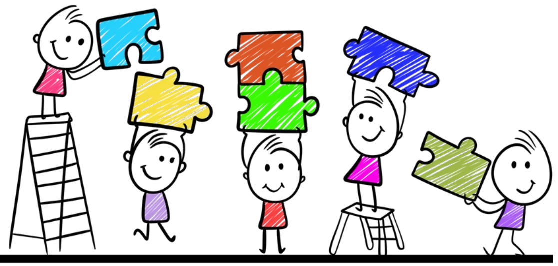
RECLAIMING CHILDHOOD

Unearthing the Roots of Child Labor

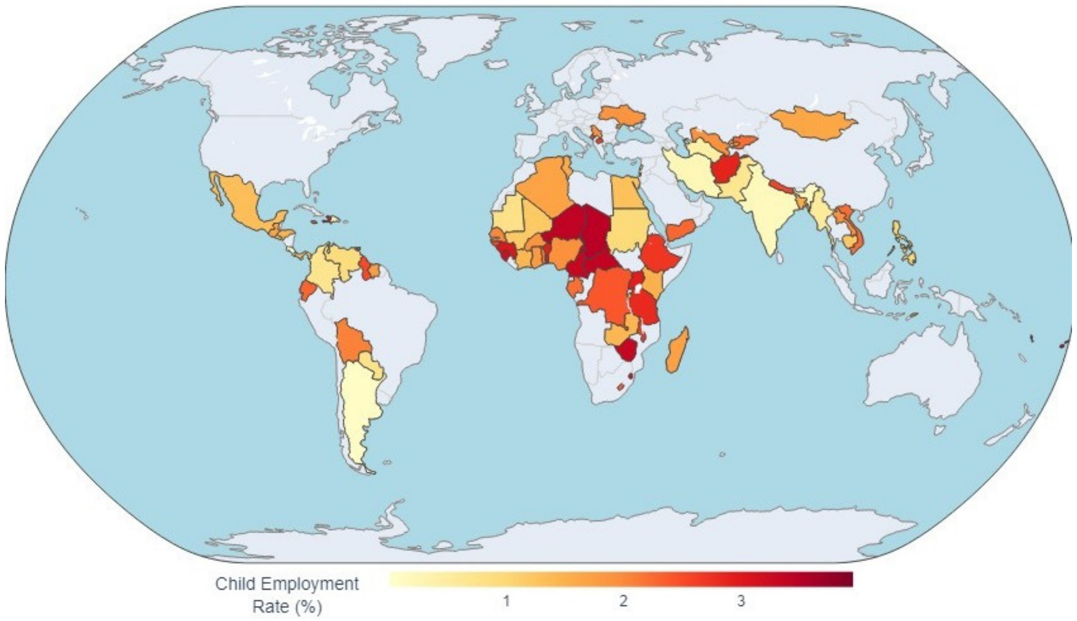


QuantaQuartet

Ahmed Önder AKKAYA | Atasagun YILMAZ | Taner YEŞİLAY | Tuana ÇERCİVE

ahmedakkaya@hacettepe.edu.tr atasagunyilmaz@hacettepe.edu.tr taneryesilay@hacettepe.edu.tr tuanacercive@hacettepe.edu.tr

