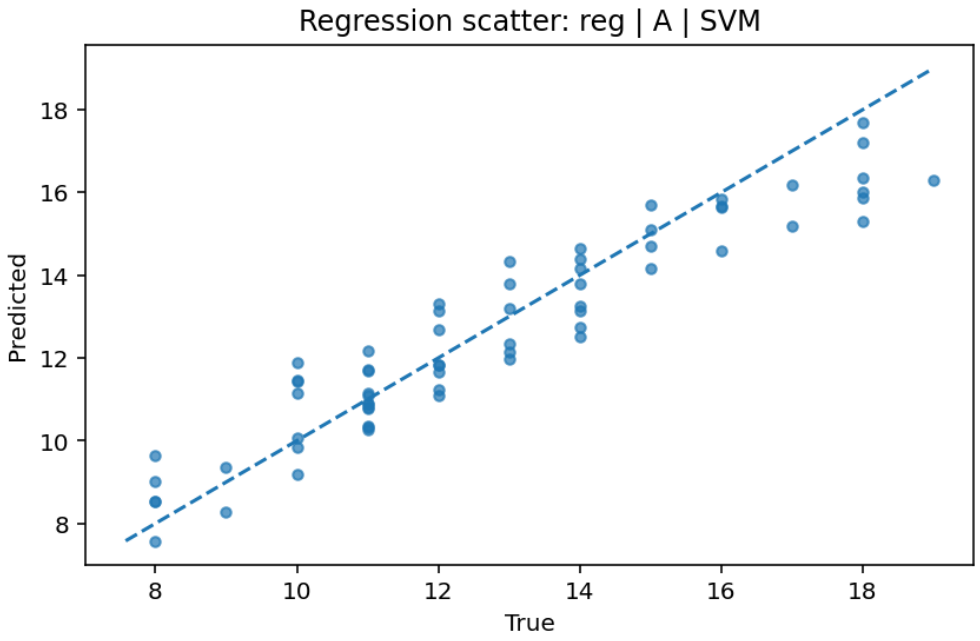


Figure A.11: Additional figure 32

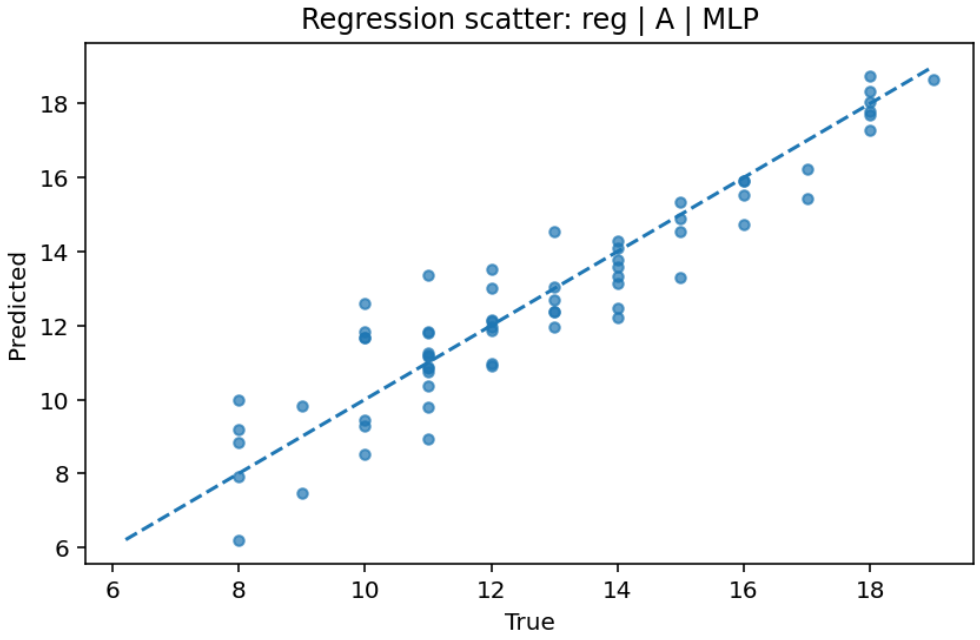


Figure A.12: Additional figure 33

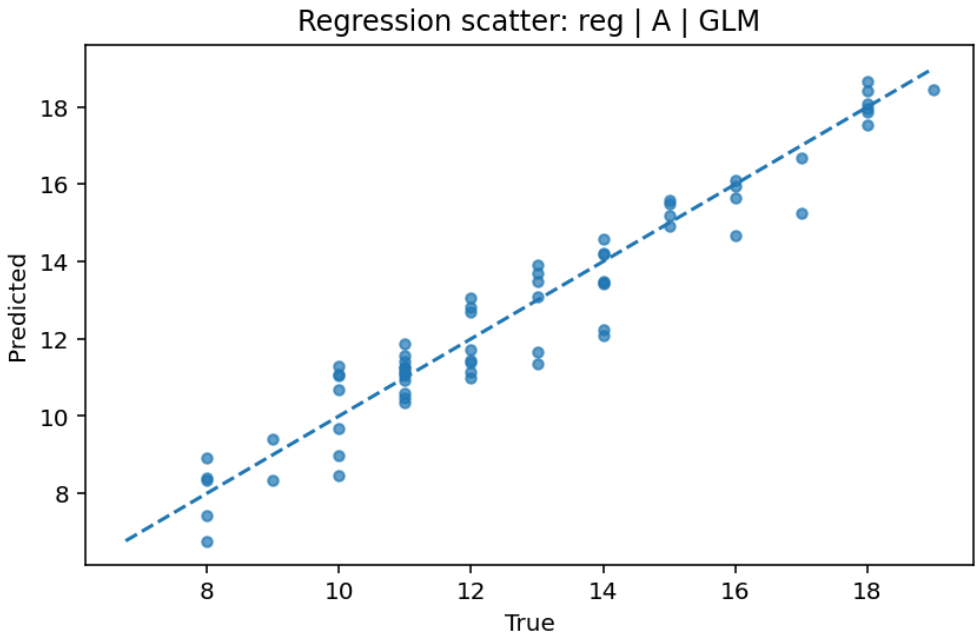


Figure A.13: Additional figure 34

