

Victoria Mendez
Hannah Miles-Kingrey
Ethan Shipman
Nathaniel Trief



Early Warning System for At-Risk Students: Leveraging Data to Reduce Drop-Out Rates



Project Proposal

- ★ Our goal is to create an **Early Warning System for At-Risk Students**.
- ★ The education dataset, collected via school reports and questionnaires, consists of over 600 individual data instances from two Portuguese classes including attributes such as
 - demographics
 - grades
 - social factors
 - support
- ★ We're hoping to identify problem areas within students' environments and potentially reduce the total drop-out rate.



Key Questions: Potential Risk Factors

- extra educational support
 - parent education level
 - health
 - social tendencies



Methodology



Data Model Implementation

- cleaned two CSVs (math class data, Portuguese class data)
- SQLite database creation

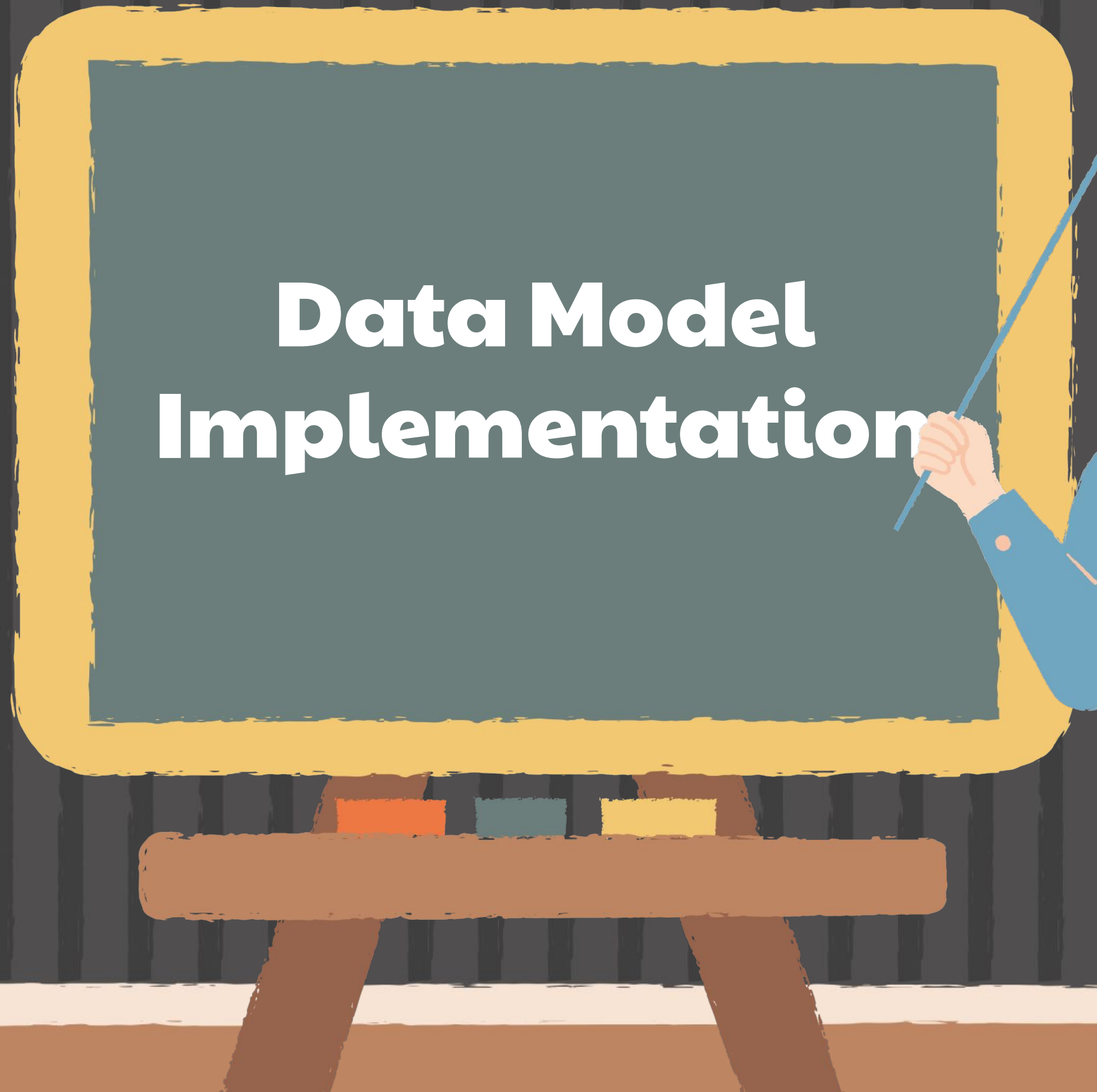
Data Model Optimization

- feature engineering
- multilabel classification
 - modifying target
- experimenting with model type

Prediction & Visualization



Data Model Implementation



Implementation Process

- Minimal cleaning
- Feature Engineering
 - Sudden Grade Drop (from G1 to G2)
 - Attendance Engagement (Balances absences with grades)
 - Low Engagement Flag (Flags students whose engagements levels are in the highest quartile)
- Model Types
 - Logistic Regression - Baseline
 - Decision Tree - Baseline
- Predictions



Classification Report for Baseline

Baseline Decision Tree Model Evaluation:
Accuracy: 0.8038277511961722
Classification Report:

