Executive Summary

Name: Abdul Oladosu

Project Summary & Problem Framing

Goal: The goal of this project is how we can spot students who may struggle early enough to help them.

I built forecasting models that estimate a student's final grade (G3) using information schools already collect (early-term grades, attendance, study habits, and a few background indicators). The purpose is support, not punishment: give counselors and teachers an early heads-up so they can reach out with the right intervention at the right time.

What I learned (at a glance).

- **Good early-term grades (G1, G2)** are the strongest signals of the final outcome.
- **Absences add up:** as missed days accumulate, final grades generally decline.
- **Study time** and **prior failures** help us identify students who benefit from coaching even before any term grades exist.

What are my Recommendations?

1. **Focus on what we can influence fast:** We should try to work on nudging attendance and the study habits of students by offering short structured support such as study plans, study groups, reminders, peer groups.
2. **Keep humans in the loop**: A predicted result should start conversations about how we can help students, letting advisors and teachers decide next steps.

Data Overview

What's in the data? Student demographics, per-term absences, self-reported study time, prior failures, and the three course grades: **G1** (Term 1), **G2** (Term 2), **G3** (Final). The file is moderate in size and workable with standard cleaning (imputing a few missing values; encoding yes/no style categories).

The **target variable** is G3 (final grade) on a 0–20 scale.

Target variable key relationships in simple terms:

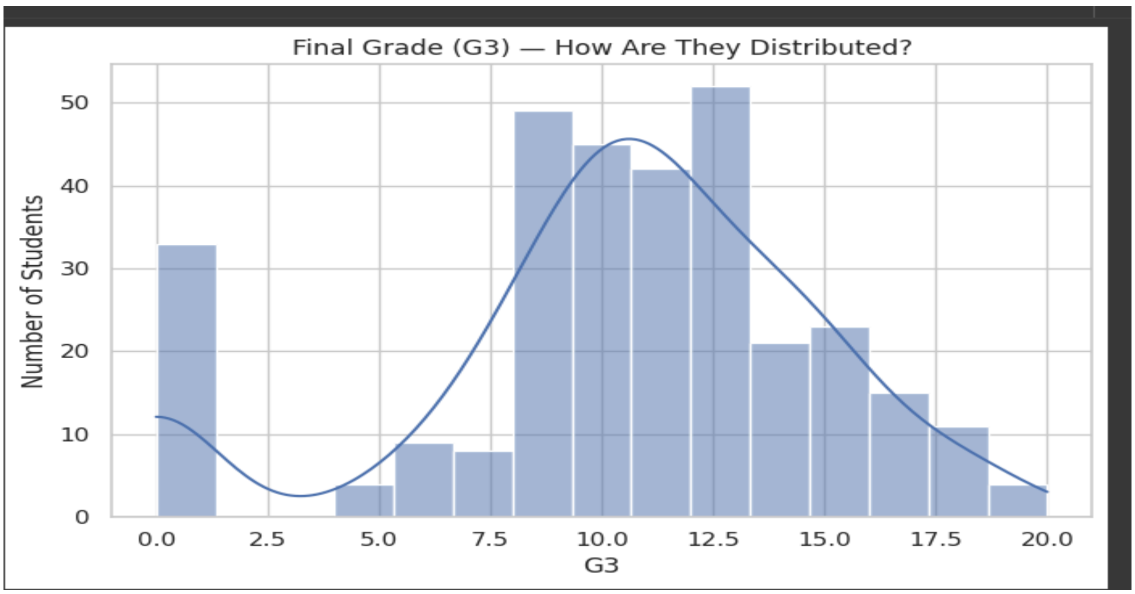
- Early performance → final performance. Students strong in G1/G2 typically finish strong in G3.
- Attendance is a steady drag. A higher total of missed classes correlates with lower G3.
- Study time helps (modestly). More study generally relates to higher G3.

Why two model versions?

- **With G1/G2:** highest accuracy, but only after those grades exist (mid-year).
- **Without G1/G2:** less accurate, but **available immediately—**when prevention can still change the trajectory.