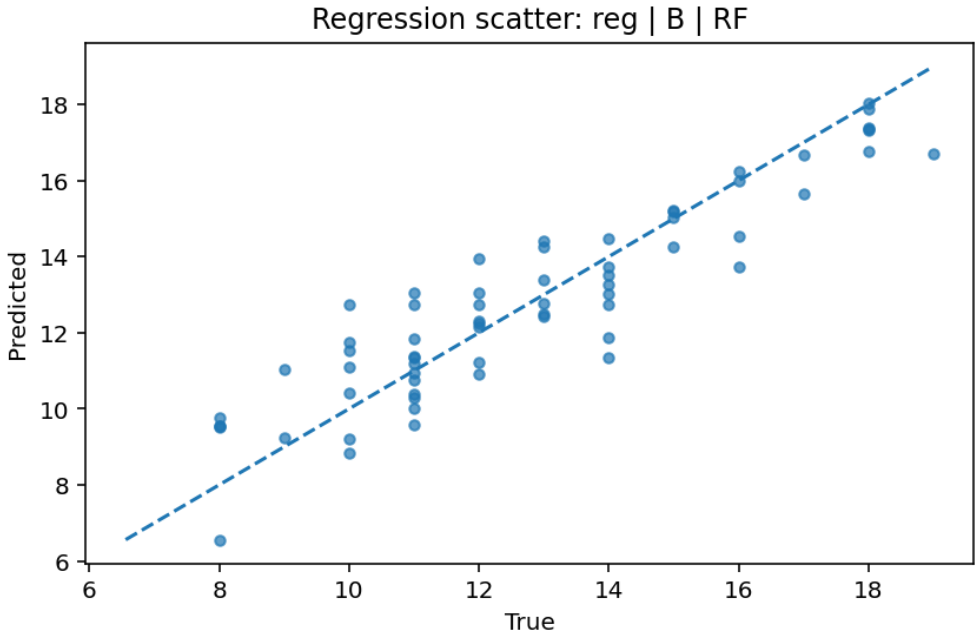


Figure A.14: Additional figure 35

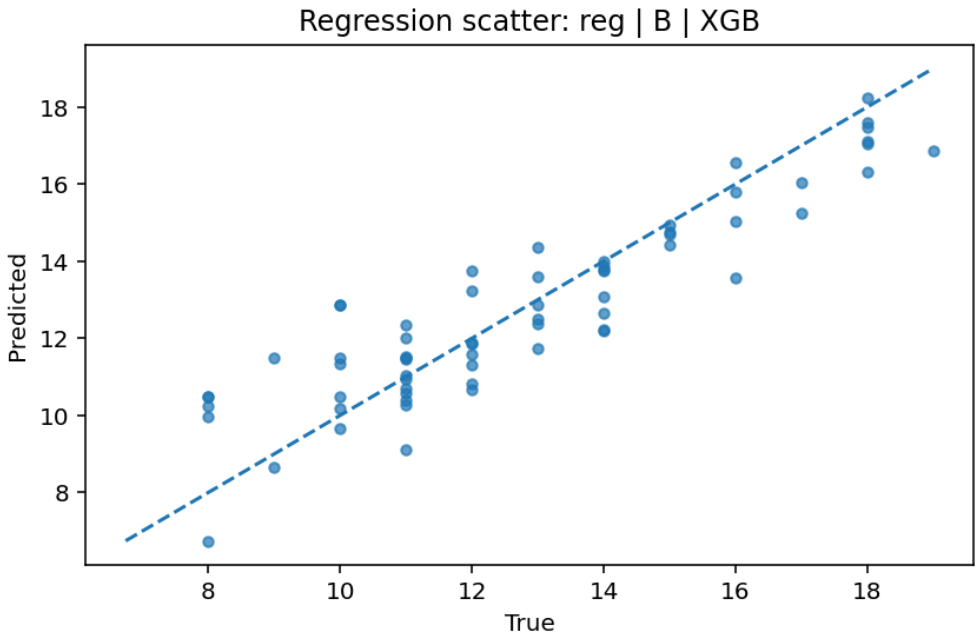


Figure A.15: Additional figure 36

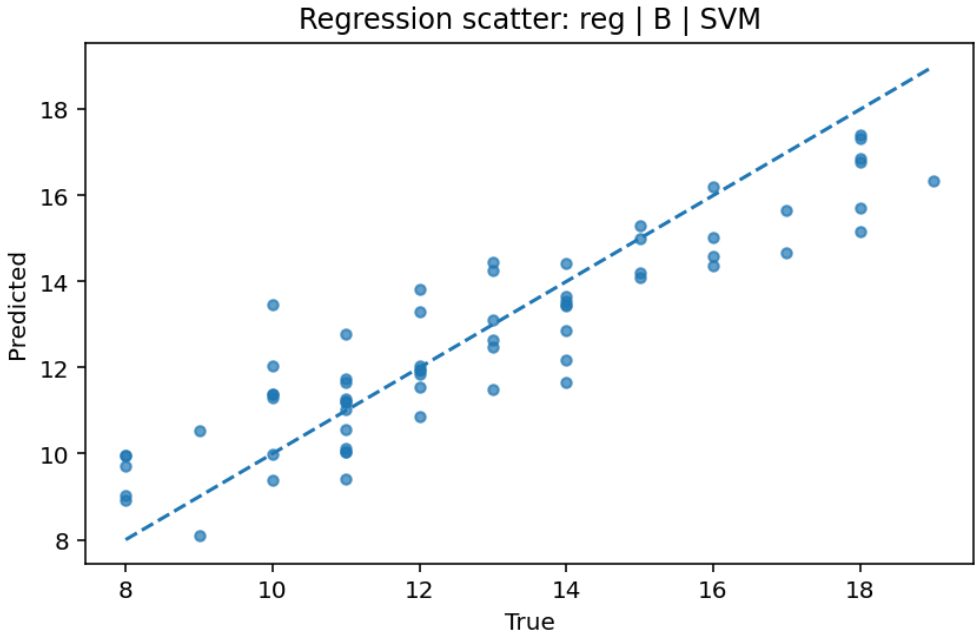