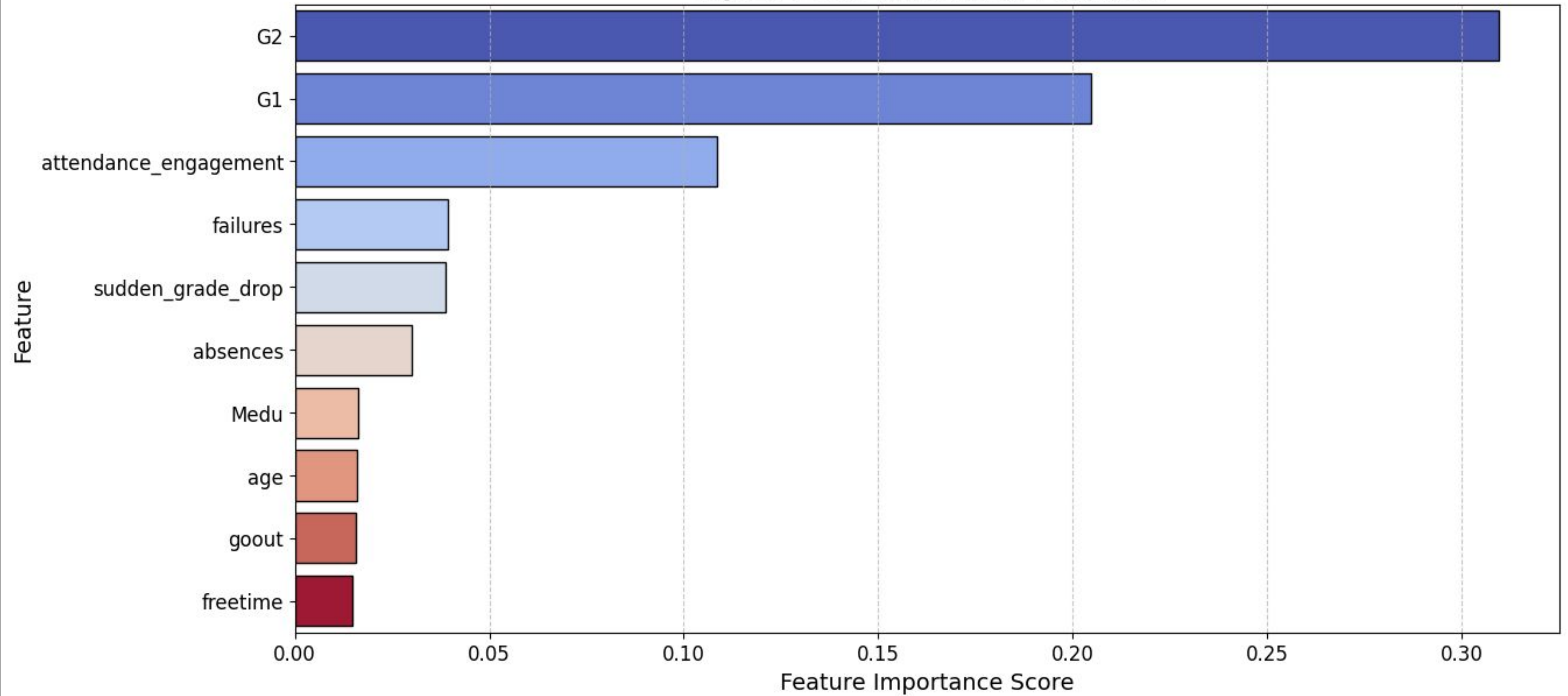
	precision	recall	f1-score	support
0	0.71	0.78	0.74	46
1	0.84	0.82	0.83	122
2	0.82	0.78	0.80	41
accuracy			0.80	209
macro avg	0.79	0.79	0.79	209
weighted avg	0.81	0.80	0.80	209

Baseline Logistic Regression Model Evaluation:
Accuracy: 0.84688995215311
Classification Report:

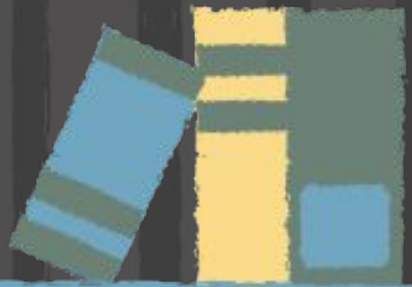
	precision	recall	f1-score	support
0	0.74	0.87	0.80	46
1	0.90	0.83	0.86	122
2	0.84	0.88	0.86	41
accuracy			0.85	209
macro avg	0.83	0.86	0.84	209
weighted avg	0.85	0.85	0.85	209



Top 10 Most Influential Features



Data Model Optimization



Optimization Process

- **Feature Engineering**
 - Extra Support
 - Health
 - Parent Education
 - Social Tendencies
- **Experimenting with different targets**
 - G3 (final grade, 0-20 scale)
 - creating a risk category (multiclass labels)
- **Experimenting with Model Types**
 - Logistic Regression
 - ✓ Highest accuracy (99%)
 - Decision Tree
 - ✓ Highest accuracy (100%)
 - Random Forest
 - Pruned Forest
 - ✓ Highest accuracy (85.65%)



Optimizations 1-3 focused on engineering features, yielding 90%+ accuracy.

Optimization 4 focused on modifying classification labels from three to **four risk categories** and making this the **new target**.



0 - 'at risk'

1 - 'low risk'

2 - 'moderate risk'

3 - 'not at risk'

Logistic Regression Model Evaluation:

Accuracy: 0.9904306220095693

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	46
1	0.97	1.00	0.98	61
2	1.00	1.00	1.00	61
3	1.00	1.00	1.00	41
accuracy			0.99	209
macro avg	0.99	0.99	0.99	209
weighted avg	0.99	0.99	0.99	209

- performs well at classifying 'moderate risk' and 'not at risk' students
- missed some 'at risk' students leading to false-positives for 'low risk' students.

Decision Tree Model Evaluation:

Accuracy: 1.0

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	61
2	1.00	1.00	1.00	61
3	1.00	1.00	1.00	41
accuracy			1.00	209
macro avg	1.00	1.00	1.00	209
weighted avg	1.00	1.00	1.00	209

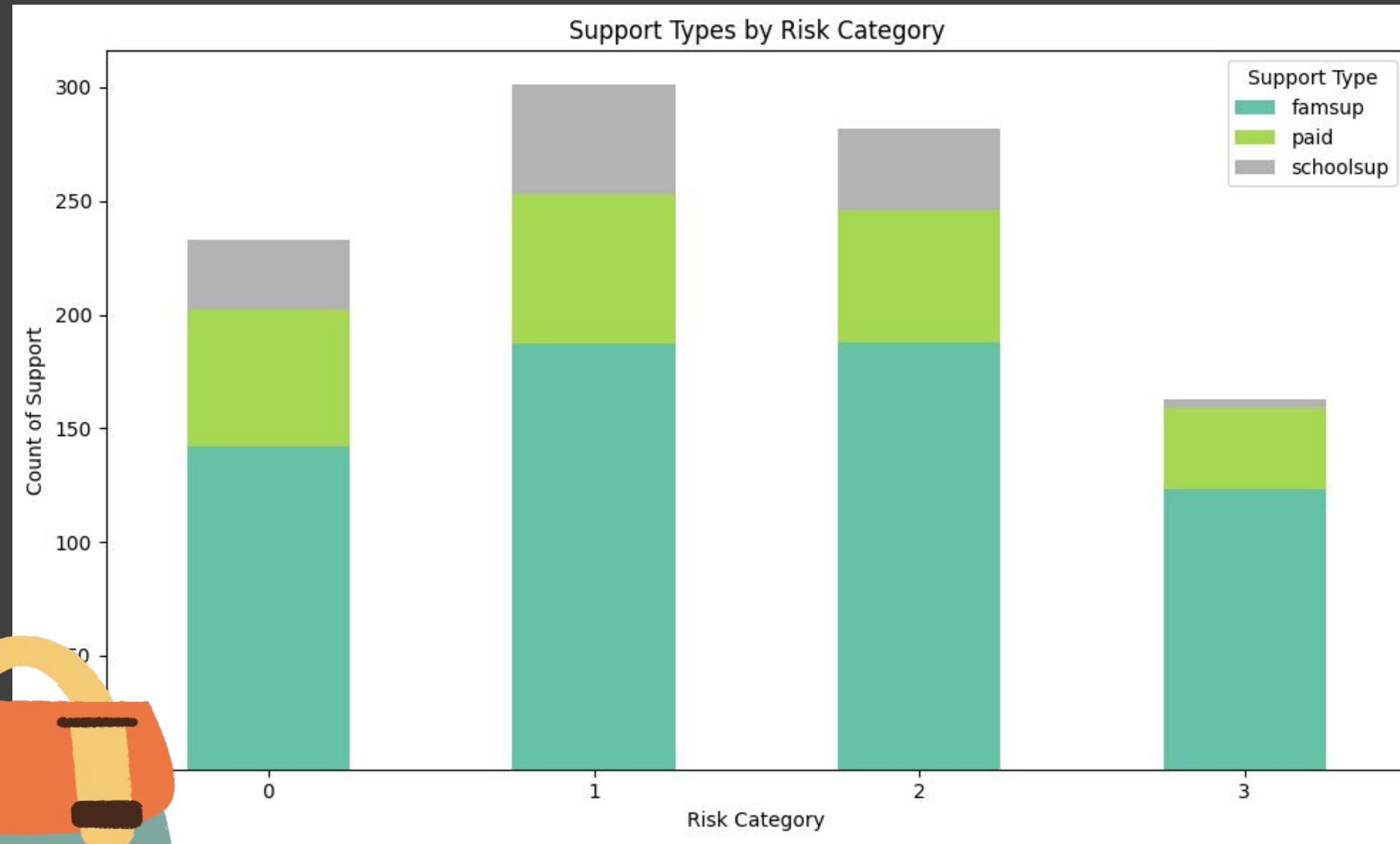
The 100% accuracy here may reflect a complex or non-linear dataset.

