Data Processing and Analysis Experiment Report

1. Dataset: PISA (Programme for International Student Assessment)

PISA is an international assessment that measures the abilities of **15-year-old students**, covering three subject areas: **Mathematics, Science, and Reading**.

In this study, we focus on the **2018 assessment results**, which include the scores of **hundreds of thousands of students worldwide**.

2. Research Objectives

- **To demonstrate that national gender inequality has a greater impact on academic performance than other national variables** (e.g., **international Gini coefficient, national GDP per capita**). The hypothesis suggests that the greater the gender imbalance in a country, the lower the student performance.
- **To explore whether a growth mindset can mitigate the negative impact of gender inequality**, examining the interaction between a growth mindset and national gender imbalance.

3. Data Analysis Requirements

- Feature importance ranking will be conducted using **Lasso, Random Forest, Permutation Feature Importance, and XGBoost**.
- Given the inherent uncertainty in machine learning methods, the results should be interpreted as one possible scenario, serving as a supplementary analysis for conclusions.

4. Experiment Results

4.1 Data Overview

- **Two levels of data were used: school level and student level**.
- **The dependent variable (academic performance) was measured using PV2, PV5, and PV9**.

- **Initial dataset size:** 612,004 records
- **After data cleaning:** 444,238 records
    - **Male students:** 221,826
    - **Female students:** 222,412

4.2 Correlation Analysis

- **GII (Gender Inequality Index) and GINI (Gini coefficient) both have a negative impact on academic performance**, with **GII showing the strongest negative correlation**.

4.2.1 Lasso Regression Method

(a) Introduction

Lasso (Least Absolute Shrinkage and Selection Operator) is a **regularization method** that **adds a penalty term to the empirical risk function** to control model complexity. When the penalty

term is an **L1 norm**, it forces certain feature coefficients to become zero, effectively performing feature selection.

$$\widehat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

As α **decreases**, the **predictive power of the model improves**, indicating that all seven feature variables contribute to student performance.

(b)Feature Importance Ranking (Lasso method)

- **As λ increases, coefficients shrink**, and features are dropped one by one.
- **The last remaining features are the most important ones**.

Table: Overall Feature Importance Ranking (Mixed-Gender Analysis)

Experiment	**Feature Importance Ranking (Most important → Least important)**
Lasso-PV2-Math	**GII, Gender, gmc, ESCS, GDP, GINI**
Lasso-PV5-Math	**GII, Gender, gmc, ESCS, GDP, GINI**
Lasso-PV9-Math	**GII, Gender, gmc, ESCS, GDP, GINI**
Lasso-PV2-Reading	**GII, ESCS, gmc, Gender, GINI, GDP**
Lasso-PV5-Reading	**GII, ESCS, gmc, Gender, GINI, GDP**
Lasso-PV9-Reading	**GII, ESCS, gmc, Gender, GINI, GDP**
Lasso-PV2-Science	**GII, gmc, ESCS, GDP, Gender, GINI**
Lasso-PV5-Science	**GII, gmc, ESCS, GDP, Gender, GINI**
Lasso-PV9-Science	**GII, gmc, ESCS, GDP, Gender, GINI**

(c) Conclusion:

- **GII (Gender Inequality Index) has a stronger influence than GINI (Gini coefficient) and GDP.**

- **GII negatively affects academic performance, as indicated by negative coefficients.**

(d) Gender-Specific Analysis

- **Male Students**

- **Top Influencing Features: GII, gmc, ESCS, GDP, GINI** (Consistently ranked highest across different subjects)

- **Female Students**

- **Top Influencing Features: GII, ESCS, gmc, GDP, GINI**

- **GII remains the most critical factor**, showing that **gender inequality significantly impacts female students' academic performance**.

4.2.2 Random Forest Method

(a) Introduction

Random Forest (RF) is an **ensemble learning method** based on the **Bagging** approach. It introduces **random feature selection** during training to enhance model robustness.

(b) Feature Importance in Random Forest

- **GII ranked among the top three important features across all subjects.**

- **Gender-based analysis revealed that GII had a greater impact on female students than male students.**

4.2.3 XGBoost Method

(a) Introduction

XGBoost (Extreme Gradient Boosting) is an **optimized gradient boosting algorithm** widely used in machine learning competitions. It is an improved version of **GBDT (Gradient Boosting Decision Trees)**, designed for high efficiency.

(b) Feature Importance in XGBoost

- **Similar to Random Forest, XGBoost identified GII as one of the top influencing factors.**

- **GII consistently ranked high across both male and female student analyses.**

(c) Experimental Conclusion

This experiment utilized **Lasso regression, Random Forest, and XGBoost** to evaluate the impact of multiple independent variables (**ESCS, Gender, gmc, GII, GDP, GINI**) on student performance.

- **Across different methods, Gender Inequality Index (GII) consistently ranked as a highly influential factor, confirming that national gender inequality has a more significant impact on academic performance than GDP or the Gini coefficient.**
- **Countries with higher gender inequality tend to have lower student performance.**
- **Both male and female students are negatively affected by gender inequality, with female students experiencing a stronger impact.**

---

5. Final Thoughts

The results **reinforce the argument that reducing gender inequality at the national level could improve academic outcomes**. Additionally, **fostering a growth mindset may help mitigate the negative impact of gender inequality on student performance**.

Appendix

Dataset details

Student Variables

Variable (English)	**Variable**
CNTSCHID	International School ID
CNTRYID	Country Identifier
CNT	Country/Region Code (Three Letters)
CNTSTUID	International Student ID
gender	Student Gender (1 = Male, 2 = Female)
growth_mindset_origin	Growth Mindset Dimension

Variable (English)	Variable
ESCS	Economic, Social, and Cultural Status Index
W_FSTURWT, W_FSTURWT1-80	81 Base Grade Weight
W_SCHGRNRABWT	GRADE NONRESPONSE ADJUSTED SCHOOL BASE WEIGHT
W_FSTUWT_SCH_S	Sum of W_FSTUWT (W_FSTUWT 的总和)
SENWT	Senate Weight (5000 per country)
PV1MATH – PV10MATH	10 math grade
PV1READ – PV10READ	10 read grade
growth_mindset_Continuous	0-4 from small to big, represent the student's growth mind extent

---

National Variables

Variable (English)	Variable (Chinese)
CNTRYID	Country Identifier
GDP per capita	GDP per capita (Unit: USD
GII	Gender Inequality Index
GINI	Gini Coefficient