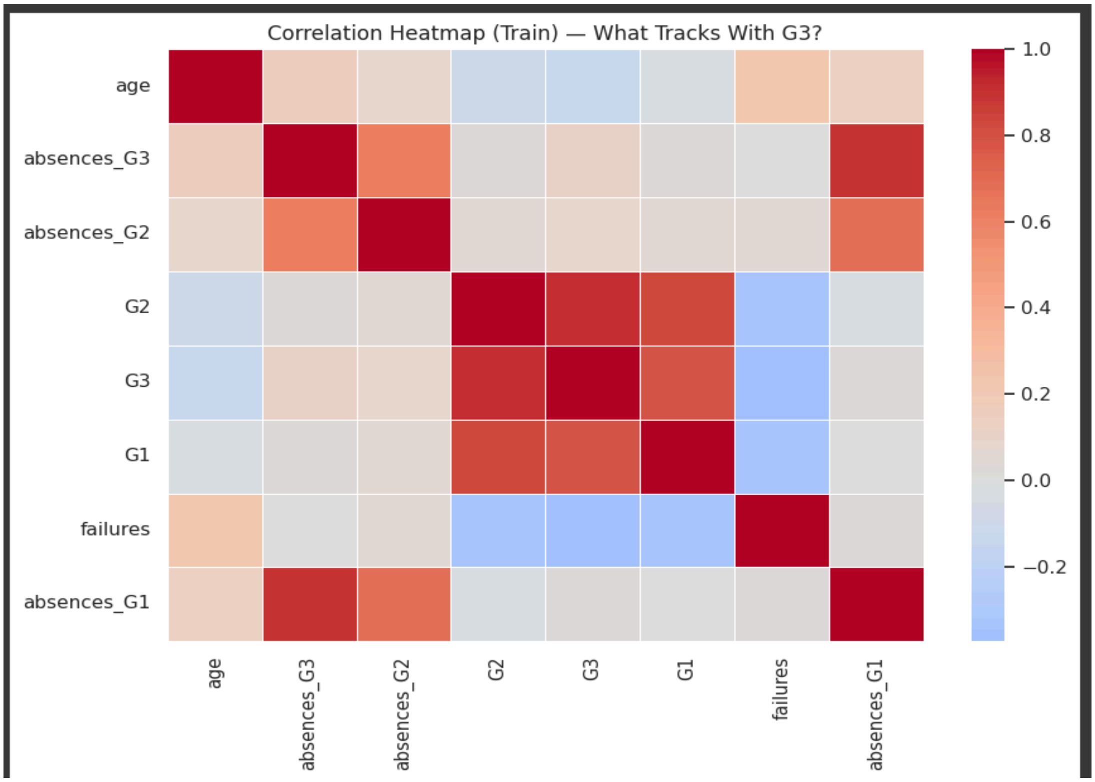
Analytical Insight

Figure 1 — Final Grade (G3) Distribution



What it shows. A histogram of final grades with a smooth curve overlaid. We see a broad spread, with **clusters near pass thresholds**.
Why does it matter? Students just above/below these cutoffs are **high-leverage**: small, timely supports can move them into a safer range.