Figure 6: Global log(Child Employment Rate) – Latest Available Data by Country

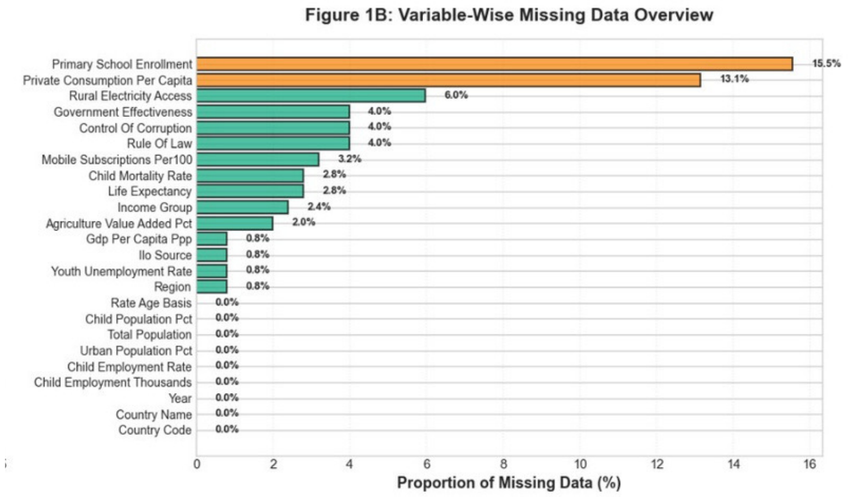


The reason we decided to focus on the issue of child labor—among the extensive data provided to all participants—is that, in recent years this topic has not received the attention it deserves due to other global challenges dominating public attention.

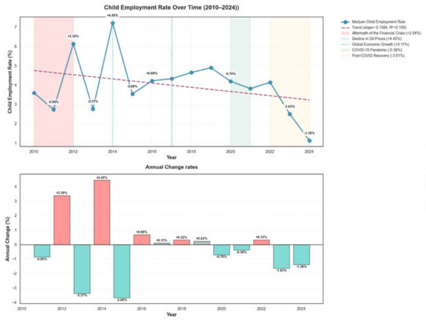
In the beginning of our presentation, We'd like to state that our dataset only includes countries where surveys were conducted and covers data collected *after* 2010.

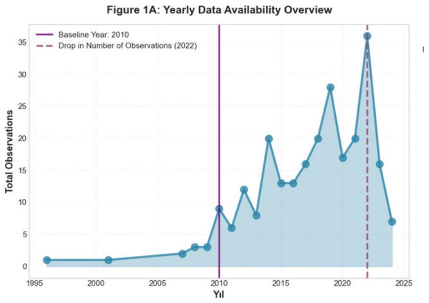
However, as can be observed from the world map chart on the left side, child labor remains a persistent and unresolved problem, particularly concentrated in the African region. Despite global efforts, the data clearly states that this issue continues to pose a significant social and economic challenge in many parts of the world.

QuantaQuartet



On the left, you can see the indicators we used and their proportion of missing data. Some key variables, such as Primary School Enrollment and Private Consumption per Capita, show higher data gaps — reflecting and demostrating the challenges of collecting consistent global data on social issues like child labor.





On the left, the charts shows that survey activities began to rise notably after 2010, which is our baseline year. However, after peaking around 2020, the number of observations sharply declined.

We believe this drop is not because the problem was solved, but because global priorities have shifted to economic crises, the COVID-19 pandemic, and to other issues have drawn attention away from child labor, which sadly still remains a persistent global challenge.

QuantaQuartet

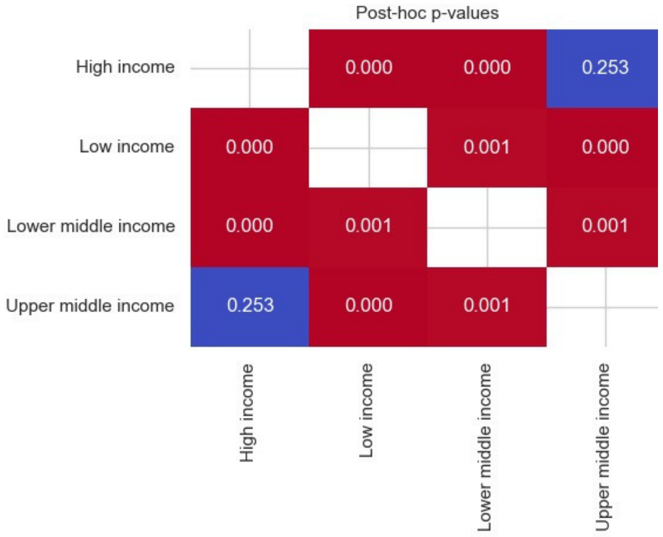
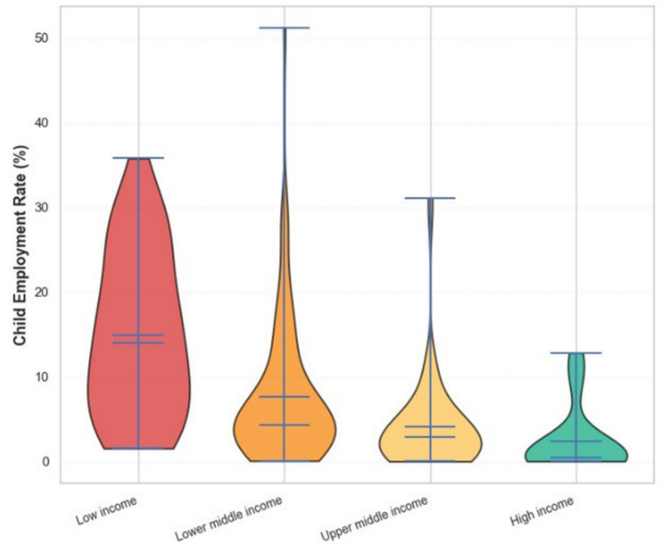


