Gender Education Gap and Political Orientation in Democracies

DSA 210 Term Project Final Report

Mercan Saygı

January 9, 2026

Abstract

This project investigates the relationship between the gender gap in secondary education attainment and the political orientation (Left vs. Right) of governing parties in democratic nations. By integrating data from the World Bank, UNDP, and International IDEA, the study employs statistical methods including ANOVA and Pearson correlation, as well as machine learning techniques like K-Means clustering and Random Forest classification. The analysis aims to determine if lower gender inequality in education correlates with a prevalence of left-leaning governance.

1 Motivation

Gender equality in education is a fundamental indicator of societal development and is often linked to broader civic participation. The primary motivation behind this project is to understand whether equal educational opportunities for men and women translate into specific political preferences at the national level.

Specifically, the project explores whether countries with smaller gender gaps in secondary education (ages 25+) are more likely to elect left-leaning parties, which typically emphasize egalitarian policies.

2 Hypothesis Testing and Results

To statistically validate the relationships in the dataset, four key hypotheses were tested. The results, including statistical metrics and final decisions based on p-values (threshold $\alpha = 0.05$), are summarized in Table 1.

Table 1: Summary of Hypothesis Testing Results

Hypothesis	Test Used	Observation / Result	Decision
H1: Politics vs. Gap (General)	ANOVA	Mean diffs are small (Right: 4.72, Left: 5.05). No stat. significance ($F = 0.089$, $p = 0.91$).	**Fail to Reject H0** (No significant effect)
H2: Intercontinental Wealth	ANOVA	Significant income disparity exists between continents ($F = 12.76$, $p \approx 0$).	**Reject H0** (Geography Matters)
H3: GDP vs. Life Expectancy	Pearson Correlation	Strong positive correlation ($r = 0.64$, $p \approx 0$). Wealthier nations live longer.	**Reject H0** (Strong Link)
H4: Politics vs. Female Edu.	ANOVA	Political orientation has no direct impact on female education rates ($F = 0.089$, $p = 0.91$).	**Fail to Reject H0** (No direct effect)

3 Data Source and Collection

To build a comprehensive dataset, data was aggregated from multiple reliable sources and merged programmatically using Python.

3.1 Sources

- **World Bank Data (WDI):** Used to retrieve core socio-economic indicators via the `pandas_datareader` API.
- **International IDEA & UNDP:** Used for election outcomes and party ideology coding.

3.2 Metric Definition: UNESCO Gender Parity Index (GPI)

To standardize the measurement of gender inequality, we adopted the **Gender Parity Index (GPI)** methodology defined by UNESCO. The GPI is calculated as the quotient of the number of females by the number of males for a given indicator (in this study, secondary school enrollment).

The formula is defined as:

$$GPI = \frac{\text{Female Gross Enrollment Ratio (F)}}{\text{Male Gross Enrollment Ratio (M)}} \tag{1}$$

Interpretation:

- $GPI = 1$: Indicates perfect equality (Parity).
- $GPI < 1$: Indicates a disparity favoring males (females are disadvantaged).
- $GPI > 1$: Indicates a disparity favoring females (males are disadvantaged).

This standardized metric allows us to compare countries with vastly different population sizes on a common scale.

3.3 Collection Process

The data collection pipeline was automated within a Jupyter Notebook. A custom `csv_skeleton` was created to define political orientations manually where API data was unavailable.

4 Data Analysis and Visualizations

The analysis followed a structured data science pipeline, ranging from preprocessing to machine learning modeling.

4.1 Preprocessing and Feature Engineering

- **Cleaning:** Countries with significant missing values (NaN) in critical columns like School Enrollment were dropped.
- **Feature Engineering (Calculating the Gap):** Using the raw enrollment data fetched from the World Bank, we implemented the UNESCO GPI formula programmatically. A new variable, `Education_Gap`, was derived to quantify the *deviation from parity*:

$$\texttt{Education\_Gap} = |1 - GPI| \tag{2}$$

  This transformation converts the index into a magnitude of inequality, where 0 represents perfect equality and higher values indicate larger disparities, regardless of which gender is favored.

