UNB

University of New Brunswick

CS4403

Interim Report

Instructor: Jake Van Der Laan

Author

Yasin Rahman

ID: 3733078

1 Introduction

This interim report provides an update on the progress made toward analyzing global education trends using data obtained from Kaggle. The study aims to investigate relationships between educational attainment, employment levels, birth rates, and gender disparities.
Link: Kaggle

2 Summary of Collected Data

The dataset used for this project is a global education dataset obtained from Kaggle. It contains data from 202 different locations and nations worldwide and provides a comprehensive overview of educational trends and their socioeconomic impacts. The dataset has been structured to allow for meaningful analysis of key educational indicators and their relationships with various economic and demographic factors.

Data Structure

- **Number of Rows (Instances):** 202 (Each representing a country or region)
- **Number of Columns (Attributes):** 29

Types of Attributes and Their Importance

The dataset includes 29 attributes covering different aspects of education and socioeconomic factors. Some of the key features include:

Educational Indicators

- **Literacy Rate (%):** Measures the percentage of people who can read and write.
- **Primary, Secondary, and Tertiary School Enrollment Rates (%):** Tracks the proportion of students enrolled at different educational levels.
- **Average Years of Schooling:** Represents the average number of years spent in formal education.

- **Completion Rates:** Measures the percentage of students completing primary, secondary, or higher education.

Socioeconomic Indicators

- **GDP per Capita ($):** Measures the economic output per person, providing an economic context to education levels.
- **Employment Rate (%):** Assesses the percentage of the population that is employed.
- **Unemployment Rate (%):** Evaluates the portion of the workforce that is jobless.
- **Poverty Rate (%):** Provides insights into how education correlates with economic hardships.

Demographic Factors

- **Population Size:** Determines the total number of people within each region.
- **Gender-Based Educational Data:** Compares male vs. female literacy rates, school enrollment, and educational attainment.
- **Birth Rates (Per 1,000 People):** Helps analyze the impact of education on population growth and family planning trends.

Initial Observations from Data Collection

- The dataset contains a wide variation in literacy rates, school enrollments, and GDP per capita across different countries.
- There are missing values in some fields, particularly in regions with less reliable data collection mechanisms.
- Some variables show strong correlations, such as GDP per capita and school enrollment rates.
- There are noticeable gender disparities in education, with some regions exhibiting lower female literacy and school completion rates.

- The dataset is relatively clean but required some preprocessing to standardize formats and handle missing data.

This dataset is large enough to support advanced statistical analysis and machine learning models while avoiding excessive risk of overfitting. The information contained within it has strong potential to reveal patterns that can inform educational policy-making and socioeconomic planning.

3 Progress Toward Project Goals

Significant progress has been made in processing and analyzing the dataset to extract meaningful insights related to global education trends and their socioeconomic impacts. This section outlines the key steps taken so far, including data preprocessing, exploratory data analysis (EDA), and the application of initial data mining techniques.

Data Preprocessing

Before conducting any analysis, the dataset was carefully preprocessed to ensure consistency, accuracy, and completeness. The following steps were performed:

Handling Missing Values

- Some attributes, particularly those related to **education completion rates** and **socioeconomic indicators**, contained missing values.
- Missing numerical data was addressed using **mean imputation**, where missing values were replaced with the average value for that attribute.
- Records with excessive missing data that could not be reliably imputed were **removed** to maintain data integrity.

Standardization of Numerical Values

- Since attributes such as **GDP per capita, literacy rates, and school enrollment rates** were on different scales, numerical values were **standardized** to bring them to a common range.

- Standardization helps ensure that models relying on **distance-based metrics** (e.g., clustering) are not biased by attributes with larger numerical ranges.

Encoding Categorical Variables

- Some categorical attributes, such as **region classification**, were **encoded into numerical values** to facilitate analysis.
- **One-hot encoding** was applied where necessary to preserve categorical information without introducing unintended ordinal relationships.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to better understand the dataset, detect patterns, and identify potential relationships between variables.

Descriptive Statistics

- Summary statistics (**mean, median, standard deviation**) were computed for each numerical attribute to assess their distributions.
- The analysis highlighted key disparities in education levels between different **countries and regions**.

Data Visualization Techniques

- **Histograms and Density Plots:** Used to examine the distribution of numerical variables such as **literacy rates and GDP per capita** across different regions.
- **Scatter Plots:** Used to visualize relationships between **education indicators and economic factors** (e.g., how school enrollment correlates with GDP per capita).
- **Correlation Heatmaps:** Helped identify **strong correlations between variables**, such as the link between **higher GDP per capita and improved literacy rates**.

Outlier Detection

- Outliers were identified in **literacy rates, education completion rates, and GDP per capita**.
- Some countries displayed **extremely high or low values** compared to the global average, potentially indicating either exceptional success in education policies or severe educational disparities.
- The decision on handling outliers (**removal or transformation**) is still being evaluated.

4 Data Mining strategies and Findings

Mnay data mining techniques have been explored to uncover trends and relationships in the dataset. These techniques have helped in identifying key factors influencing education and socioeconomic conditions worldwide. The main strategies applied so far include correlation analysis, clustering using K-Means, regression modeling, decision trees, principal component analysis (PCA), and association rule mining.