Figure 4B: Distribution by Income Groupm



Almost every pairwise comparison is significant($p \approx 0.000–0.001$), confirming that child employment rates strongly differ between income levels.
The only *non-significant* difference($p = 0.253$) is between high income and upper-middle income groups. Meaning these two groups have relatively similar, low child employment patterns. This pattern implies a nonlinear decline: most of the reduction in child labor happens as countries move from low to middle income, with diminishing returns afterward.

Low Income|Broken Symmetry

Child labor is not an exception but a norm. In some nations, nearly half of children work, not study. High variance shows crisis-driven volatility: poverty is chronic, but fragility makes it worse.

Lower Middle Income|The Transition Zone

The shape narrows, but the long tail remains. Some countries reform while others resist. Median declines, yet extremes still persist. This group is a turning point for nations.

Upper Middle Income|Fading Echoes

Variance shrinks and rates drop below 10%. Child labor becomes marginal, becoming a *statistical noise* rather than a systemic issue.
Still, rural or informal sectors hide residual exploitation. We think that this masks the issue on upper-middle

High Income|The Bimodal Paradox

The first peak near zero shows protection through welfare and education. But a small second peak reveals *invisible* children — migrants and refugee childeren are still working.
In the case of these countries, even prosperity leaves a shadow.

QuantaQuartet

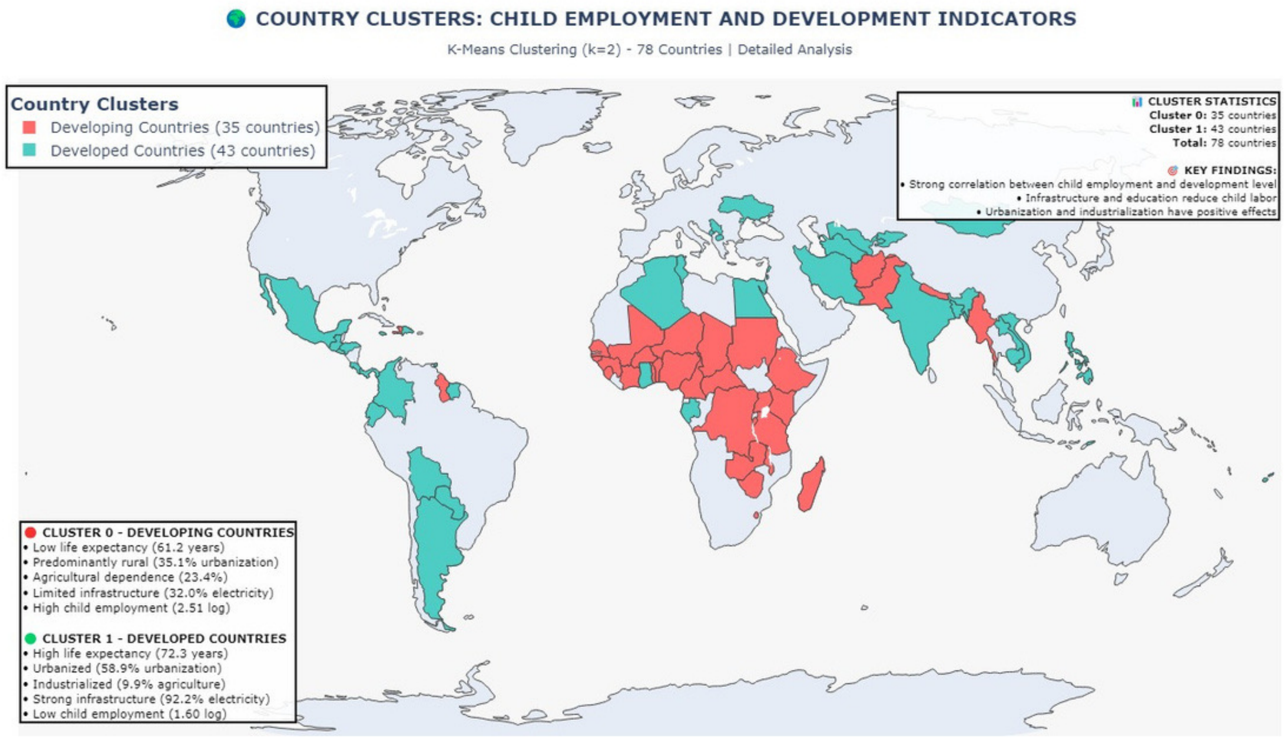


Figure: Country Clusters – Child Employment and Development Indicators