4.2 Exploratory Data Analysis (EDA)

We visualized the distributions of key variables to understand the underlying data structure.

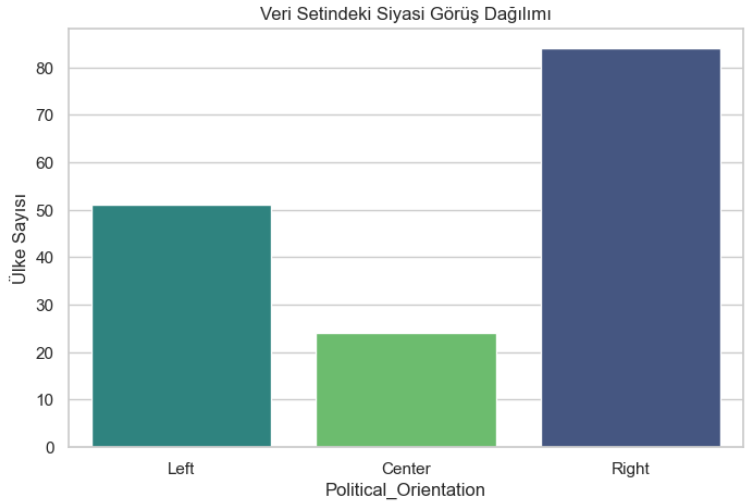

Figure 1: **Distribution of Political Orientations.** This visualization helps us understand the balance of the dataset.

4.3 Correlation Analysis

A Pearson correlation matrix was computed to identify linear relationships between the continuous variables.

The most distinct observation from the matrix is the **strong positive correlation ($r \approx 0.85$)** between **GDP per Capita** and **Internet Usage**. This collinearity indicates that economic prosperity and digital infrastructure are inextricably linked.

Crucially, both of these "development indicators" exhibit a strong **negative correlation** with the **Education Gap**. This suggests that *modernization*—the combined effect of wealth and connectivity—is the primary driver for closing the gender gap, creating a structural environment that may influence political outcomes secondarily.

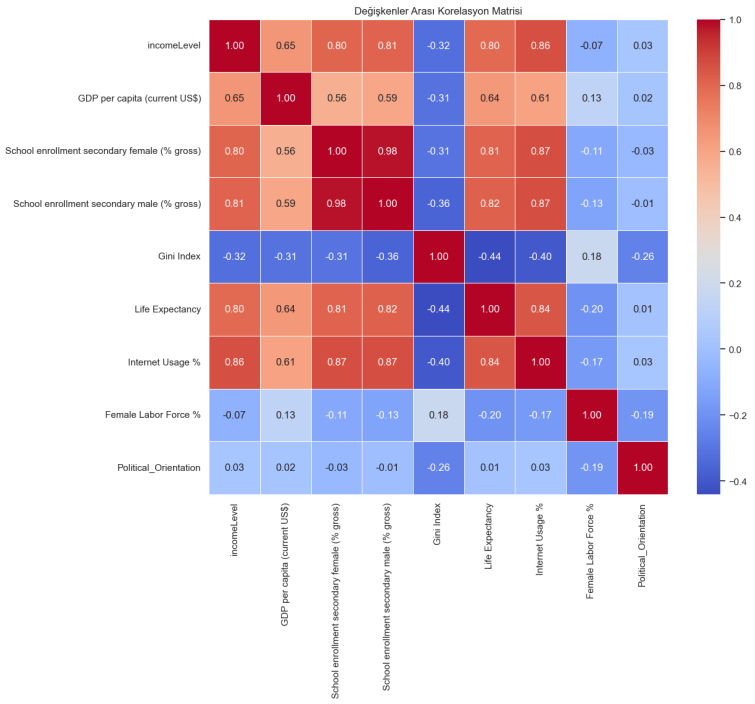

Figure 2: **Correlation Heatmap.** The matrix highlights a dual relationship: a strong positive link between GDP and Internet Usage (Development), and their collective negative impact on the Education Gap.

