STAT 385 Final Project

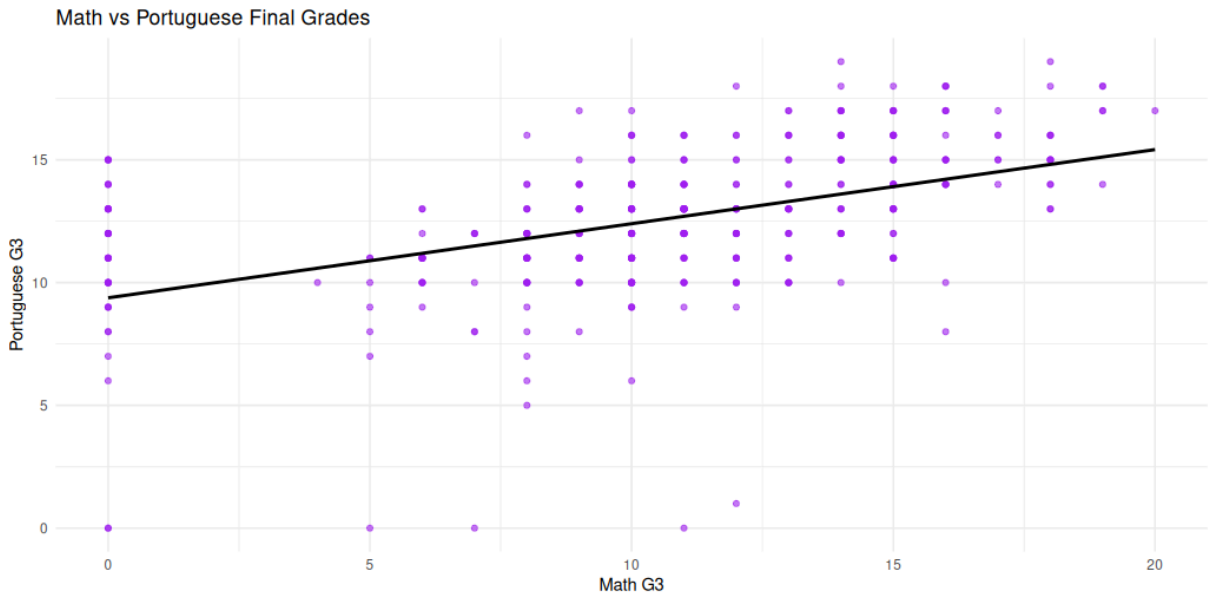
Larry Barrett, Daniel Mroz, Al Pakrosnis