Figure A.16: Additional figure 37

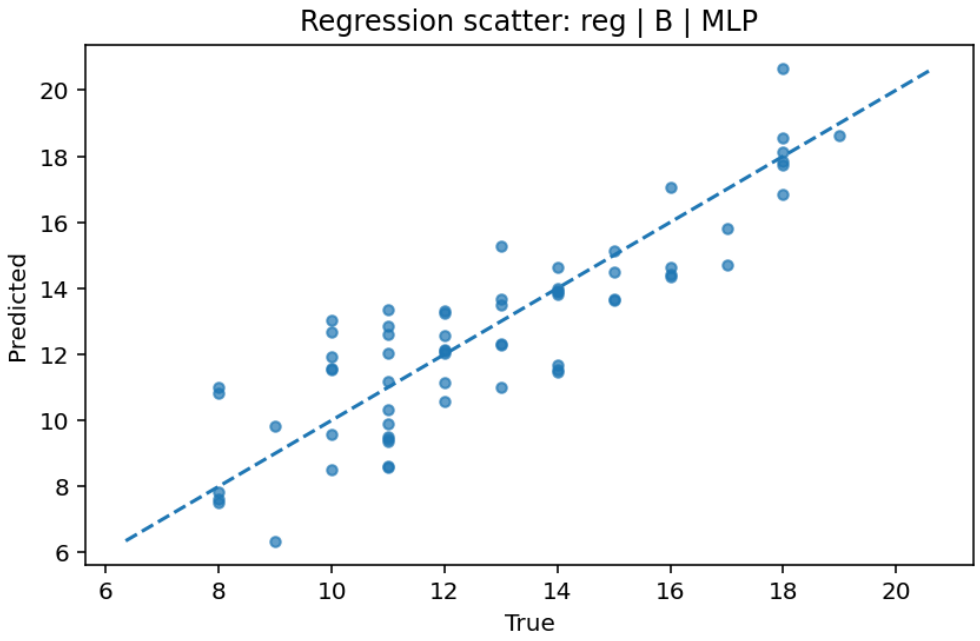


Figure A.17: Additional figure 38

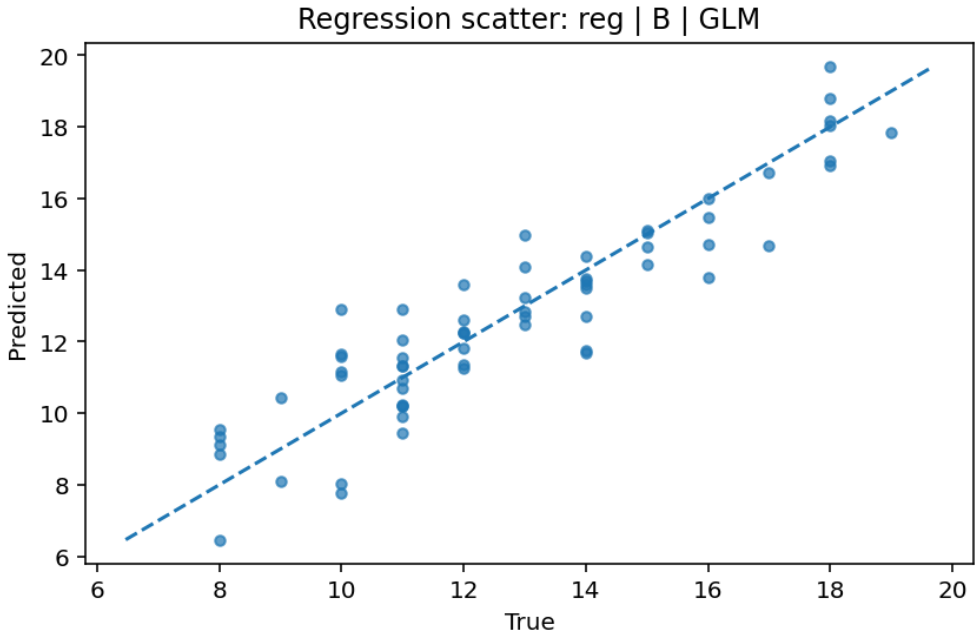


Figure A.18: Additional figure 39

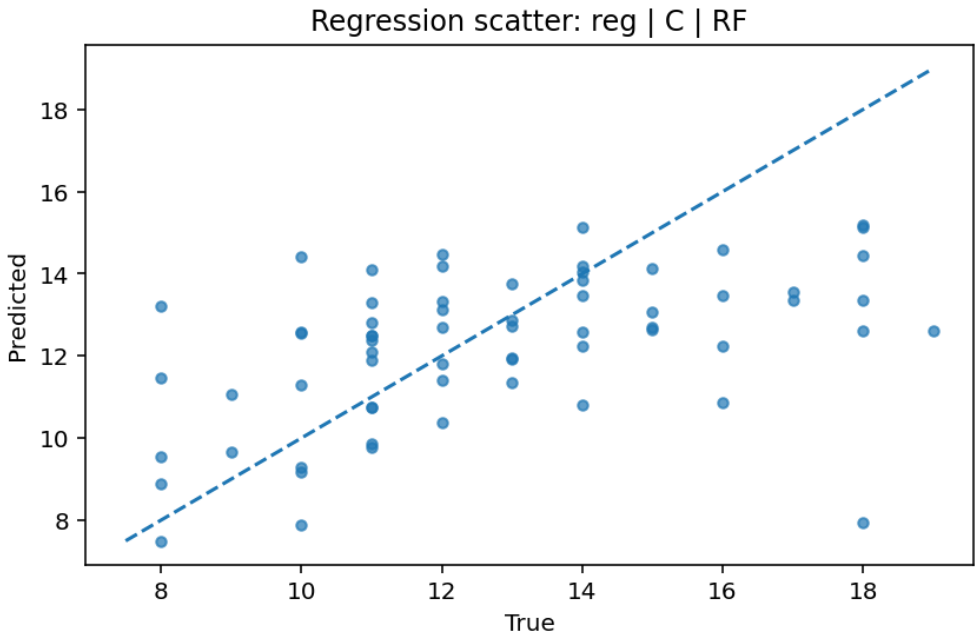