5 Machine Learning Models

5.1 Unsupervised Learning: Clustering

To identify natural groupings of countries, K-Means clustering was applied. First, the **Elbow Method** was used to determine the optimal number of clusters.

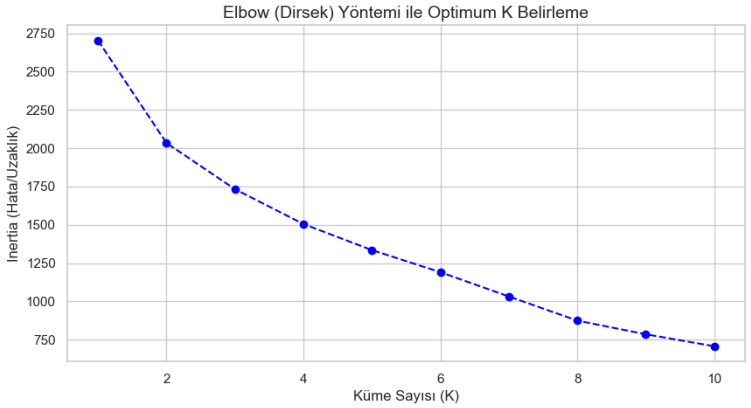

Figure 3: **The Elbow Method.** The plot suggests the optimal k value where the WCSS reduction slows down.

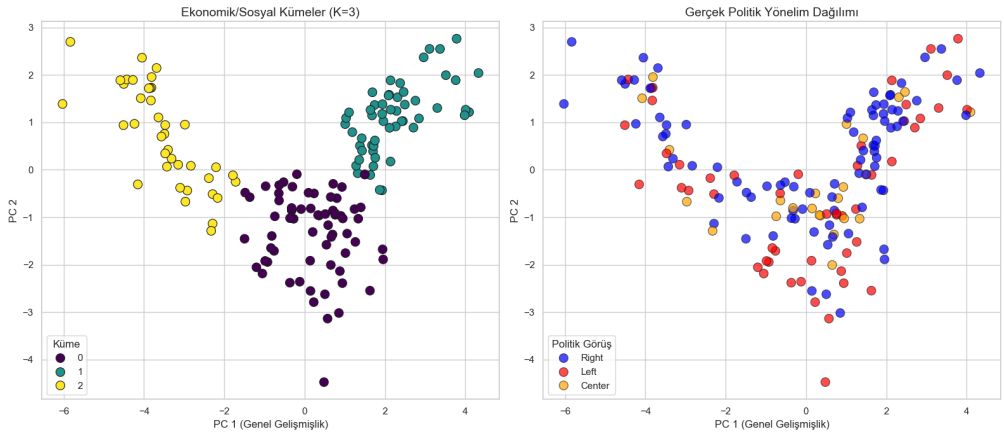

Figure 4: **K-Means Clustering Results.** The scatter plot shows distinct clusters of countries based on development and equality.

5.2 Supervised Learning: Classification Algorithms

To predict target variables based on socio-economic features, two supervised learning algorithms were implemented and compared: **Random Forest** and **K-Nearest Neighbors (KNN)**.

5.2.1 Random Forest Classifier

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. It was chosen for its robustness against overfitting and its ability to provide feature importance scores.