Extra support features didn't make a significant difference.

Extra Supports



Is there a relationship between a student being at-risk and access to extra educational support, family educational support, or extra paid classes within the course subject? *Hannah*



risk category

- 0 - 'at risk'
- 1 - 'low risk'
- 2 - 'moderate risk'
- 3 - 'not at risk'

Students identified as being '**at risk**' receive

- more total support than '**not at risk students**'
- less than '**moderate**' and '**low risk**' students.

Influence of Parental Education on Student Performance and Higher Education Interest

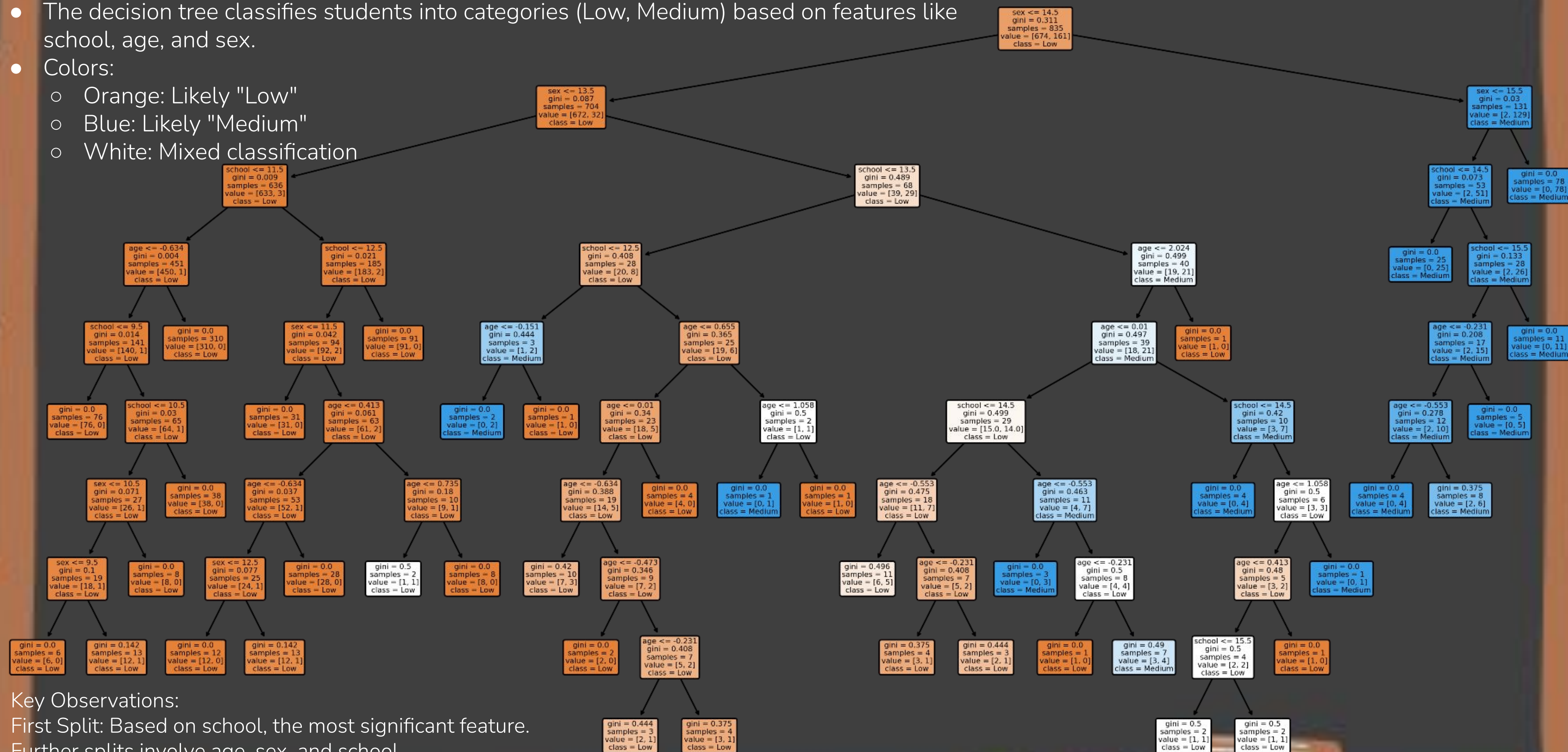


Decision Tree Model for Predicting Student Outcomes

Victoria

Visualization Overview:

- The decision tree classifies students into categories (Low, Medium) based on features like school, age, and sex.
- Colors:
 - Orange: Likely "Low"
 - Blue: Likely "Medium"
 - White: Mixed classification



Key Observations:

First Split: Based on school, the most significant feature.
Further splits involve age, sex, and school.
Nodes with a Gini index of 0 indicate perfect classification.

Do students with higher-educated parents perform better in final exams (G3)?