Figure A.19: Additional figure 40

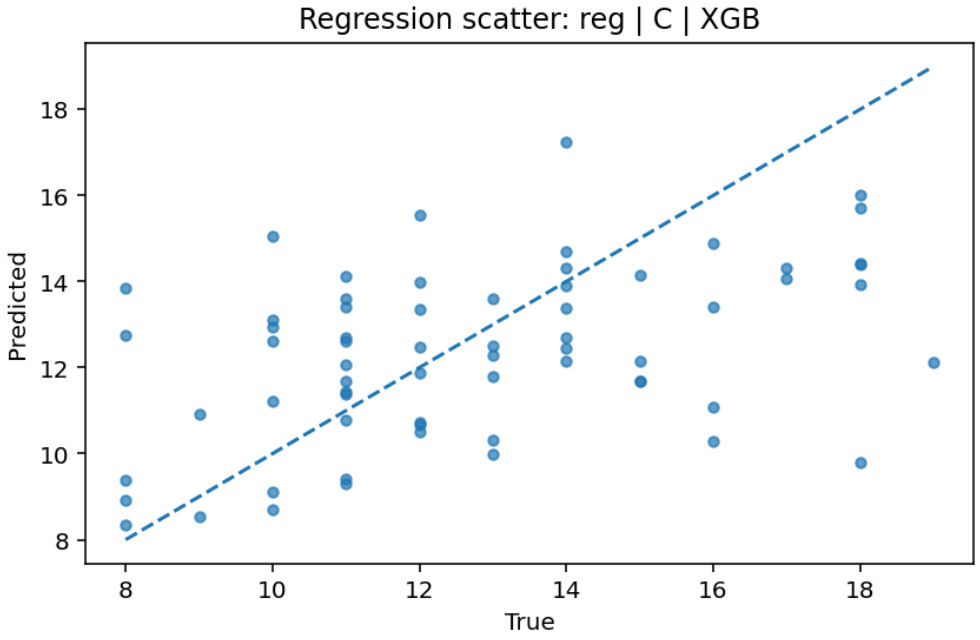


Figure A.20: Additional figure 41

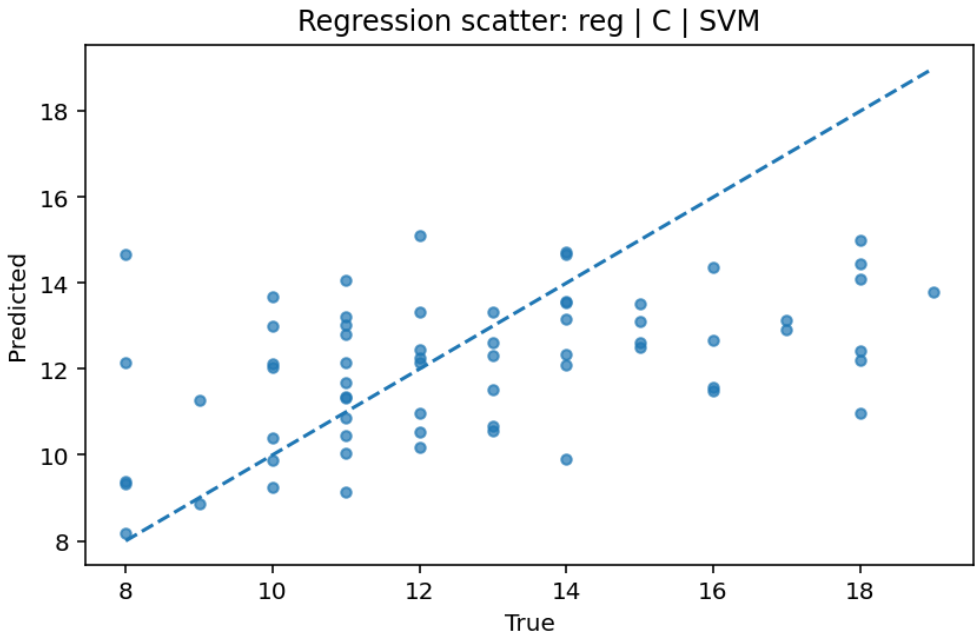


Figure A.21: Additional figure 42

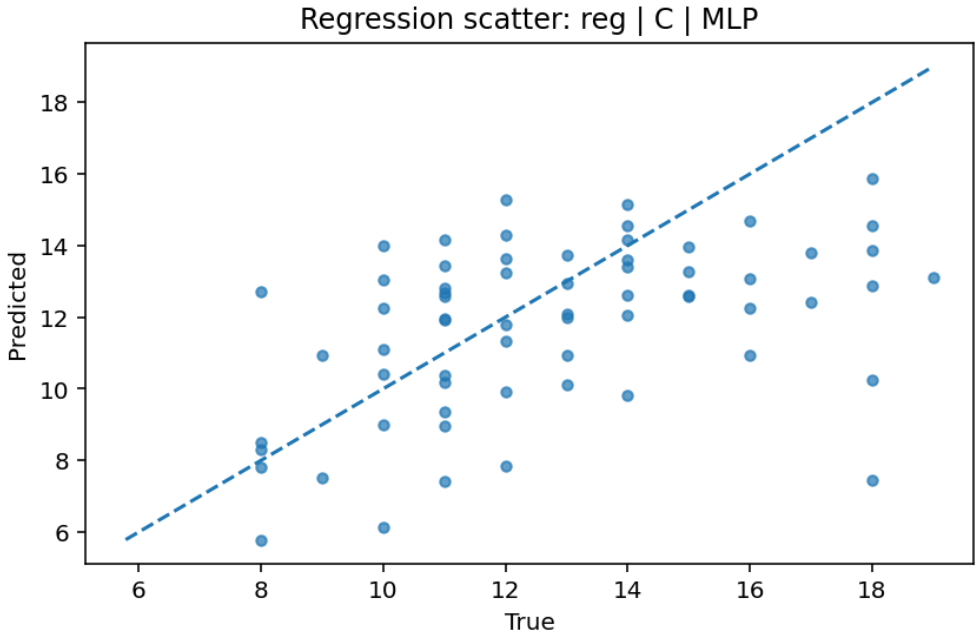


