Introduction

In professor Morawski's class students were asked to fill out a survey that would indicate their habits related to exam grade. The goal of this analysis was to understand how different study habits and behaviors relate to student performance, using a dataset containing survey responses and grade data from 188 students. Each record included both numerical and categorical data such as study hours, sleep before the exam, time spent on social media, class attendance, and whether students left the exam early. Three types of predictive models were built:

1. Predict whether a student passes or fails.
2. Predict the student's letter grade A through F
3. Predict the student's exact numerical score.

For each part, several machine learning algorithms were trained. The dataset was preprocessed by filling missing values, scaling numerical features, and one-hot encoding categorical variables. Models were evaluated using accuracy, precision, recall, and regression metrics with RMSE and R^2.

Feature Engineering

Each student's performance was expressed as Total score / Max Points:

Grades were then binned into both binary Pass/Fail and letter grade categories:

- Pass (1) for A, B, C (≥70%)
- Fail (0) for D, F (<70%)

Features were divided into:

- Numerical: hours studied, hours of social media use, and hours of sleep before the exam.
- Categorical: attendance frequency, prior ML experience, completion of readings, early departure, and year (grade level)

Categorical variables were converted to binary indicators using pd.get_dummies() and numerical features were standardized with the StandardScalar function.

The final dataset contained 188 responses

Predicting Pass Vs Fail

Three models were trained: KNN, Logistic Regression, and Decision Tree

A 60/40 train test split was used, stratified by the pass/fail label. Evaluation metrics included accuracy, precision, and recall.

Model	Accuracy	Precision	Recall
K-Nearest Neighbors	0.66	0.58	0.58
Logistic Regression	0.71	0.76	0.42
Decision Tree	0.55	0.45	0.48

Analysis:

Logistic Regression performed best overall with 71% accuracy, indicating a reasonably strong linear relationship between student habits and passing outcomes. However, the recall of 0.42 shows that the model missed many failing students. It's better at confirming who passed than detecting who might fail.

The KNN model had more balanced precision and recall, meaning it's less confident but more even handed. The Decision Tree underperformed, possibly due to overfitting small subsets or the limited dataset size. Confusion matrices were printed to visualize classification balance. Logistic Regression showed very few false positives and a relatively balanced split between pass/fail predictions.

Students who studied more hours, slept adequately, and consistently attended lectures tended to fall into the "Pass" group. However, nonlinear interactions between habits along with other outliers likely limited the models' predictive power.

Predicting Letter Grades A–F

For the second task, letter grades were predicted using KNN, Logistic Regression, and Random Forest

The data was split 85/15 train/test and evaluated using accuracy and the classification report (precision, recall).

Model	Accuracy
K-Nearest Neighbors	0.38
Logistic Regression	0.52
Random Forest	0.41

Analysis:
All models struggled to differentiate between individual letter grades. The Logistic Regression model achieved the best performance at 52% accuracy. Most misclassifications occurred among the mid grade categories C and D, which had overlapping feature distributions. It was found that larger train sets for the data yielded better results.

Letter grades represent fine grained distinctions that are difficult to separate using behavioral data alone. Features like "study hours" and "sleep" may correlate weakly with small score differences, making it hard for models to predict specific letters. The dataset also has class imbalance such as very little A's, which reduces model performance for underrepresented classes like A.

Predicting Exact Grade

Two regression models were tested:

Model	RMSE	R^2
Linear Regression	8.50	0.18
Ridge Regression	8.49	0.18

Analysis:
Both regression models explained roughly 18% of the variance in total scores, suggesting that the selected study behavior features only modestly predict performance. The RMSE being 8.5 points indicates that the models' grade predictions typically deviate by about one letter grade from the true score. Utilizing Ridge did not meaningfully improve results, confirming that overfitting was not the main issue; rather, the available features lacked strong predictive power to the level that would be required to predict exact scores.

Conclusions

Across all models, Logistic Regression consistently provided the best accuracy and interpretability for predicting whether a student passes or fails. For multiclass grading, models struggled due to overlapping categories and limited data size.

From a behavioral perspective, the analysis suggests:

- Sleep duration are weak but positive predictors of success
- Screen time has negligent correlation to grade
- Study hours has a slight negative correlation to grade (likely due to outliers)

However, the modest performance metrics imply that grades depend on additional factors not captured in the survey (cmsc gpa , motivation, studied with friends?)

To improve predictive power we could incorporate more statistical and behavioral data (study schedules, GPA history) and analyze the importance of each category to identify which behaviors most affect grades. In addition, simply having more data would also assist in finding the best results.

Overall, this project demonstrates how machine learning can reveal patterns in student behavior, even if precise grade prediction remains challenging. Logistic Regression proved to be the most reliable model for broad outcomes pass/fail, while regression models were less effective for predicting exact scores.