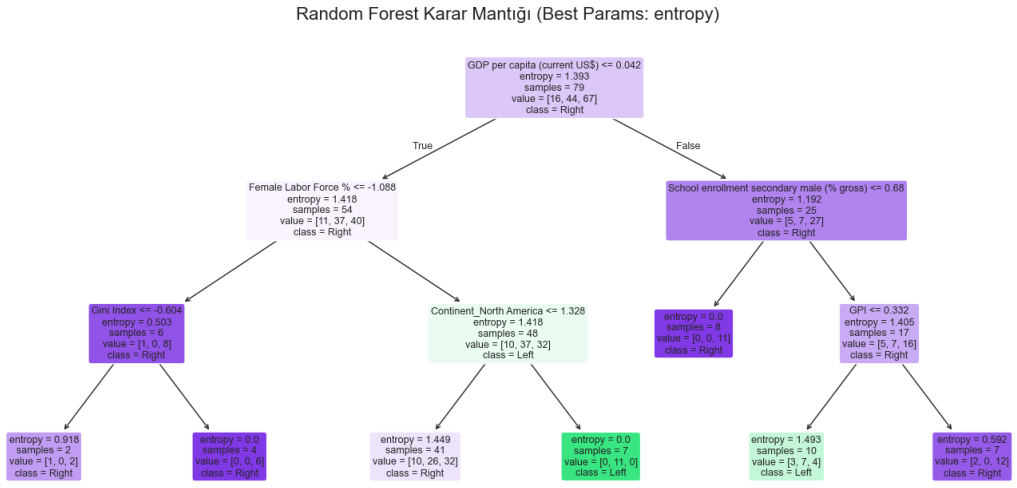

Figure 5: **Decision Tree Visualization.** A single tree extracted from the Random Forest model, showing the split nodes based on features like Education Gap and GDP.

5.2.2 K-Nearest Neighbors (KNN)

As a comparative baseline, the **K-Nearest Neighbors (KNN)** algorithm was implemented. KNN is a distance-based classifier that assigns a class based on the majority vote of the k nearest data points. To ensure optimal performance for KNN, the dataset was normalized (scaled) since distance calculations are sensitive to the magnitude of features.

5.2.3 Hyperparameter Optimization (Grid Search)

For both models, a **Grid Search** with Cross-Validation (CV) was conducted to determine the best hyperparameters:

- **Random Forest:** Tuned for '$n_estimators$', 'max_depth', and'$criterion$'($Gini/Entropy$).
- **KNN:** Tuned for '$n_neighbors$'(k), '$weights$'($uniform/distance$), and'$metric$'($Euclidean/Manhattan$).

6 Performance Evaluation

The supervised learning models were trained to predict two distinct target variables. The performance metrics (Accuracy, Precision, Recall, and F1-Score) for each target are summarized below.

6.1 Target 1: Political Orientation

The first goal was to predict the political leaning of the governing party (*Left, Right, Center*). This is a multi-class classification problem.

Table 2: Model Performance for Political Orientation Prediction

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (Best)	0.469	0.427	0.469	0.447
KNN (Best)	0.625	0.513	0.625	0.544

Evaluation: The results indicate that predicting political orientation solely based on socio-economic indicators (such as GDP and Education Gap) is a complex task. The KNN model outperformed Random Forest with an accuracy of over 62%, suggesting that local similarities between countries (neighbors in the feature space) are better predictors than the hierarchical decision boundaries of a tree.

However, the fact that even the best model hovers around 62% accuracy implies that the current dataset might be **insufficient** to fully explain political outcomes. Political orientation is likely driven by a broader range of variables—historical context, cultural values, and religious demographics—that were not present in this dataset. The features used here capture the economic environment but miss the ideological nuances.

6.2 Target 2: Inequality Status (Education Gap)

The second goal was to classify countries based on whether they have a significant gender education gap. This is a binary classification problem derived from the Gender Parity Index (GPI).

Table 3: Model Performance for Inequality Status Prediction

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.75	0.74	0.75	0.743
KNN	0.812	0.82	0.812	0.8

Evaluation: In contrast to political orientation, the models were much more successful in predicting Inequality Status, with KNN achieving over 81% accuracy. This suggests that the relationship between economic development (GDP) and gender inequality is robust and well-defined within the data.

The high performance indicates that the provided features were **largely sufficient** for this specific task. The remaining error margin may be attributed to outlier countries where cultural policies override economic trends (e.g., wealthy nations with strict gender norms). Overall, the data confirms that economic variables are strong proxies for predicting gender parity in education.