Figure A.22: Additional figure 43

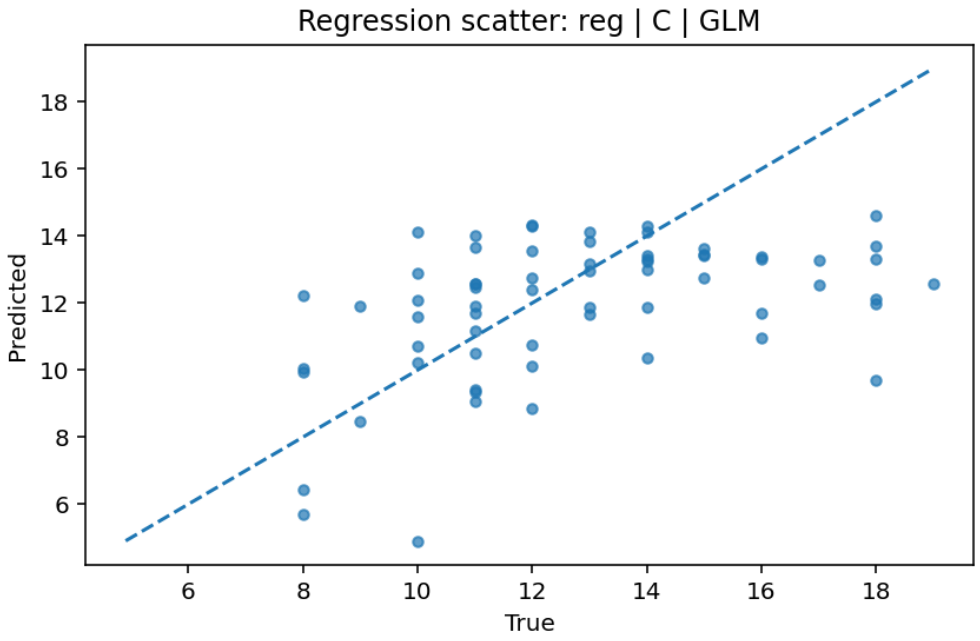


Figure A.23: Additional figure 44

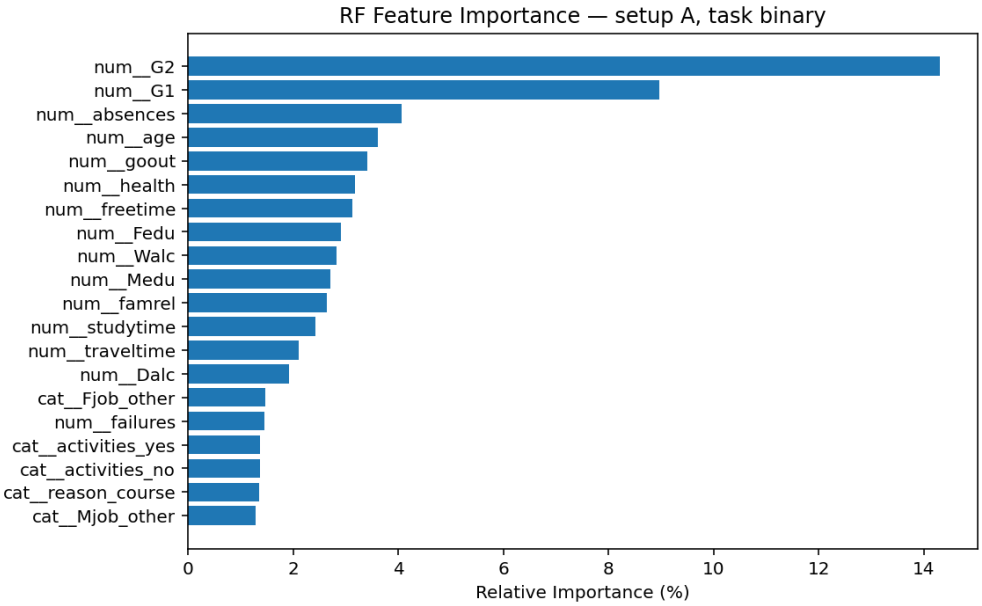


Figure A.24: Additional figure 45

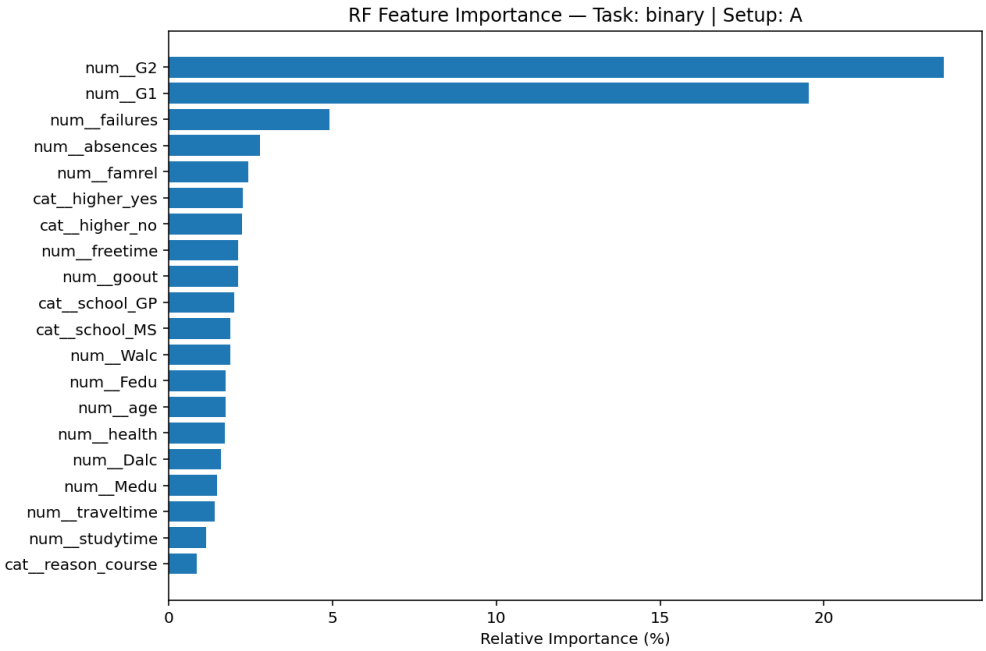


