Research Paper: Predicting Student Academic Performance

Lesson: Data Foundations for Machine Learning
Student Name: [Salim Sacad Muse]

1. Title & Collection Method

Title: "University Student Performance Dataset"

Collection Method: This dataset was created to simulate the real-world process of data collection for an educational study. The data synthetically represents a survey conducted among undergraduate students. The values for features like Weekly_Study_Hours and GPA were generated to reflect plausible correlations (e.g., higher study hours generally correlate with a higher GPA, but other factors like a long commute or a part-time job can mitigate this effect). This approach ensures a realistic and messy dataset that is perfect for practicing preprocessing techniques.

2. Description of Features & Labels

Features (X - Input Variables)

1. Age**:** Age of the student (Numerical, Integer)
2. Major**:** Field of study (Categorical: Computer Science, Engineering, Business, Arts)
3. Weekly_Study_Hours**:** Total hours spent studying per week (Numerical, Continuous)
4. Part_Time_Job**:** Whether the student has a part-time job (Categorical: Yes, No)
5. Commute_Distance**:** Distance from home to university in kilometers (Numerical, Continuous)

Label (y - Output Variable)

- GPA**:** The student's current Grade Point Average (Numerical, Continuous from 0.0 to 4.0). This is the value we want to predict, making this a **regression** problem.

3. Dataset Structure

- **Number of Samples (Rows):** 100
- **Number of Features + Label (Columns):** 6 (5 features + 1 label)

Sample Table (First 10 Samples)

Age	Major	Weekly_Study_Hours	Part_Time_Job	Commute_Distance	GPA
20	Computer Science	25	Yes	5.2	3.8
22	Business	15	No	12.5	3.2
21	Engineering	30	Yes	20.0	3.5
19	Arts	10	No	3.0	2.9
23	Computer Science	28	No	15.8	3.9
20	Engineering	18	Yes	8.3	2.8
21	Business	22	Yes	10.0	3.4
20	Computer Science	15	No	7.5	3.1
22	Arts	12	Yes	25.0	2.5
19	Computer Science	35	No	4.2	4.0

4. Quality Issues

The dataset has been intentionally created with several real-world data quality problems that must be addressed during preprocessing.

1. **Categorical Text Data:** The Major and Part_Time_Job features are in text format ("Computer Science", "Yes"). These need to be converted into numerical values using techniques like **Label Encoding** or **One-Hot Encoding** before a model can process them.
2. **Different Scales (Requires Feature Scaling):** The numerical features (Age, Weekly_Study_Hours, Commute_Distance) are on vastly different scales (e.g., Age ~20, Study Hours 25, Commute Distance ~15.0). An algorithm would be heavily biased towards features with larger ranges unless **scaling** (e.g., Standardization or Normalization) is applied.
3. **Potential Outliers:** A few data points could be considered outliers. For example, one student has a Weekly_Study_Hours value of **48**, and another has a Commute_Distance of **34.1 km**. I will need to investigate if these are valid extreme values or errors.

4.Class Imbalance: The Major feature has a roughly equal number of samples for "Computer Science", "Engineering", and "Business", but fewer samples for "Arts" (approximately 10% of the data). This imbalance might cause a model to be less accurate for predicting the GPA of Arts students.

5. Use Case

Machine Learning Task: Regression
The goal is to predict a continuous numerical value (**GPA**).

Potential Application: This dataset could be used to build a predictive model for student academic advisors. By inputting a student's profile (age, major, study habits, etc.), the model could predict their likely GPA.

Benefit: This would help identify students who are at risk of underperforming early in the semester, allowing advisors to provide proactive support, recommend time management strategies, or connect them with tutoring resources. It could also help in understanding the key factors that contribute to academic success