The K-Means clustering (k=2) clearly separates the world into two groups based on development and child labor indicators.
Cluster-0|Developing Countries: Countries in Sub-Saharan Africa, parts of South Asia, and the Middle East dominate this group. They show low life expectancy (≈61 years), rural dominance, agricultural dependence, and limited infrastructure, resulting in high child employment(≈2.5 log rate).

Cluster-1|Developed Countries: Found mostly in Europe, North America, and East Asia. These nations are urbanized (≈59%), industrialized, and enjoy strong infrastructure and education systems, corresponding to very low child employment (≈1.6 log rate).

Interpretation: The clustering confirms a strong inverse relationship between socioeconomic development and child labor. Education, industrialization, and urbanization act as key buffers against child employment—transforming economic progress into social protection.
Where electricity, education, and urban life expand, child labor fades. The divide on this map is not geographic — it's developmental.

References

Datasets Utilized

- Competiton Dataset
- International Labour Organization-ILOSTAT

  Children in Employment by Sex and Age(By Thousands)

https://ilostat.ilo.org/data/

Closure

For more details regarding methods, code snippets and more detailed graphics, please check out our project's repo on GitHub.

https://github.com/TanerYSLY/ASA-child-labor-data-analysis

For any inquiries regarding this project and related matters, please do not hesitate to contact us.

—Contact info have been provided in the cover page.—

Due to Competition Rules, the link provided above will be available after the deadline provided to us. The repo will be turned from private to public after 31/10/2025 23:59(Türkiye Time)