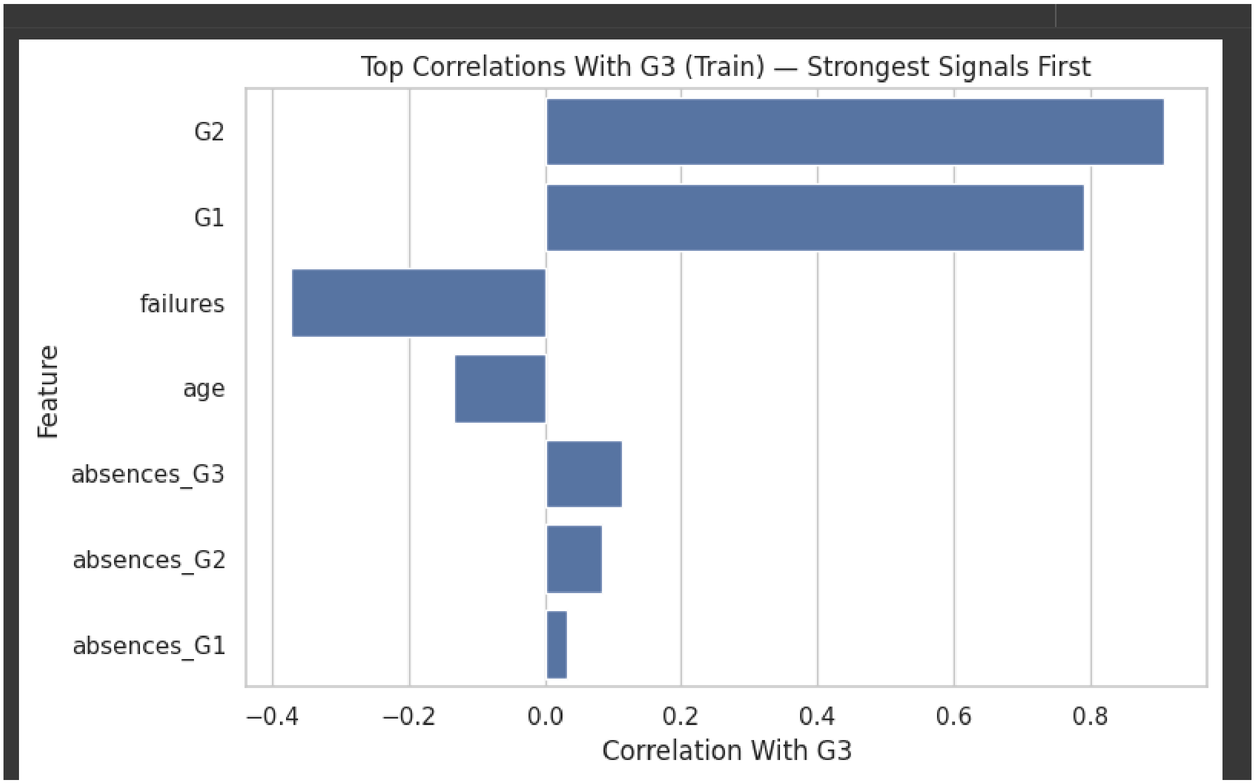
Figure 2 — Correlation Heatmap (Train) — *What Tracks With G3?*



What it shows. A “weather map” of the grade1(G1), grade2 (G2), Grade3(G3), student age, failures (previous class failed) and number of school absences in previous year’s grade(1,2,3) correlates with students final grade (G3). As shown above the **G1 and G2** grades show the strongest positive relationship with the student’s final**G3**: if a student does well in the G1 and G2 grades there is a high chance of performing well on the G3 grade,the higher the number of past classes failed the shows negative impact on how well student will perform, absences relate weakly-to-moderately negatively.

Why does it matter? This validates our intuition (early grades matter) and highlights **attendance** and **failure history** as practical risk indicators for early outreach.