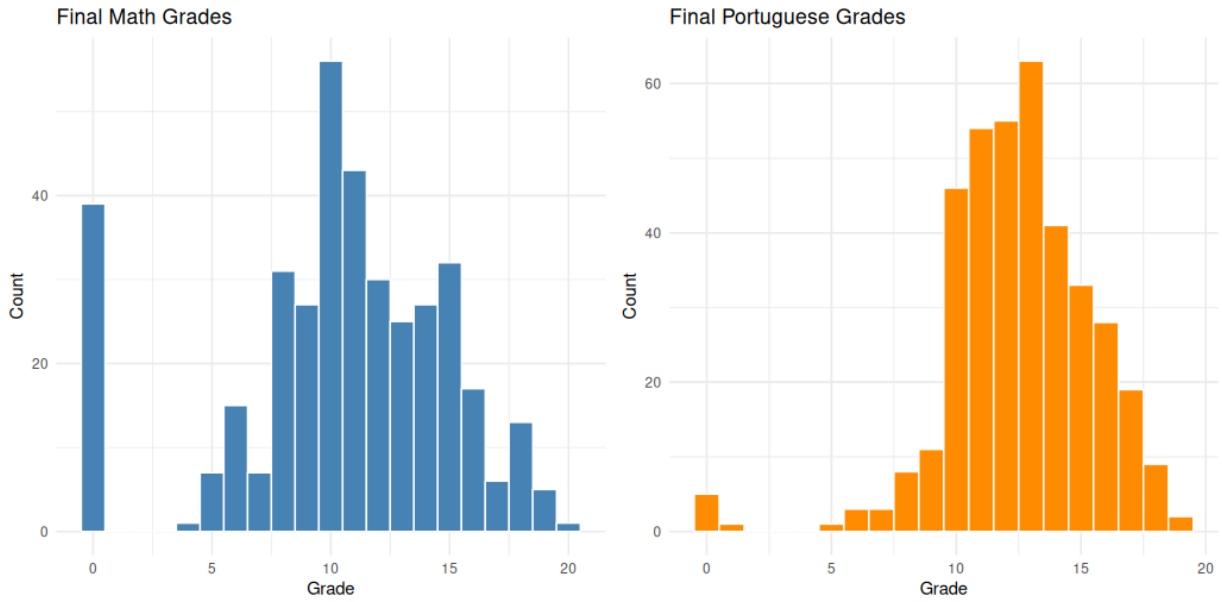
May 5, 2025

1. Question 1: How do math and Portuguese differ in their explanatory variables?

Exploratory Data Analysis:

```
> # Correlation
> cor_val <- cor(data$G3_math, data$G3_port)
> cor_val
[1] 0.4803494
```





Modeling Strategy:

I. Linear regression

A. For the coalescence of our first strategy we decided to create a linear regression manually selecting what variables to select for based on their significance after running a regression on ALL variables. (p-values below .05).

B. We decided to also exclude G1 and G2 variables to help pronounce other possible variables in explaining G3 grades.

1. We ultimately selected schoolsup_math, famrel_math, and absences_math for our math data model, as these were the significant variables shown to us on a first run of the linear regression.
2. We selected Fjob, school, reason, and failures_math for our Portuguese model, for which we selected them the same way we selected those for math.

C. We used a 75-25 data train/test split.

D. We saw better performance on Portuguese than math using our model, which lines up with other analyses on this data.

E. An easy improvement for this would be to use stepwise rather than this manual selection process.

II. Ridge regression

A. Used a 75-25 data split for training and testing

B. Target variable: G3_math & G3_portuguese

C. Predicted variables: all available variables used (besides G1 and G2 as mentioned above so not to diminish other variables)

D. Results: Measured by MSE:

1. Portuguese: 2.80
2. Math: 4.11

E. Our goal was to see what was the difference between significant predictors of math and Portuguese and these were the results. We compared the signs for the coefficients to determine which variables have a direct and inversely related:
    1. Variables with Strong Correlation for both models:
        a) G1 &G2
        b) Family relationships
        c) Higher education
        d) Workday Alcohol Consumption
        e) Failures
    2. Differences across both models:
        a) Internet (math positive, portuguese negative)
        b) School support (portuguese positive, math negative)
        c) Extracurricular activities (math positive, portuguese negative)
        d) Romantic relationship (math positive, portuguese negative)

Modeling Outcomes & Comparison:

For this section of our total analysis of the data we compared the efficacy of our ml regression as compared to a ridge regression, wherein the first variables were selected in a "common-sense" sort of matter with dampening being binary (a variable was either included or not). Our second model instead had feature and feature strength selection done via ridge, or "non-manually."

To evaluate performance and to compare the differences between the models we used MSE and Num predictions. Here were our results:

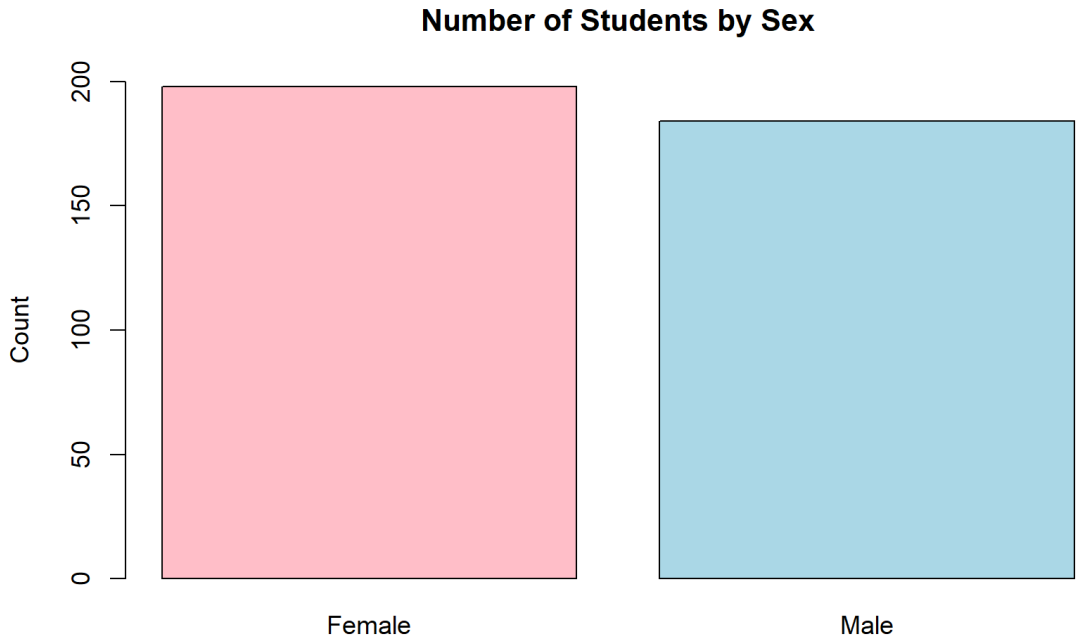
Regression style	MSE Math	MSE: Portuguese	Num preds
Multiple	22.4	8.2	M: 3, P: 4
Ridge	4.11	2.80	42

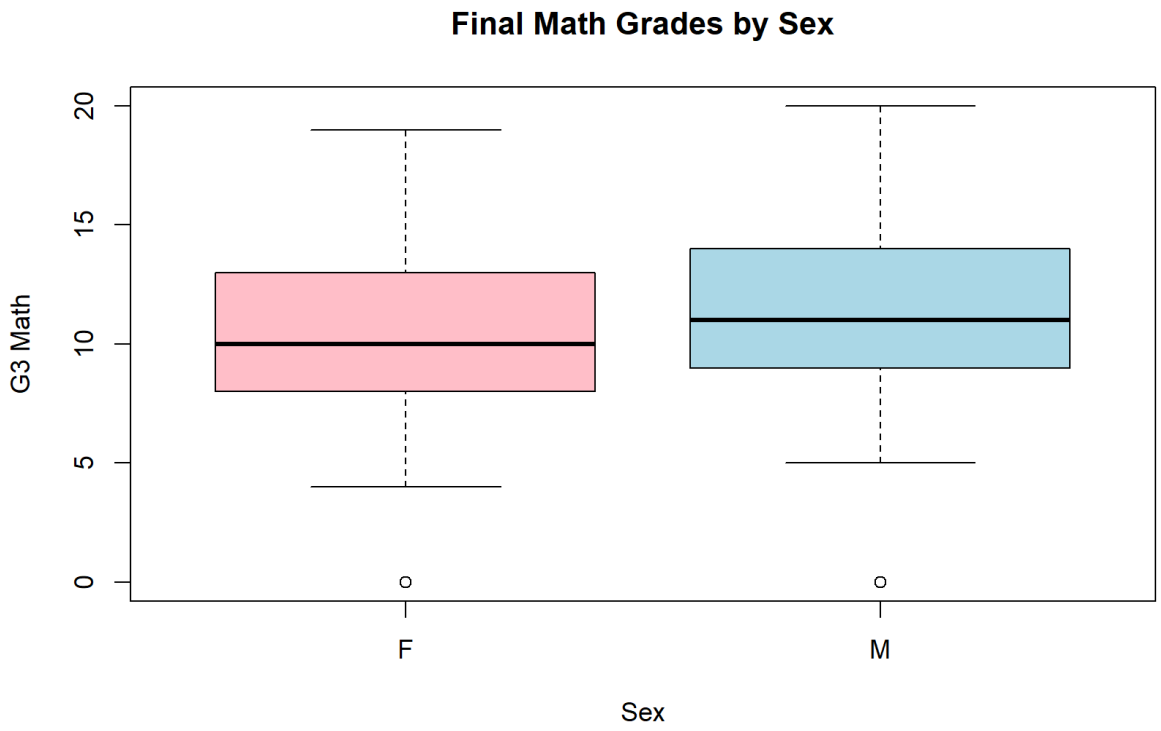
Quite clearly, the ridge regression vastly outperformed the manual feature selection multiple regression model. And although the multiple can be considered less "fitted" since it has 1/10 as many, however, this is not fully accurate as the ridge regression uses "dampening" to reduce noise from less significant features, but it doesn't actually get rid of any. So solely based on metrics ridge does outperform multiple and for further analysis we would choose to use ridge as our champion model, while also wanting to explore other models and compare them.