QuantaQuartet

Reclaiming Childhood

Unearthing the Roots of Child-Labour

Technical Presentation

Figure-1

Here, We used the median value per country to capture each nation's typical situation and reduce outlier effects.

```
df_country_median = (
    df.groupby(["country_code","country_name"], as_index=False)
      .agg({
          'child_employment_rate': 'median',
          'log_child_employment_rate': 'median',
          'log_gdp_per_capita_ppp': 'median',
          'agriculture_value_added_pct': 'median',
          'primary_school_enrollment': 'median',
          'life_expectancy': 'median',
          'urban_population_pct': 'median',
          'rural_electricity_access': 'median',
          'child_population_pct': 'median',
          'youth_unemployment_rate': 'median',
          'log_private_consumption_per_capita': 'median',
          'governance_index': 'median'
      })
) #✅ Purpose: To reduce random fluctuations in the time series and represent the c
  #level.


#  Interaction terms

# Does governance combined with GDP per capita impact child employment?
df_country_median['governance_x_gdp'] = (df_country_median['governance_index'] *
                                         df_country_median['log_gdp_per_capita_ppp']
# Do urbanization and agriculture jointly influence child employment?
df_country_median['urban_x_agriculture'] = (df_country_median['urban_population_pct'
                                            df_country_median['agriculture_value_adc

# Does education combined with governance impact child employment?
df_country_median['education_x_governance'] = (df_country_median['primary_school_enr
                                               df_country_median['governance_index']

base_features = [
    'log_gdp_per_capita_ppp', 'agriculture_value_added_pct', 'primary_school_enrollm
    'life_expectancy', 'urban_population_pct', 'rural_electricity_access', 'child_pc
    'youth_unemployment_rate', 'log_private_consumption_per_capita', 'governance_inc
    'governance_x_gdp', 'urban_x_agriculture', 'education_x_governance'
]
```

Figure-2

Within this part, Ridge, Lasso, ElasticNet, and Polynomial Ridge methods were used instead of classical regression (OLS).

```
================================================================================
                              DATA PREPARATION
================================================================================

Sample (n): 58
Number of features (p): 13
n/p ratio: 4.46 (ideal: >10)

✅ All features standardized (mean=0, std=1)

================================================================================
                        ⚠️ MULTICOLLINEARITY ANALYSIS (VIF)
================================================================================

VIF Interpretation: <5 (good), 5-10 (moderate), >10 (bad), >100 (catastrophic)

                          Feature         VIF
                 governance_index  209.185369
                 governance_x_gdp  138.017937
          education_x_governance  118.959618
             log_gdp_per_capita_ppp   17.058354
log_private_consumption_per_capita   12.556855
        agriculture_value_added_pct    8.955503
            rural_electricity_access    7.901274
                child_population_pct    6.952630
                urban_population_pct    6.823608
                     life_expectancy    6.053609
                  urban_x_agriculture    4.591081
              youth_unemployment_rate    1.584173
            primary_school_enrollment    1.361424
```

Figure-3

```
lasso.fit(X_scaled, y)
selected_features = X.columns[lasso.coef_ != 0].tolist()
print(f"Lasso selection ({len(selected_features)} features)::")
print(selected_features)


X_reduced = X_scaled[selected_features]
ridge_reduced = RidgeCV(alphas=alphas, cv=cv)
ridge_reduced.fit(X_reduced, y)
cv_r2_reduced = cross_val_score(ridge_reduced, X_reduced, y, cv=cv, scoring='r2').mean()
print(f"Reduced model CV R²: {cv_r2_reduced:.4f}")
```

✓ 16.6s

```
Lasso selection (4 features)::
['agriculture_value_added_pct', 'life_expectancy', 'urban_population_pct', 'rural_electricity_access']
Reduced model CV R²: 0.2886
```

In this part, We used Lasso for variable selection.

Figure-4

Here, we re-filtered the data

```
df = pd.read_csv('analyze data.csv', index_col=0)

df_country_median = (
    df.groupby(["country_code","country_name"], as_index=False)
      .agg({
          'log_child_employment_rate': 'median',
          'agriculture_value_added_pct': 'median',
          'life_expectancy': 'median',
          'urban_population_pct': 'median',
          'rural_electricity_access': 'median',
      })
)
base_features = ['agriculture_value_added_pct',
          'life_expectancy',
          'urban_population_pct',
          'rural_electricity_access',]
target = "log_child_employment_rate"
df_clean = df_country_median.dropna(subset=[target]+ base_features)
```