Figure A.25: Additional figure 46

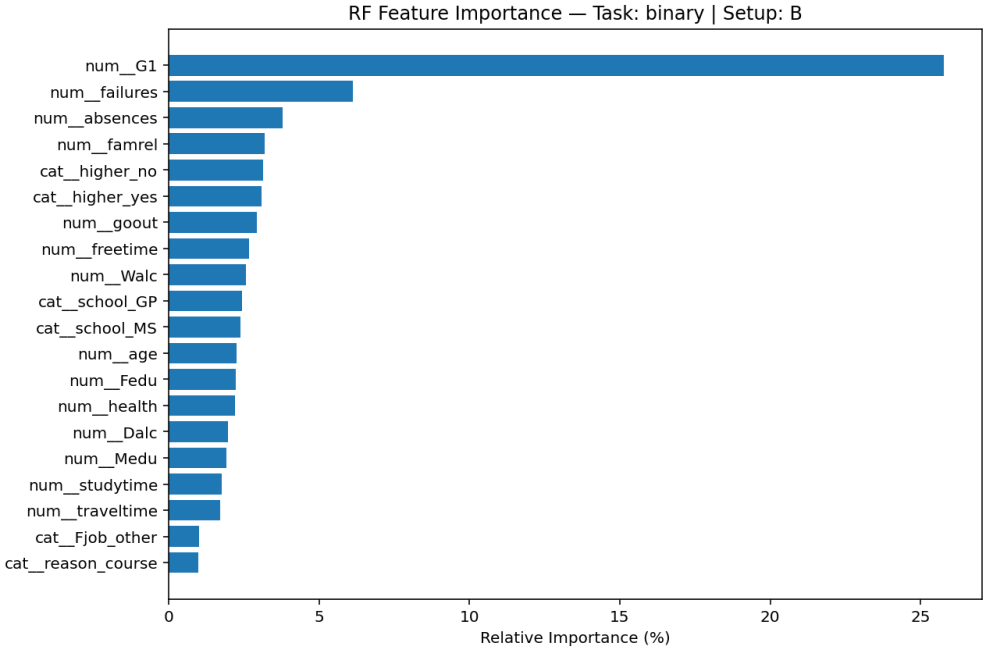


Figure A.26: Additional figure 47

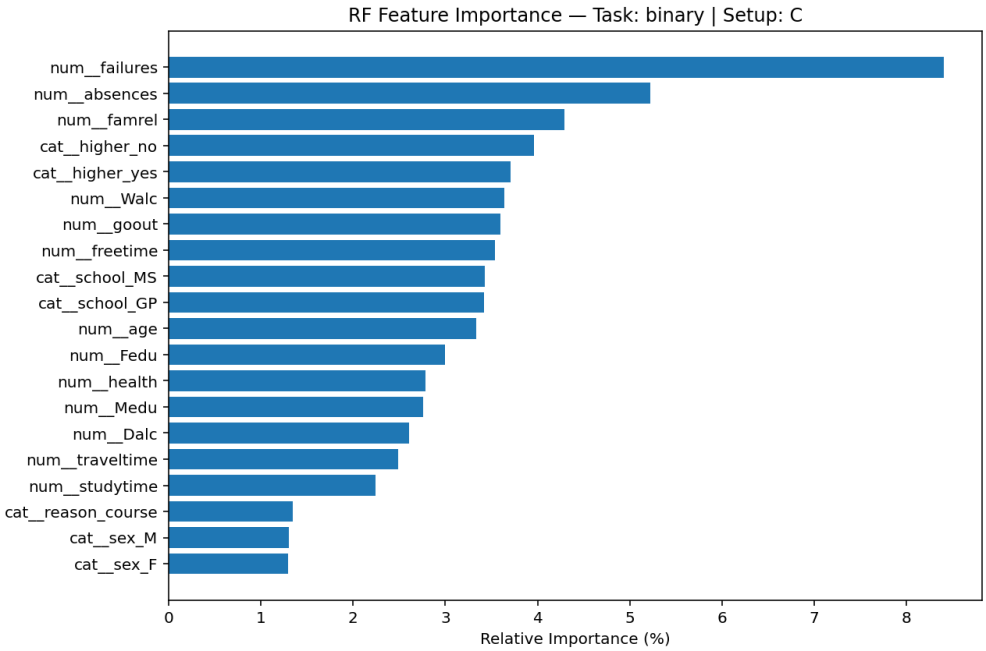


Figure A.27: Additional figure 48