Victoria

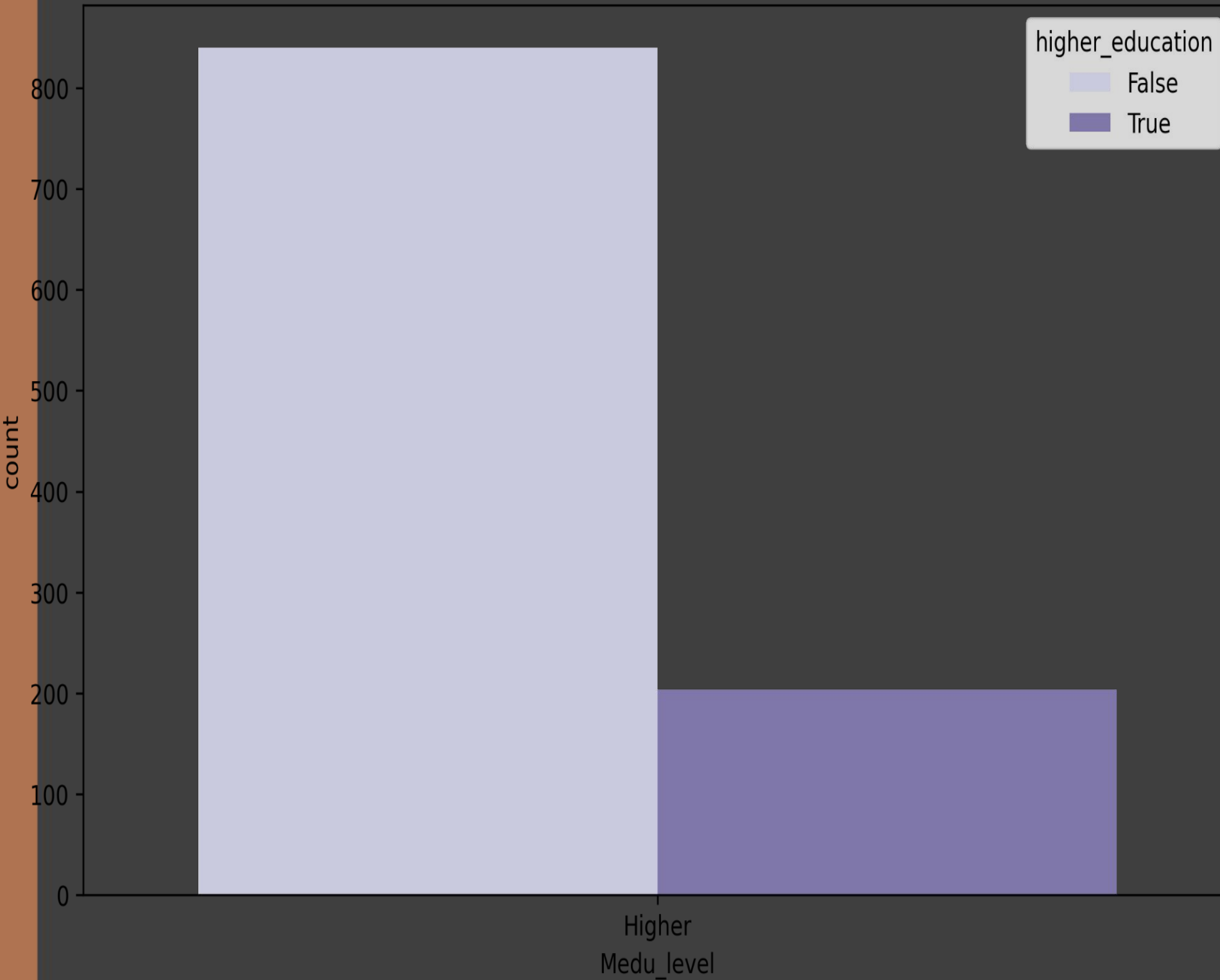
Do they show stronger interest in higher education?

Key Takeaways:

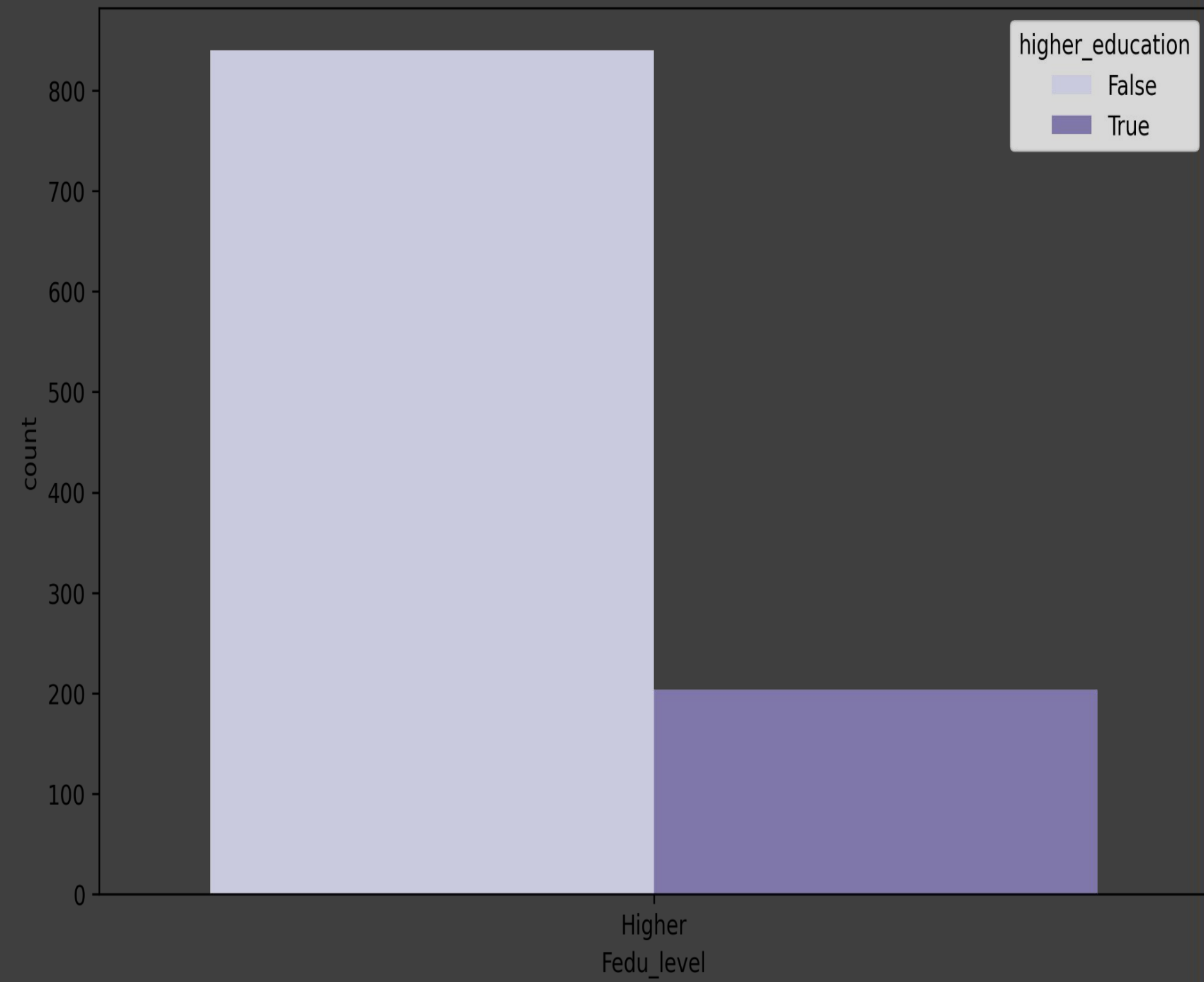
- ❖ Class 0 (Not Pursuing Higher Education):
 - 100% precision & recall – no false negatives 🚀
- ❖ Class 1 (Pursuing Higher Education):
 - 100% precision – all predicted students belong to this class 🎓
 - 86% recall – some students may be missed (false negatives) ⚠
- ❖ F1-Score Insights:
 - High F1-scores for both classes 🔥
 - Slight preference for precision in predicting higher education students 📊