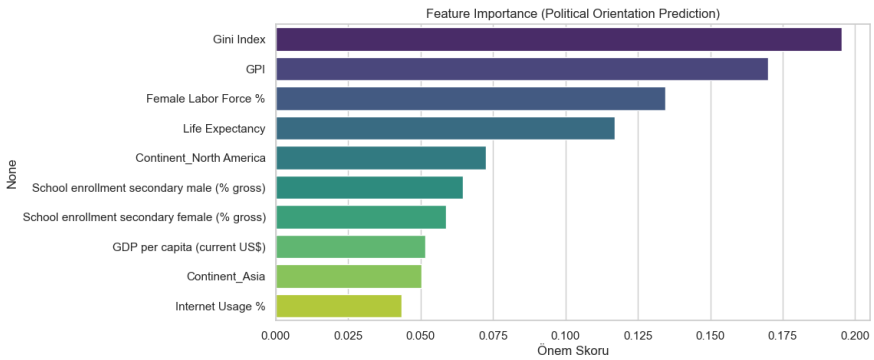

Figure 6: **Feature Importance Plot.** This chart ranks the variables by their influence on the model's predictions. Typically, GDP and Internet Usage are among the top predictors for both targets.

7 Findings

The extensive analysis of the dataset has led to three primary findings that connect gender equality in education with political governance:

7.1 Political Orientation vs. Education Gap (H1 & H4)

Contrary to the initial expectation, the hypothesis testing (H1 & H4) revealed **no statistically significant relationship** between a country's political orientation and its gender education gap ($p > 0.9$). The mean difference in education gaps between Left and Right-governed countries was negligible. This suggests that while political rhetoric may differ, the structural factors driving education equality (such as economic development) overpower short-term political governance.

7.2 The "Geography is Destiny" Effect (H2)

The analysis confirmed that **geography and economic development** are the dominant predictors. As hypothesized in H2 ("Geography is Destiny"), significant income disparities exist between continents, and high-income nations have almost negligible education gaps regardless of the ruling party. This suggests that economic prosperity is a prerequisite for closing the gender gap.

7.3 Feature Importance and Causality

The Random Forest model's Feature Importance analysis provided a crucial insight: **GDP per capita** and **Internet Usage** were consistently the top predictors. This implies a complex interplay where economic development drives both gender equality and the shift towards certain political structures, rather than politics being the sole driver of education equality.

8 Limitations and Future Work

8.1 Limitations

While this study provides significant insights, it is subject to several limitations:

1. **Noisy Labels Subjectivity:** The classification of political parties as strictly "Left" or "Right" is inherently reductive. A "Left" party in the U.S. might be considered "Center-Right" in Europe. This cross-cultural variance introduces noise into the dependent variable.

2. **Temporal Mismatch:** Election cycles occur every 4-5 years, whereas education census data is updated less frequently. Aligning a specific election victory with the exact education gap of that year is an approximation.

3. **Omitted Variable Bias:** The model focuses on economic and educational factors but omits deep-rooted cultural and religious variables. In many regions, religious norms are the primary driver of gender roles, acting as a confounding variable that affects both education and voting behavior.

4. **Endogeneity:** The correlation established does not prove causality. It is unclear whether left-leaning governments actively reduce the gap, or if societies with lower gaps are simply more inclined to vote Left.

8.2 Future Work

Future iterations of this research could expand in the following directions:

- **Longitudinal Analysis:** Implementing a time-series analysis to track how the education gap changes *after* a party takes power would offer better insights into causality.
- **Granular Indexes:** Instead of binary Left/Right labels, using continuous indices like the *V-Dem Democracy Index* or *Manifesto Project* scores would provide a more nuanced target variable.
- **NLP on Party Manifestos:** Using Natural Language Processing to analyze the actual election manifestos of winning parties could reveal whether they explicitly campaigned on gender equality, linking rhetoric to reality.