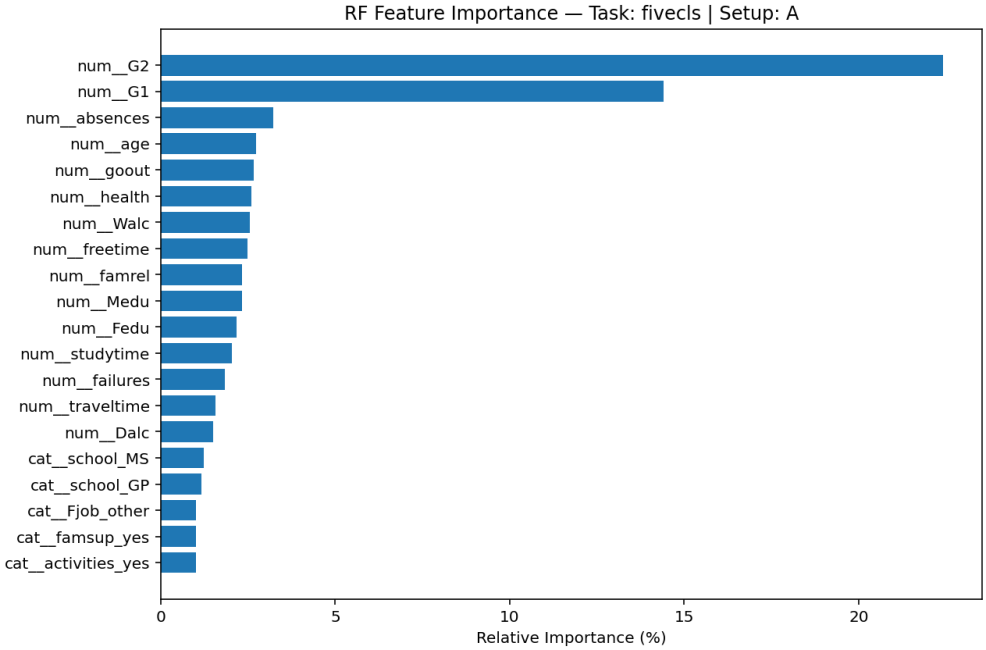


Figure A.28: Additional figure 49

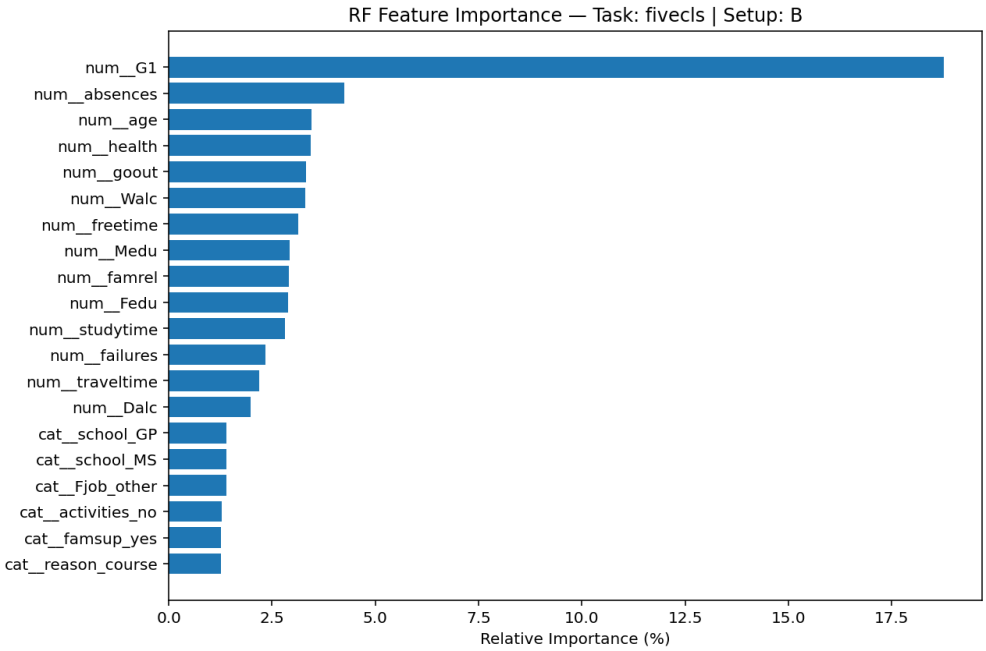


Figure A.29: Additional figure 50

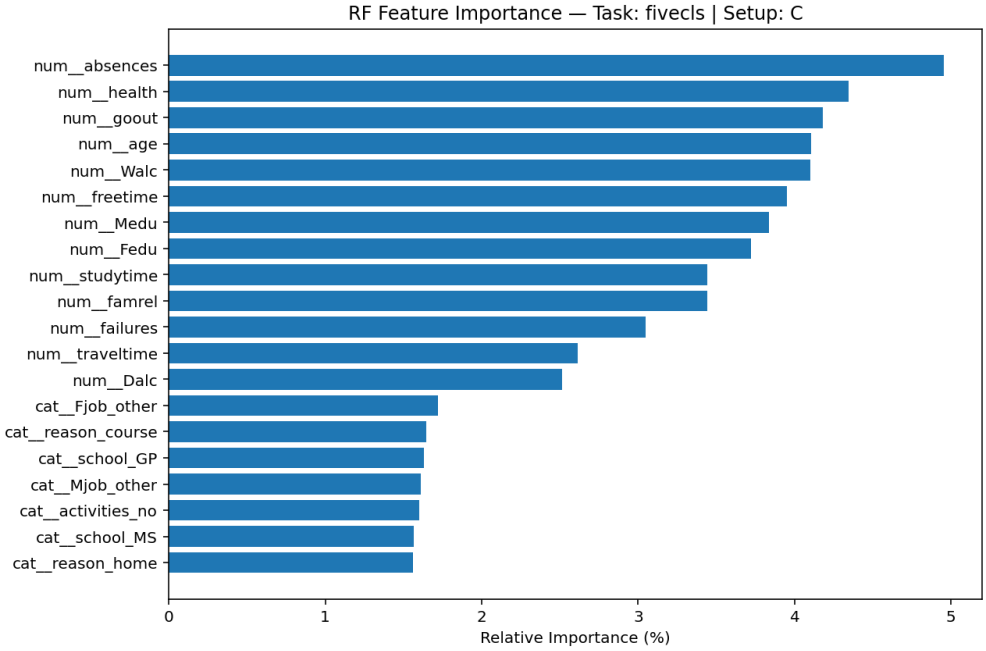


Figure A.30: Additional figure 51