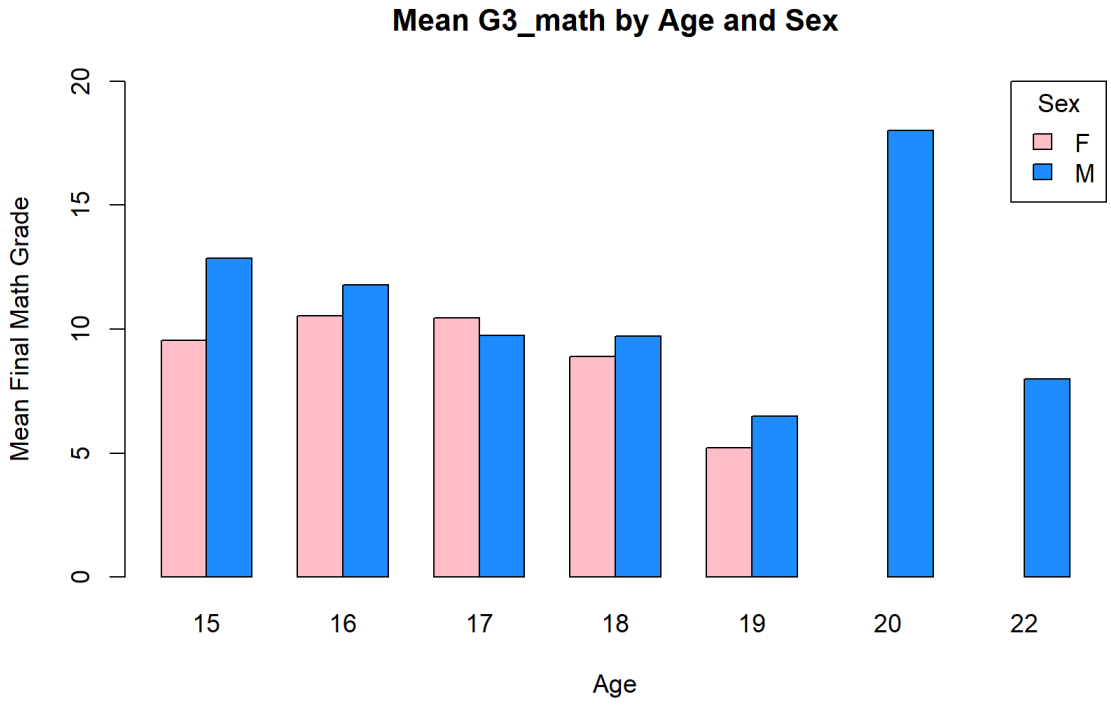
The most significant outcome from this is learning about the importance of using the advanced statistical techniques we learn in this class for model building. Particularly how doing so yields significantly better results than just trying to bootleg model creation by naively selecting the best looking features like how we did in the multiple linear regression section of this question.