Higher parental education levels (likely contributing to students' higher academic performance) are strongly associated with students' interest in pursuing higher education.

Interest in Higher Education vs Mother's Education



Interest in Higher Education vs Father's Education



- A large proportion of students whose parents have lower education levels are not pursuing higher education (lighter bars dominate).
- In contrast, students with more highly educated parents show a greater interest in continuing their education (darker bars increase).
- This suggests a strong correlation between parental education and a student's likelihood of seeking higher education.



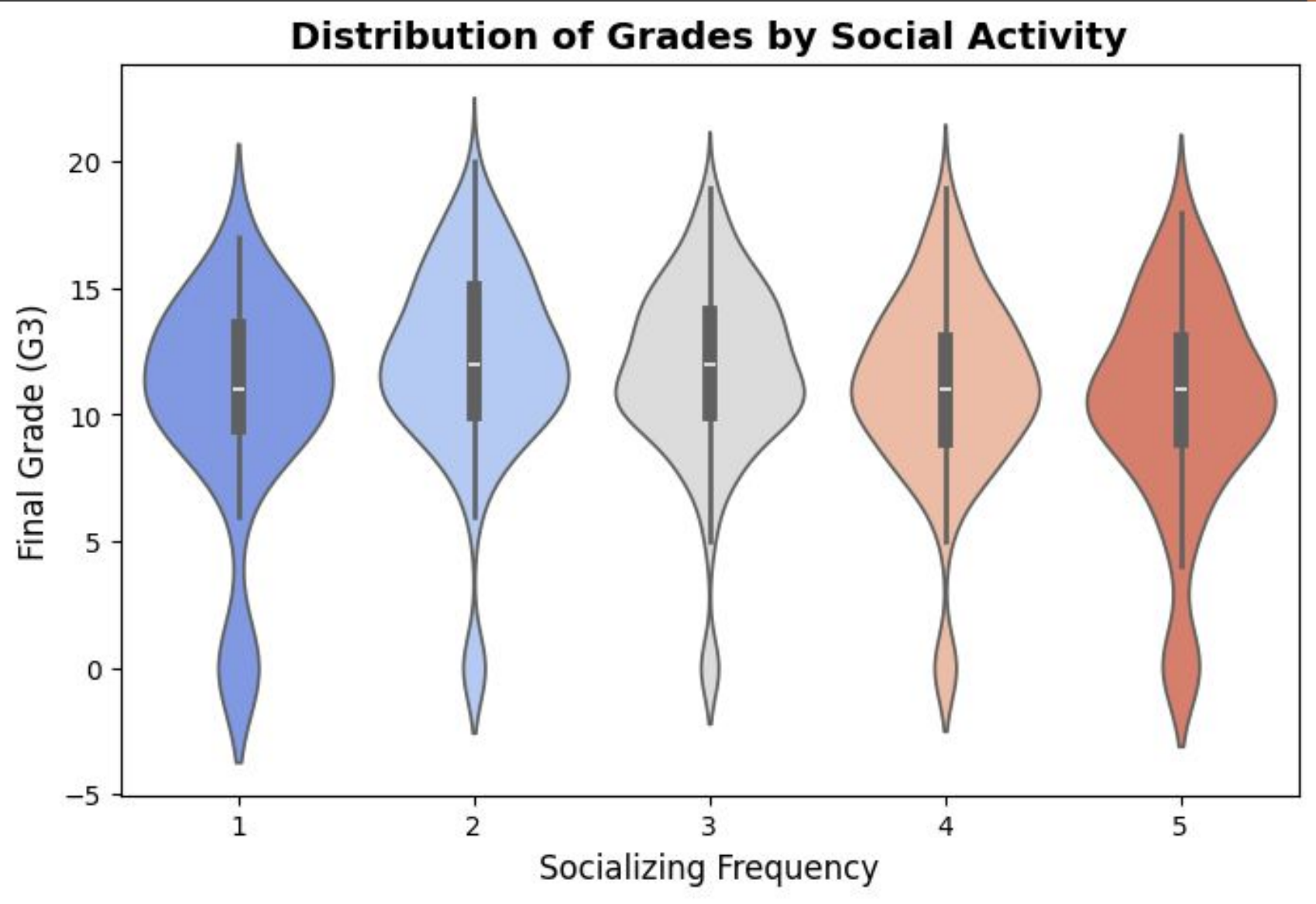
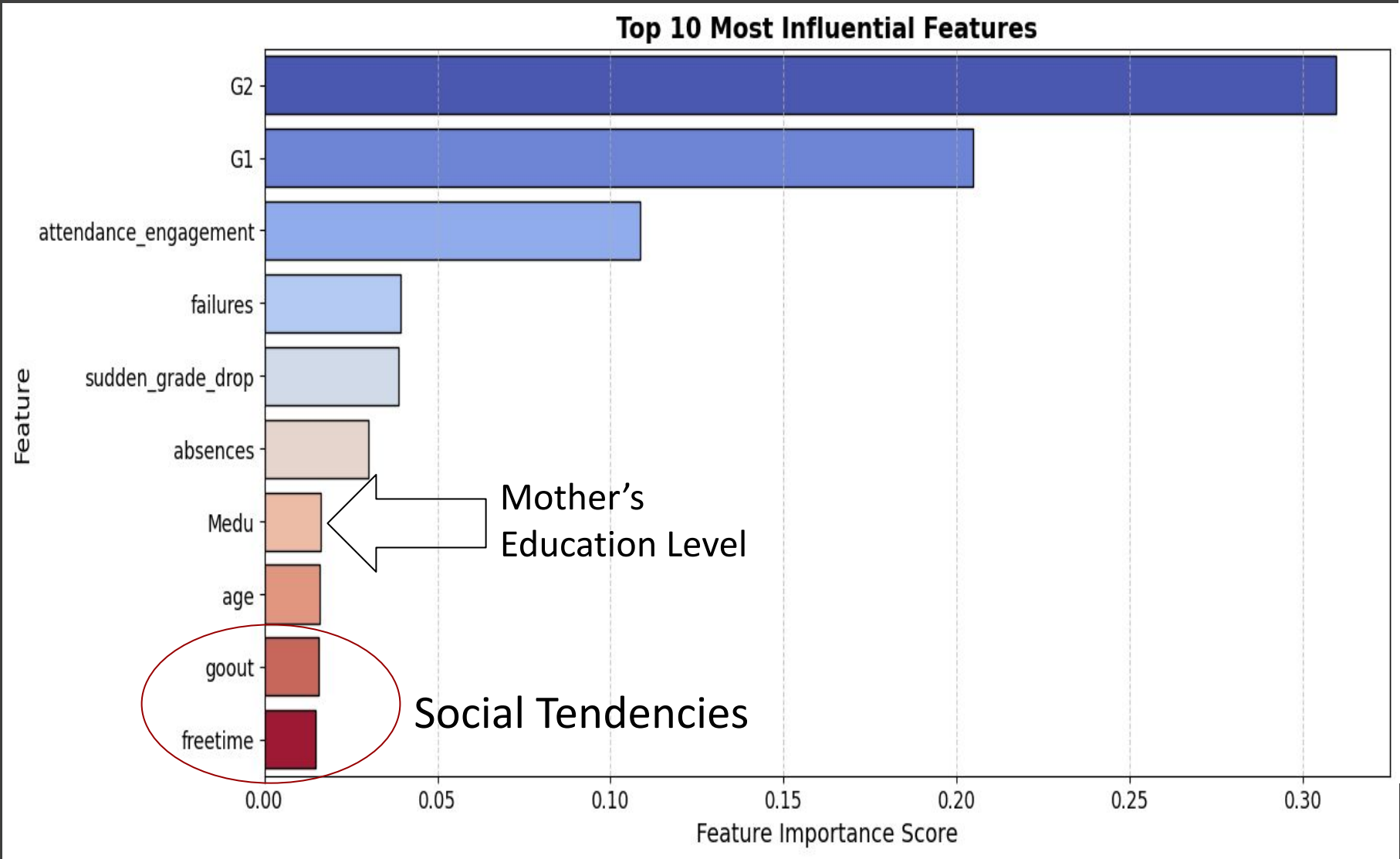