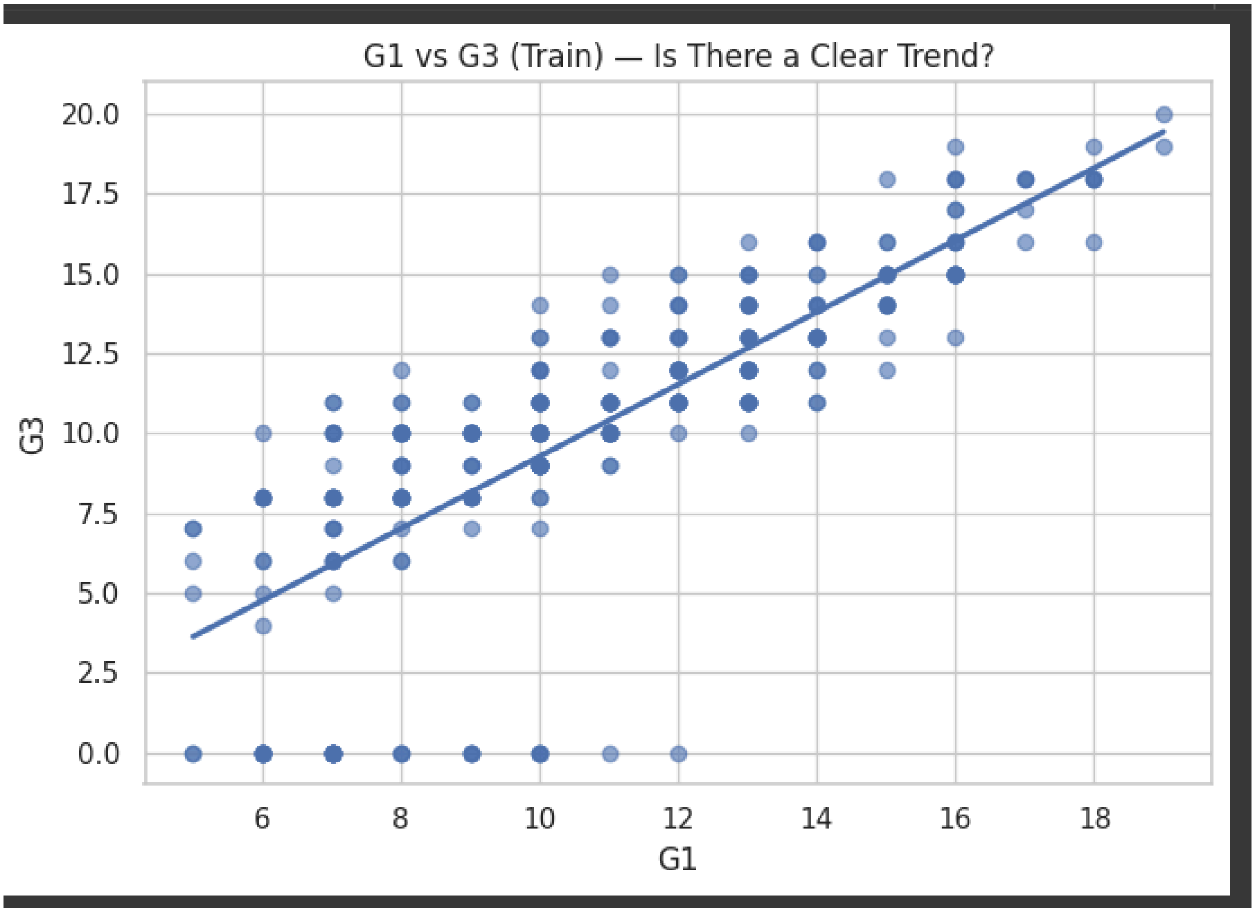
Figure 3 — Top Correlations With G3 — *Strongest Signals First*



What it shows. A ranked bar chart: **Grade 2** and **Grade 1** lead,showing how strongly it relates to Grade 3, followed (negatively) by **failures**; **age has a smaller effect on Grade3** and individual term absences show low to moderate .effect on grade 3

Why does it matter: This makes it obvious **to prioritize** the Grade 1 and 2 in modeling an intervention for students.

G1 vs G3 (Train) — *Is There a Clear Trend?*



higher G1 tends to mean higher G3.
Why does it matter: This is a simple, persuasive picture for staff: early performance is a reliable signal.

Beyond the charts (extra insights).

- **Prior failures** flag students for early, subject-specific remediation.
- Context signals (e.g., access or support indicators) have smaller effects; they don't define outcomes but help shape tone/timing of outreach.

Features I learned (and why).

- **High-signal academic:** Upon exploring the numeric attributes G1, G2 gave strongest predictors.
- **Actionable behavior:** engineered absences_sum and studytime — levers we can influence during the term.

- **History:** failures will help prioritize coaching.