2. Question 2:What conditions do males vs females require to succeed in their math classes?

Exploratory Data Analysis:








```
> summary(male$G3_math)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    9.00   11.00   10.98   14.00   20.00 
> summary(female$G3_math)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	8.000	10.000	9.838	13.000	19.000

Modeling Strategy:

I. Multiple Linear Regression
   - Predicted value: G3_math (final grade for math)
   - Predictor variables: all available:
   - Result: measured by Mean Squared Error
     - Male: 3.42
     - Female: 7.83
   - Interpretation:
     - The multiple linear regression model was used just as a baseline for the other models we would implement. We can take away the fact that the model for females is harder to predict than for males both by MSE (7.83 > 3.42) and also adjusted R squared (0.8015 < 0.8512)

II. Stepwise Regression using AIC
   - Predicted value: G3_math (final grade for math)
   - Predictor variables: all available:
   - Result: measured by Mean Squared Error
     - Male: 2.23
     - Female: 6.80
   - Interpretation:
     - Male model ended up with 14 variables where the female model had 17.
     - We got a much better mean squared error from this model compared to the regular MLR model likely due to cutting variables that were arbitrary.
     - There are a few variables that both models had in common (previous quarter grades, famrel, absences) however variables like what we saw in our EDA like age were not significant across both models.

III. LASSO Regression with Cross Validation
   - Predicted value: G3_math (final grade for math)
   - Predictor variables: all available
   - Best lambda: this was determined using cross validation and selecting the minimum value for each model.
     - Male: 0.1415092
     - Female: 0.1131053
   - Result: measured by Mean Squared Error
     - Male: 1.44
     - Female: 5.19
   - Interpretation:
     - LASSO was the best performing model across all that were tested
     - Female model had 23 variables and the male model had 14 variables
     - There was a lot more agreement between the male and female models in terms of variables that they had in common.

        - The male model had significantly less variables likely because the coefficient for higher is very high (1.132659)

Summary

- LASSO model yielded the best results and gave us the best answer to our question.
- Predictors for good math grades that both males and females have in common
    - Age
    - Mom = teacher
    - nursery = yes
    - guardian = mother
    - famrel
    - absences
    - previous quarter grades
- Predictors for good math grades that are only significant for males:
    - Father job
    - paid classes
    - higher education
- Predictors for good math grades that are only significant for females:
    - reason=other
    - goes out with friends
    - health
    - romantic
    - travel time
    - school support

Potential improvements:

- Implementing more models or using an ensemble method to yield more results