Correlation Analysis

- A **strong positive correlation** was observed between **GDP per capita** and key education indicators such as **literacy rates and school enrollment**.
- Countries with **higher GDP per capita** tend to have **higher primary, secondary, and tertiary school enrollment rates**.
- Gender disparities in education were evident in certain regions, where **female literacy rates were consistently lower than male literacy rates**.
- These findings confirm that economic prosperity plays a crucial role in educational access and attainment, and that gender-based educational disparities remain a significant issue in certain regions.

Clustering (K-Means Clustering)

- **K-Means clustering** was applied to group countries based on their **education and socioeconomic characteristics**.
- The dataset was segmented into **three major clusters** representing different levels of educational development:
  - **High Literacy, High GDP Countries:** North America, Western Europe
  - **Moderate Literacy, Developing Economies:** South America, Southeast Asia
  - **Low Literacy, Low GDP Nations:** Sub-Saharan Africa
- The clustering analysis helped identify **regions where educational interventions may be most necessary**.
- The results suggest that countries in the **Low Literacy, Low GDP cluster** would benefit the most from targeted education programs and policies.

Regression Models

- **Linear regression models** were tested to examine the relationship between **school enrollment rates** and **employment levels**.
- Preliminary results indicate that **higher school enrollment rates correlate with lower unemployment rates**.
- This suggests that **education plays a key role in improving job prospects**, supporting the hypothesis that increased educational attainment leads to better employment opportunities.
- Future work will refine these models by incorporating **additional socioeconomic variables** to improve prediction accuracy.

Decision Trees and Random Forests

- A **decision tree model** was used to determine the most important factors influencing **literacy rates and education completion rates**.

- **Employment rates, GDP per capita, and school enrollment rates** were identified as the most significant predictors of literacy.
- A **random forest model** improved accuracy by reducing overfitting and confirmed that economic stability and gender equity in education are strong predictors of literacy improvements.
- These models reinforce the importance of financial and social policies in promoting education.

Principal Component Analysis (PCA)

- Since the dataset contains **29 features**, **PCA** was used to reduce dimensionality and highlight the most influential variables in education trends.
- The analysis revealed that **three principal components** explain the majority of variance in the dataset:
  1. Economic Factors (**GDP per capita, employment rate, and poverty rate**)
  2. Education Access (**school enrollment rates and literacy rates**)
  3. Gender Disparities (**male vs. female literacy and school completion rates**)
- PCA helps streamline predictive modeling by reducing redundancy while preserving key information.

Association Rule Mining

- **Apriori algorithm** was applied to find hidden relationships between **education, employment, and economic indicators**.
- Results indicated that:
  - **Countries with high literacy rates (above 90%)** also tend to have **high employment rates (above 70%)**, with a confidence level of 85%.

  - **Low school enrollment rates (below 50%)** are strongly associated with **high poverty rates (above 40%)**, suggesting an urgent need for educational intervention.
  - A strong association was found between **low female literacy rates and low GDP per capita**, reinforcing the role of gender equality in economic growth.

- These insights can be used to shape policy recommendations aimed at reducing gender disparities and improving economic conditions through education.

Summary of Data Mining Strategies and Results

The data mining strategies explored so far have provided key insights into global education trends:

- **Correlation analysis** has confirmed the **strong relationship between education levels and GDP per capita**, highlighting the economic benefits of higher education.

- **K-Means clustering** has successfully categorized countries into distinct **education-development groups**, helping to identify regions where educational reforms are most needed.

- **Regression modeling** has provided evidence that **higher school enrollment rates are associated with lower unemployment levels**, emphasizing the role of education in workforce development.

- **Decision trees and random forests** identified **employment rates, GDP per capita, and school enrollment rates** as the most important predictors of literacy.

- **PCA** successfully reduced the dataset to **three principal components**, helping focus the analysis on key economic, educational, and gender-based disparities.

- **Association rule mining** uncovered hidden patterns, showing that **education, employment, and poverty are deeply interconnected**.

These results have laid a strong foundation for further analysis, including testing advanced machine learning models and validating findings with real-world case studies.

5 Proposed Strategies for Further Exploration

- Apply machine learning models such as Random Forest and Support Vector Machines (SVM).
- Conduct hypothesis testing to confirm statistical significance of relationships.
- Explore geospatial visualizations to identify regional education patterns.

6 Remaining Work and Timeline

Tasks to Complete

- Optimize predictive models for accuracy.
- Finalize data visualizations for report presentation.
- Interpret model results and compare different approaches.
- Draft final report and validate findings.

Project Timeline

- **March 5 - March 15:** Model Refinement
- **March 18 - March 24:** Interpretation of Results
- **March 25 - March 31:** Final Adjustments
- **April 1 - April 7:** Final Testing
- **April 8, 2025:** Project Submission

7 Conclusion

There has been significant advancement in data preprocessing, exploratory analysis, and preliminary model testing. The data analysis illustrates fundamental relationships between education and socioeconomic factors, such as GDP per capita, employment rates, and gender disparities. The plan for the future is to develop predictive models that are more refined, validate results, and finalize the project submission.