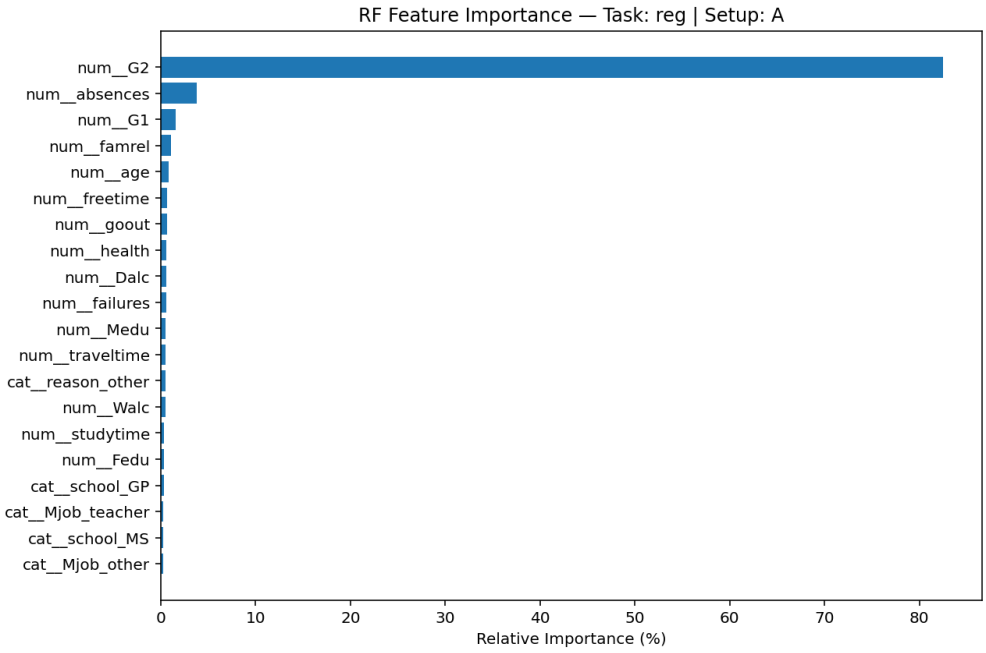


Figure A.31: Additional figure 52

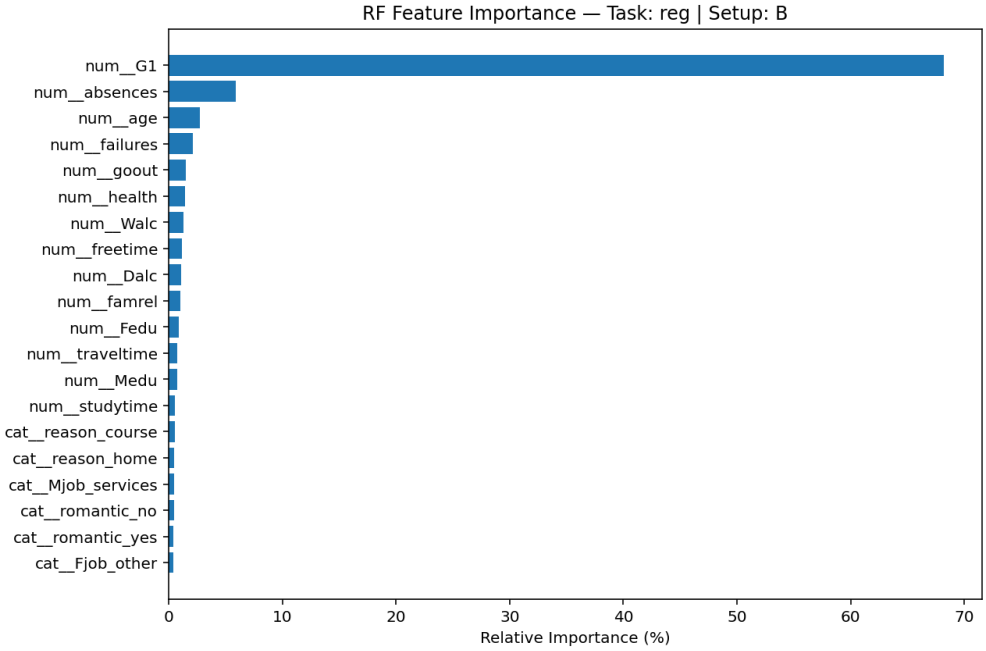


Figure A.32: Additional figure 53

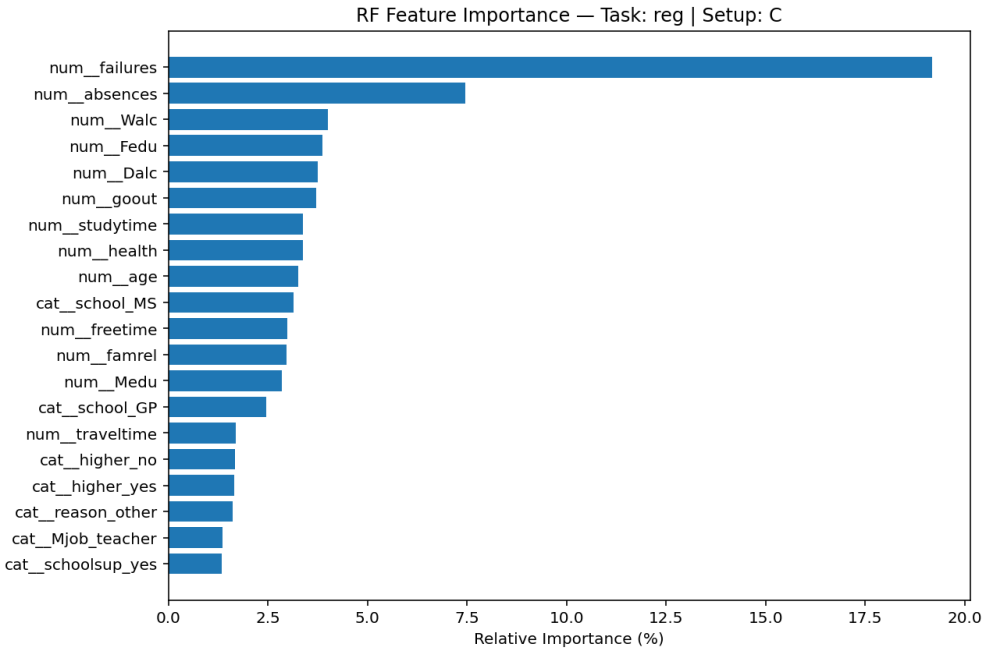


Figure A.33: Additional figure 54