- More data (collect more, bootstrap)

3. Question 3: Which students are most likely to need additional academic support, and what factors best predict this need?

1. Practical/Scientific Questions and Corresponding Solutions

- **Operational Definition:** A student "needs support" if they have at least one past class failure or have received extra school-provided educational support.
- **Goal:** Develop predictive models to classify students who need support and identify key predictors.

Statistical and Machine Learning Solutions:

- **Logistic Regression:** For interpretable, baseline classification.
- **Naive Bayes:** For a simple probabilistic approach.
- **Random Forest:** For robust, non-linear modeling and feature importance.

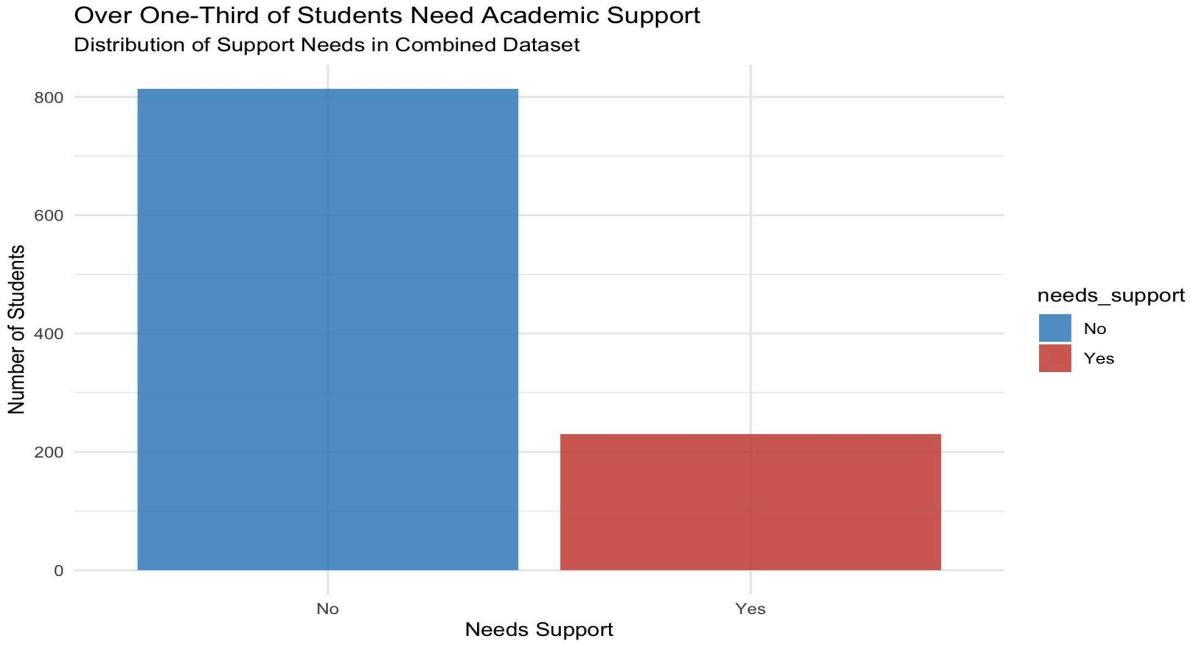
2. Preliminary Exploratory Data Analysis (EDA)

Data Preparation Steps:

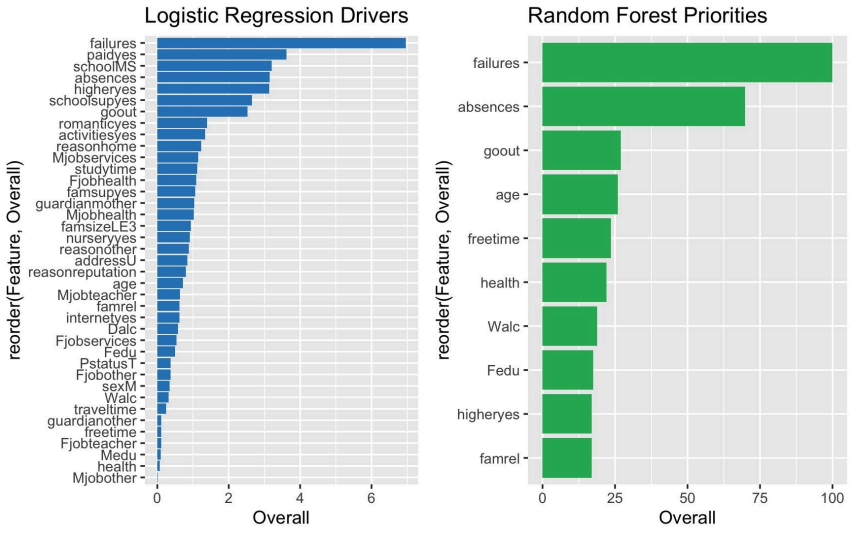
- Merged math and Portuguese datasets vertically.
- Created the binary outcome variable needs_support (1 if any failures or school support, 0 otherwise).
- Converted categorical variables to factors.
- Removed grade columns and identifiers for modeling.

Summary Statistics & Insights:

- **Distribution:** About 27% of students need support (based on the training/test split).



- Key Predictors (from Random Forest importance):
    - failures (by far the most important)
    - age, absences, Medu (mother's education), goout, freetime, health, studytime, Walc, Fedu



- **Class Imbalance:** Moderately imbalanced, but not severe.

Suggested EDA Visualizations:

- Bar plots of support need by categorical features (e.g., school, sex, address)
- Boxplots of numeric features (e.g., absences, age) by support need
- Correlation matrix for numeric predictors

These analyses guided the model choice:

- **Logistic regression** for interpretability
- **Random forest** for non-linear effects and feature ranking
- **Naive Bayes** as a simple baseline

3. Modeling Details

A. Logistic Regression

Model Specification:

$$\log\left(\frac{P(\text{needs\_support}=1)}{1-P(\text{needs\_support}=1)}\right)=\beta_0+\sum_{j=1}^{p}\beta_j X_j$$

Outcome: needs_support

- **Predictors:** All other variables except grades and identifiers (with one-hot encoding)

Variable/Model Selection:

- All predictors included initially
- Model selection via 5-fold cross-validation

Tuning:

- No hyperparameters for standard logistic regression (regularization optional)

Results:

- **Accuracy:** 92.8%
- **Recall for support class:** 73%

Interpretation:

- High specificity (no false positives), but some false negatives
- Coefficient estimates (not shown) identify factors increasing odds of needing support

Conclusion/Improvements:

- Interpretable but may underfit non-linearities
- Can improve recall with threshold tuning or regularization

B. Naive Bayes

Model Specification:

- **Outcome:** needs_support
- **Predictors:** All available (categorical as factors)
- **Assumption:** Conditional independence among predictors