Methodology

Models

- **Linear Regression:** This gives a straight-line from inputs to final grade, it is simple and transparent.
- **SVR (Support Vector Regression):** This model can fit **curves** when a straight line isn't enough.
- **Lasso Regression:** this is a type of linear model that **shrinks weaker inputs**, keeping the model focused and less noisy.

How we judged success.

- **RMSE:** "On average, how far off are we, in grade points?" The **smaller is better**.

Key Results

How we measured accuracy (in simple terms),We used two easy-to-understand numbers:

- **RMSE** — "How far off are we, on average, in grade points?" (smaller is better).
- **R^2** — "How much of the ups and downs in grades does the model explain?" (closer to 1 is better).

Models Result before fine tuning and performance measure

With G1 & G2 (mid-year onward)

- Linear Regression Score = 1.9
- SMV = 2.4
- Lasso = 2.2

Best model: Linear Regression
RMSE: 1.9

Without G1 & G2 (early-year)

- Linear Regression Score = 4.6
- SMV = 4.3
- Lasso = 4.4

Best model: SVR (SVM Regression)
RMSE: 4.3

What are the models delivered on the test set?

- **When G1 & G2 are available (mid-year and later)**
  Best model: Linear Regression
  RMSE: 2.22 | R^2: 0.76

  Once we have Term 1/2 grades, the model is usually within about **2.2 points** of the final grade. That's solid accuracy for making confident, targeted decisions.

- **When G1 & G2 are not available (early in the term)**
  Best model: SVR (SVM Regression)
  RMSE: 4.18 | R^2: 0.15

  Early in the year, before any grades exist, the model is usually within about **4.2 points**. It's less precise (expected), but it's available when prevention matters most.

Should we use this in practice?

Yes—thoughtfully.

- **No grades yet:** Use the **SVR** score to spot students early for light, supportive outreach like attendance nudges, study-skills check-ins.

- **After Term 1 (grades available):** Switch to **Linear Regression** score for **precise targeting**, tutoring invitations, tailored study plans, family outreach as needed.

- Keep it **human-led**: predictions start caring conversations; counselors and teachers decide what to do next.

Suggested Action

1. **Watch attendance momentum.** If total absences start creeping up, step in quickly (friendly reminders, counselor check-ins).

2. **Run short "study sprints."** Offer planning sessions and peer-study groups for students flagged early by the SVR model.

3. **Prioritize repeaters.** Students with prior failures should get early, subject-specific coaching.

4. **Keep it visible.** A simple dashboard per counselor/grade team: who's flagged, why they were flagged (e.g., absences, low study time), what support was given, and what happened next.

Production recommendation.

- **Phase 1:** Run SVR without grades weekly to flag gently (attendance reminders, study-skills check-ins).
- **Phase 2 (after Term 1):** Switch to Linear Regression with grades for precise support (tutoring invites, teacher conferences, family outreach).

- Keep predictions human-led—they should prompt caring conversations, not automatic decisions.

Business actions informed by the data.

1. **Attendance momentum:** Track absences_sum (and recent-weeks trend). Escalate supports as soon as the trend worsens.
2. **Study-skills sprints:** Short planning sessions and peer-study groups for students with low study time or borderline predictions.
3. **Early remediation for repeaters:** Prior failures merit early, subject-specific coaching.
4. **Lightweight dashboards:** For each counselor/grade team: who's flagged, why, what we tried, and what happened.

Conclusion

What worked. The two-scenario design lets us act early (no G1/G2) and tighten accuracy mid-year (with G1/G2). Linear Regression excelled with grades; SVR gave a useful early signal without them. The figures make the story intuitive for non-technical staff.

What didn't. The no-grades model is naturally less accurate; some categorical inputs added little and needed careful handling. This underscored the need for human judgment and simple, explainable rules alongside the model.

What's next?

- Pilot the two-stage workflow with one grade level; log flags → supports → outcomes.
- A/B-test interventions (tutoring vs. coaching vs. study prompts) to learn what helps which students.
- Retrain each term and audit fairness (e.g., by gender/address) to ensure equitable support.
- Add recency features (last 2–4 weeks of attendance) to sharpen early-term predictions.