Figure-5

```
================================================================================
                              DATA PREPARATION
================================================================================

Sample (n): 78
Number of features (p): 4
n/p ratio: 19.50 (ideal: >10)

✅ All features standardized (mean=0, std=1)

================================================================================
                       ⚠️ MULTICOLLINEARITY ANALYSIS (VIF)
================================================================================

VIF Interpretation: <5 (good), 5-10 (moderate), >10 (bad), >100 (catastrophic)

                  Feature      VIF
          life_expectancy 3.639950
   rural_electricity_access 2.986257
agriculture_value_added_pct 2.514534
       urban_population_pct 1.800986
```

Here the n/p ratio improved, and multicollinearity was eliminated

Figure-6

After the model comparison, Ridge performed the best.

```
==========================================================================================
🤖 MODEL TRAINING
==========================================================================================

Running 5-fold CV...

✓ OLS completed
✓ Ridge completed
✓ Lasso completed
✓ ElasticNet completed
✓ Polynomial Ridge completed

==========================================================================================
📊 MODEL COMPARISON
==========================================================================================
           Model  Train_R²    CV_R²  CV_RMSE  Bias_Var_Gap Overfitting
           Ridge  0.393773 0.152979 0.740115      0.240794        High
      ElasticNet  0.399229 0.146451 0.741209      0.252778        High
           Lasso  0.404536 0.119018 0.747182      0.285517        High
             OLS  0.404653 0.116223 0.745828      0.288430        High
Polynomial Ridge  0.417248 0.053076 0.773474      0.364172        High

==========================================================================================
🏆 BEST MODEL: Ridge (CV R² = 0.1530)
==========================================================================================
```

Figure-7

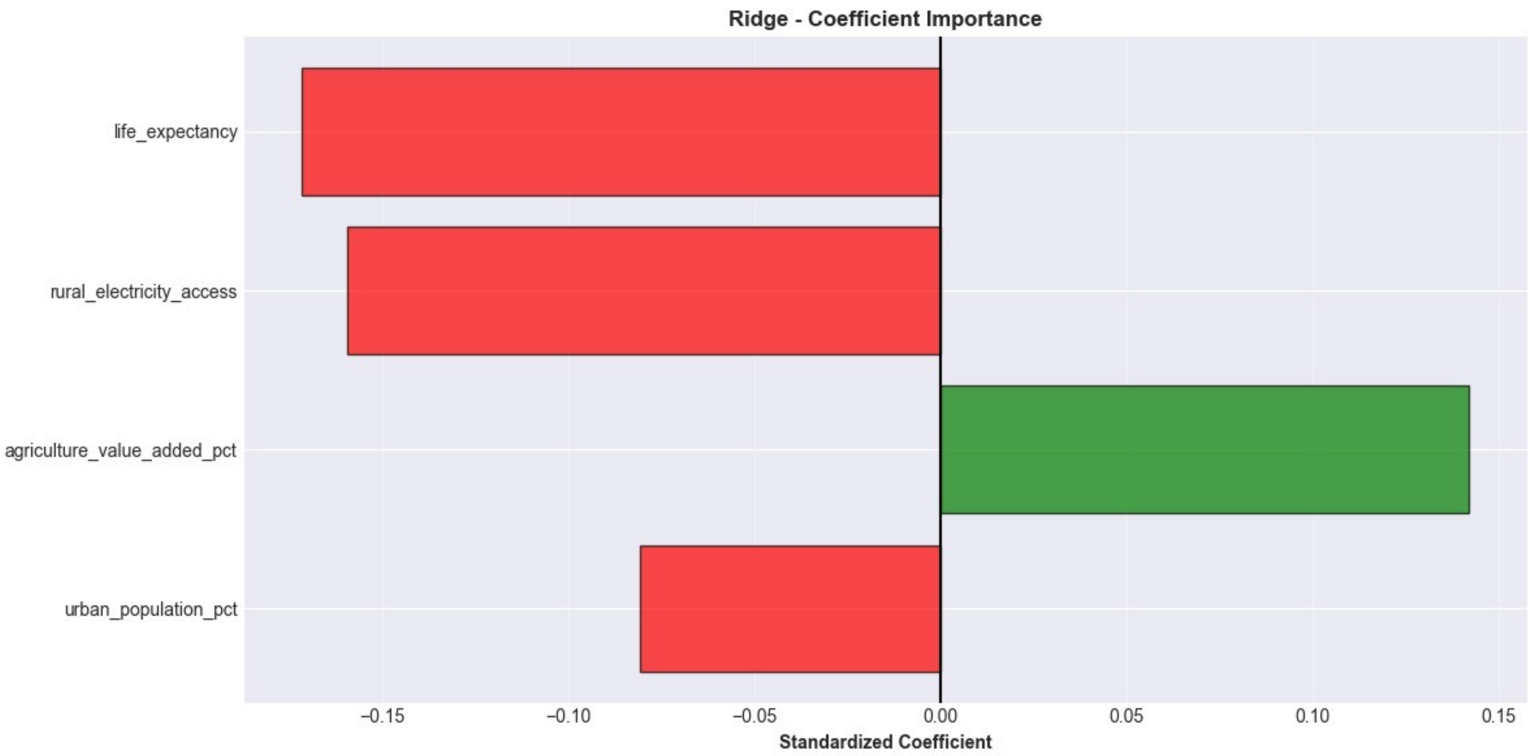
The Ridge model retained all variables while shrinking their coefficients and demonstrated the most balanced performance.

```
================================================================
                                 🔍 DETAILED ANALYSIS: Ridge
================================================================

📌 Ridge Regularization:
   • Optimal α: 40.3702
   • All 4 features used (coefficients shrunk)

📊 Performance:
   • Training R²: 0.3938
   • CV R²: 0.1530
   • RMSE: 0.6947
```