Social Tendencies



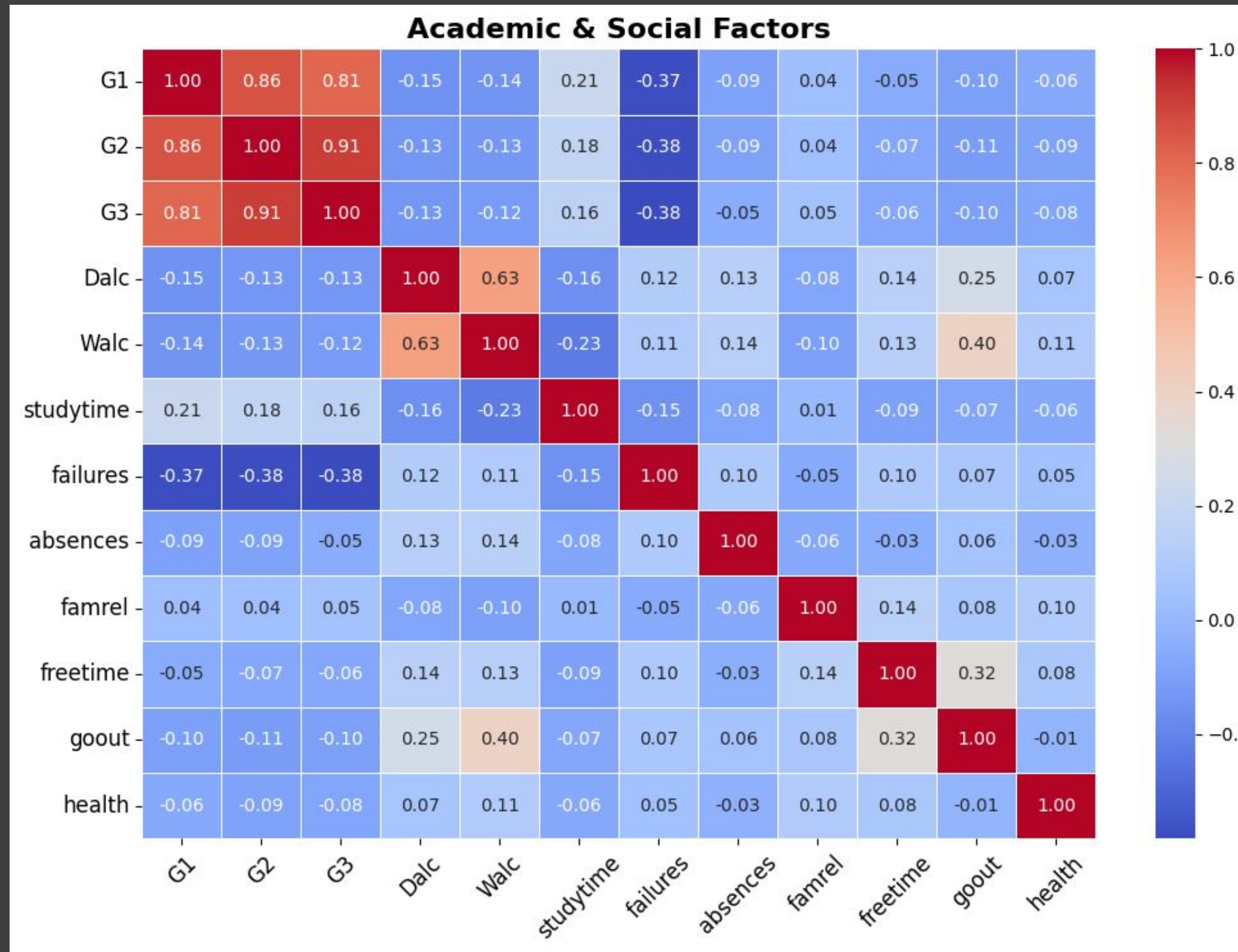
Social Tendencies and their correlations

What impact do certain **social tendencies** have on the students' likelihood to succeed? (e.g. Frequency of alcoholic consumption, social lives, extracurricular activities, romantic relationships, etc.)