Variable/Model Selection:

- All predictors included

Tuning:

- No hyperparameters for standard Gaussian Naive Bayes

Results:

- **Accuracy:** 90.9%
- **Recall for support class:** 73%

Interpretation:

- Similar to logistic regression, but with slightly more false positives
- Useful baseline model

Conclusion/Improvements:

- Fast and simple, but limited by independence assumption
- Not ideal when high accuracy or complex feature interactions are needed

C. Random Forest

Model Specification:

- **Outcome:** needs_support
- **Predictors:** All available (categorical variables one-hot encoded)
- **Model:** Ensemble of decision trees with bootstrapping and random feature selection

Variable/Model Selection:

- All predictors included
- Hyperparameter tuning via nested 5-fold cross-validation (tuneLength = 3)

Tuning:

- Number of trees and number of features per split tuned via cross-validation

Results:

- **Accuracy:** 97.1%
- **Recall for support class:** 89%

Feature Importance (Top Predictors):

1. failures (most important)
2. age, absences, Medu, goout, freetime, health, studytime, Walc, Fedu

Interpretation:

- Captures non-linearities and interactions well
- Best overall performance with very few false negatives

Conclusion/Improvements:

- Robust and accurate
- Less interpretable than logistic regression
- Could explore more advanced ensemble methods

4. Method Comparison and Recommendations

Model	Accuracy	Recall (Sensitivity)	Specificity	Main Strengths	Main Weaknesses
Logistic Regression	92.8%	73%	100%	Interpretability, baseline	Lower recall for positives
Naive Bayes	90.9%	73%	97%	Simplicity, speed	Independence assumption
Random Forest	97.1%	89%	100%	Best accuracy, robust	Less interpretable

Best Performing Model:

- **Random Forest** – due to its high accuracy, sensitivity, and ability to rank feature importance

Recommended Use:

- **Random Forest:** For deployment and identifying at-risk students
- **Logistic Regression:** When interpretability is crucial (e.g., for communicating with stakeholders)

Reproducibility

- All models were trained/tested on the same data splits (random_state = 123 or set.seed(123))
- Ensures fair comparison and reproducibility

Summary

- The main question—identifying students who need academic support—was addressed using logistic regression, naive Bayes, and random forest classifiers.
- EDA and feature importance analyses identified academic failures, age, and absenteeism as key predictors.
- Random forest achieved the best performance (97% accuracy, high recall), making it the top recommendation.
- Logistic regression is still valuable for transparency and explanation to non-technical audiences.
- All results are reproducible.

Potential Improvements:

- Explore other models (e.g., gradient boosting, SVM)
- Address class imbalance if it worsens in future data
- Evaluate the effect of dropping or merging less informative features

---

We state that each person in our group did an equivalent amount of work and that no one person unfairly contributed.