Figure-8



Ridge - Coefficient Importance
life_expectancy
rural_electricity_access
agriculture_value_added_pct
urban_population_pct
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Standardized Coefficient

Suggestions for Sustainable Development

Our regressions shows that child labour is mainly driven by:

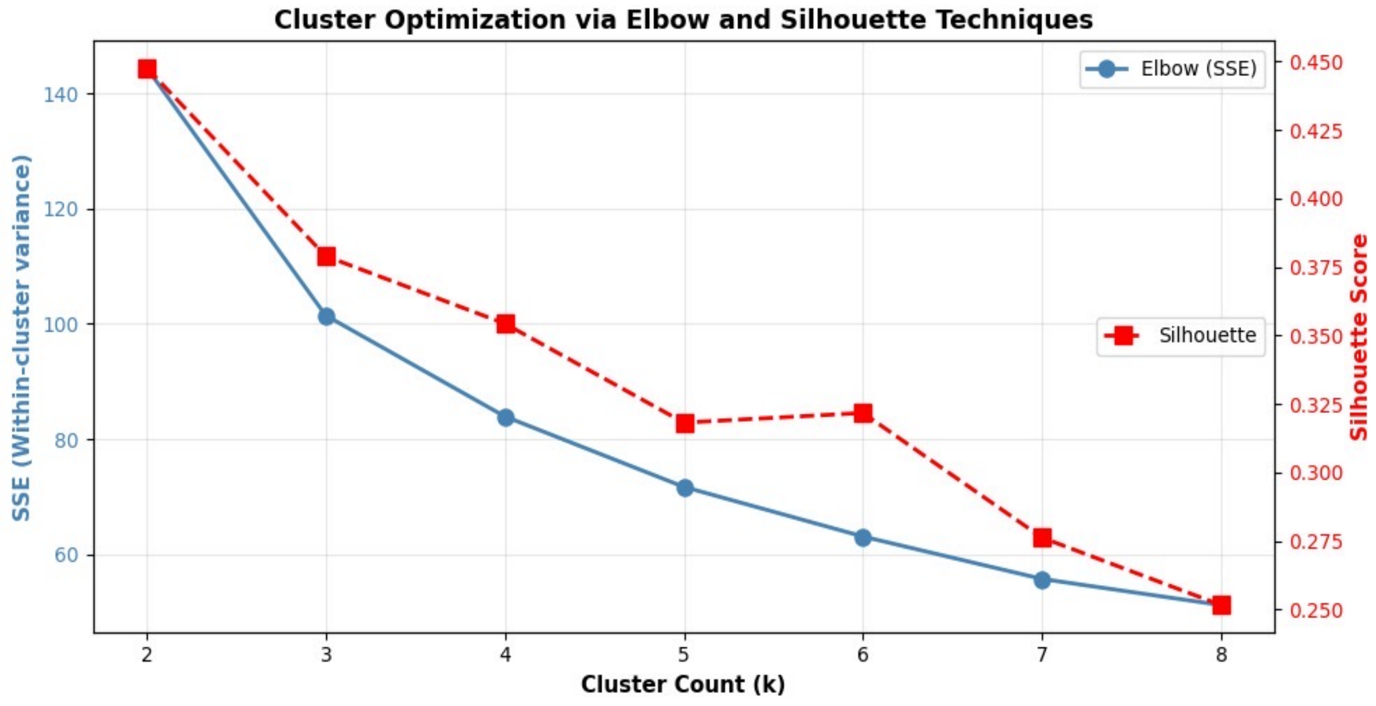
- low health
- weak infrastructure
- agricultural dependence