Social Tendencies and their correlations

Ethan





Health and Family



Nathaniel

How does health impact a person's likelihood to succeed?



01

Does poor health often result in more absences and poorer grades?

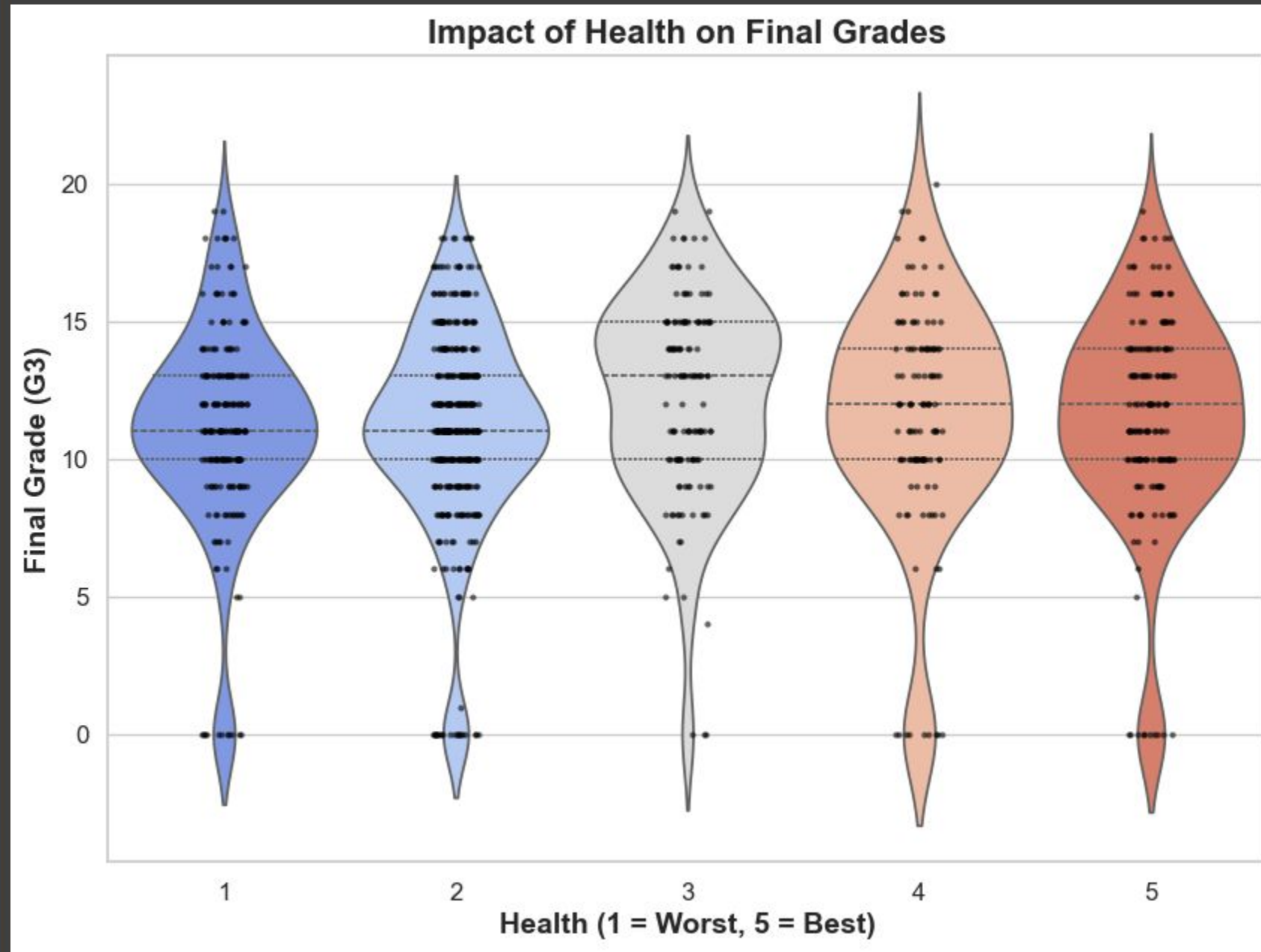
02

Is there a high incidence of absence and failure among students classified in poor health?

03

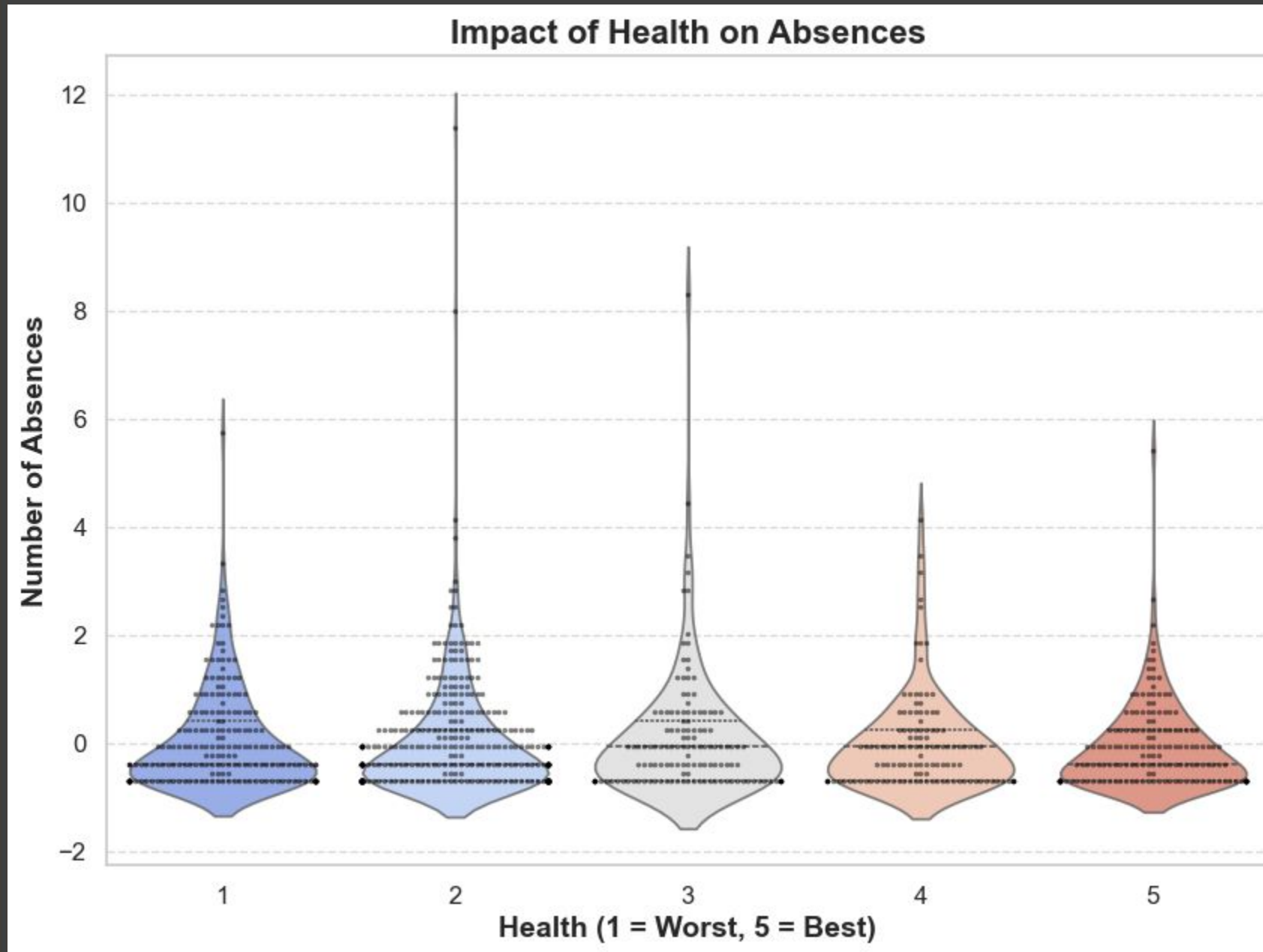
Is there a correlation between health and quality of family relationships?





Health & Success

- There is a weak negative correlation with grades ($\rho = -0.081$, $p = 0.009$), suggesting that poor health slightly lowers grades.
- Conclusion: While health may have some impact, it is not the main predictor of success.



Health and Absences

- No strong correlation ******($p = -0.034$, $p = 0.267$)****** → Poor health ****does not**** increase absences.
- ****Conclusion:**** Health ****affects grades more than attendance****.

OLS Regression Results

```

=====
Dep. Variable:          famrel    R-squared:                0.011
Model:                  OLS       Adj. R-squared:           0.010
Method:                 Least Squares    F-statistic:          11.42
Date:                  Tue, 18 Mar 2025    Prob (F-statistic):    0.000755
Time:                  18:31:13    Log-Likelihood:        -1475.7
No. Observations:      1044    AIC:                   2955.
Df Residuals:          1042    BIC:                   2965.
Df Model:               1
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	-3.469e-17	0.031	-1.13e-15	1.000	-0.060	0.060
health	0.1041	0.031	3.379	0.001	0.044	0.165

```

=====
Omnibus:                155.627    Durbin-Watson:          1.952
Prob(Omnibus):           0.000    Jarque-Bera (JB):       240.357
Skew:                    -1.008    Prob(JB):               6.41e-53
Kurtosis:                 4.207    Cond. No.                1.00
=====

```

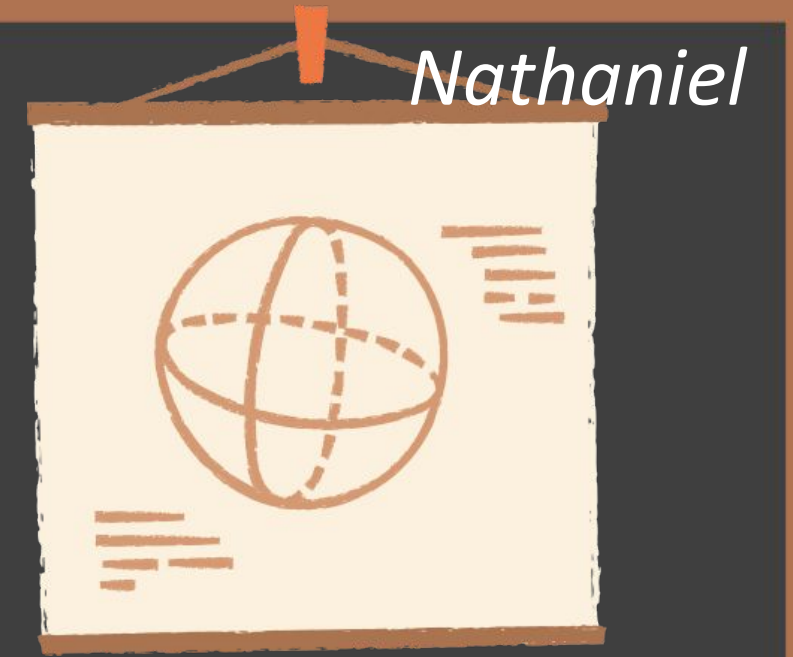
Health & Family Relationships

- Weak positive correlation between health and quality of family relationships. ****($p = 0.090$, $p = 0.0037$)**** → Healthier students report better family relationships.

- **Conclusion:** Family support may contribute to both **better health and academic success.**



Implications



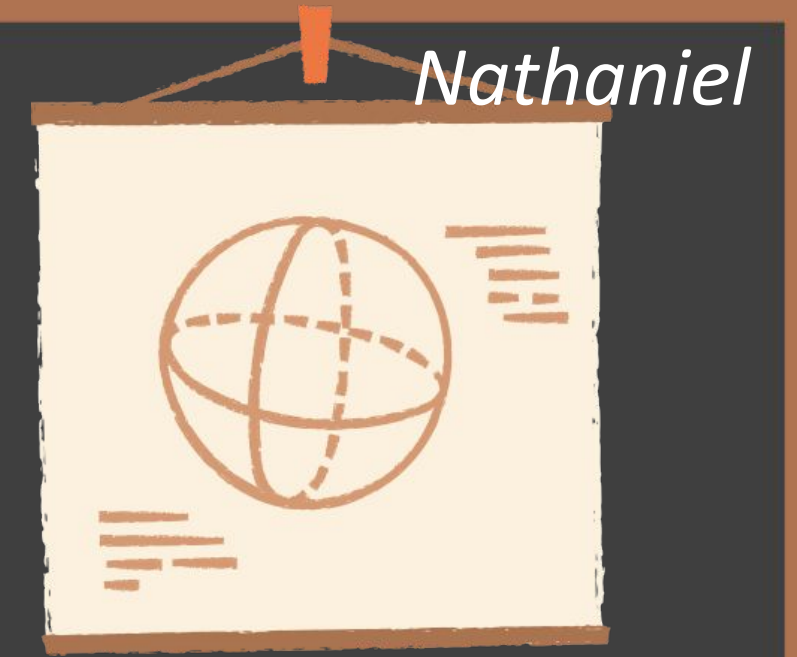
If schools want to identify at risk students, they should be looking at:

- ✓ grades: sudden drops, failures
- ✓ attendance: absence, engagement
- ✓ mother's education





Summary & Conclusions



- ★ Successfully optimized our baseline model
- ★ A Decision Tree Model was most accurate
 - Less linear relationships
 - Some features heavily skewed the chances of a student being at risk
- ★ We realized that we could have made a features prediction earlier on to ensure we were analyzing the most impactful features



**Thank
you!**