and all three are rooted in governance and development inequality.

Clustering & Validity Test

In this part, after identification of factors which influence child-labour, we used clustering to see how countries group together based on couple indicators.

Figure-9



Cluster Optimization via Elbow and Silhouette Techniques
Elbow (SSE)
Silhouette
SSE (Within-cluster variance)
Silhouette Score
Cluster Count (k)
140
120
100
80
60
0.450
0.425
0.400
0.375
0.350
0.325
0.300
0.275
0.250
2
3
4
5
6
7
8

Figure-10

Here, we used K-Means model for clustering.

```python
# 4️⃣ Final K-Means model
kmeans = KMeans(n_clusters=best_k, random_state=42, n_init=10)
clusters = kmeans.fit_predict(X_scaled)

# Cluster bilgilerini dataframe'e ekle
df_clusters = X.copy()
df_clusters['cluster'] = clusters

print(f"\n✅ The K-Means model was fitted")
print(f"   • Cluster Count: {best_k}")
print(f"   • Total Country: {len(df_clusters)}")
print(f"\n📊 Cluster distribution:")
print(df_clusters['cluster'].value_counts().sort_index())
```

✓ 0.0s

```
✅ The K-Means model was fitted
   • Cluster Count: 2
   • Total Country: 78

📊 Cluster distribution:
cluster
0    35
1    43
Name: count, dtype: int64
```

Figure-11

We used a one-way ANOVA test to check if the two clusters are statistically different.

```python
# 6️⃣ One-way ANOVA

print("\n" + "="*90)
print(" "*30 + "📊 ANOVA TEST")
print("="*90)
print("\n Feature-wise differences across clusters\n")

for col in features:
    groups = [df_clusters[df_clusters['cluster']==i][col] for i in range(best_k)]
    f_stat, p_val = f_oneway(*groups)
    significance = "✅ Statistically significant" if p_val < 0.05 else "⚠️  Statistically insignificant"
    print(f"{col:.<50} F={f_stat:>6.2f}, p={p_val:.4f} {significance}")

print("\n  p < 0.05 → Clusters show a statistically significant difference in this feature.")
print("  High F-statistic → Large variance between clusters. ")
```

✓ 0.0s

```
==========================================================================================
                              📊 ANOVA TEST
==========================================================================================

 Feature-wise differences across clusters

agriculture_value_added_pct....................... F= 86.22, p=0.0000 ✅ Statistically significant
life_expectancy................................... F=117.05, p=0.0000 ✅ Statistically significant
urban_population_pct.............................. F= 42.72, p=0.0000 ✅ Statistically significant
rural_electricity_access.......................... F=141.28, p=0.0000 ✅ Statistically significant

  p < 0.05 → Clusters show a statistically significant difference in this feature.
  High F-statistic → Large variance between clusters.
```