CxC Report

Challenge: SAP

Group Members: Kira, Sharanya, Angel, Gracie

Introduction

Poverty is one of the most pressing global challenges, affecting billions of people across different regions and socioeconomic contexts. While traditional poverty measures often focus solely on income, there are many other factors that contribute to it, such as access to electricity, education, and healthcare. Our goal is to study the relationships between variables to create a Multidimensional Index that reflects key aspects at the country-level. By doing so, we hope to uncover the most influential factors affecting poverty and propose targeted policy recommendations to drive meaningful change.

The Data

The dataset consists of 23,141 rows and 32 columns, with a mix of categorical and numerical data. The first few columns, such as "Country Name," "Country Code," "Indicator Name," and "Indicator Code," provide metadata about each data entry, indicating that the dataset contains multiple indicators for each country. This suggests that the "Country Name" column is repeated for different indicators rather than representing unique rows. Additionally, there are missing values in several columns, most notably in the "short description" and the yearly data columns. The missing values in the time series data (from 2000 to 2023) indicate that some indicators or countries lack recorded values for specific years, which must be addressed through imputation, removal, or other data-cleaning techniques.

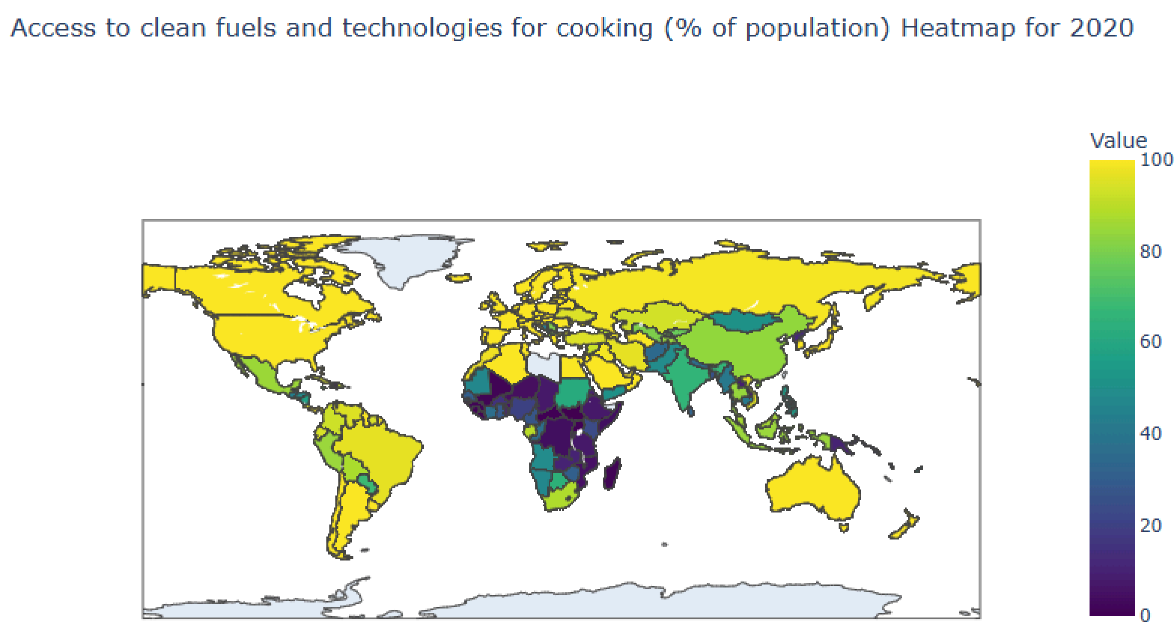
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23141 entries, 0 to 23140
Data columns (total 32 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Country Name       23141 non-null  object 
 1   Country Code       23141 non-null  object 
 2   Indicator Name     23141 non-null  object 
 3   Topic              23141 non-null  object 
 4   short description  1594 non-null   object 
 5   long description   22875 non-null  object 
 6   Indicator Code     23141 non-null  object 
 7   Unit of measure    23141 non-null  object 
 8   2000               11376 non-null  float64
 9   2001               9576 non-null   float64
 10  2002               10009 non-null  float64
 11  2003               10012 non-null  float64
 12  2004               10229 non-null  float64
 13  2005               10798 non-null  float64
 14  2006               10662 non-null  float64
 15  2007               10542 non-null  float64
 16  2008               10641 non-null  float64
 17  2009               10815 non-null  float64
 18  2010               12255 non-null  float64
 19  2011               11313 non-null  float64
 20  2012               11455 non-null  float64
 21  2013               11078 non-null  float64
 22  2014               11355 non-null  float64
 23  2015               12132 non-null  float64
 24  2016               11038 non-null  float64
 25  2017               10838 non-null  float64
 26  2018               10889 non-null  float64
 27  2019               11462 non-null  float64
 28  2020               10434 non-null  float64
 29  2021               9974 non-null   float64
 30  2022               8455 non-null   float64
 31  2023               1965 non-null   float64
dtypes: float64(24), object(8)
memory usage: 5.6+ MB
```

Data Cleaning

Our imputation approach was extensive. We wanted to make sure we dealt with null values appropriately while also making sure the data was in a suitable, rectangular form for visualization and further aggregations if needed. To begin, we manually dropped columns that were empty, redundant and/or ambiguous. These included ‘short description, ‘long-description’, ‘indicator-code and unit of measure’’. Next, we pivoted the variables so that all countries in the country column were unique, providing a unique index to the data. The rest of our data cleaning approach can be found in the attached 'Data_Cleaning.ipynb' file in our submission. This file further details our imputation process.

EDA

Before we began the modelling process, we wanted to get a sense of which indicators may be worth investigating further, so we made a geospatial heat map that uncovered the trends in the data overtime. This animated plot shows how the "Access to clean fuels and technologies" indicator changes overtime

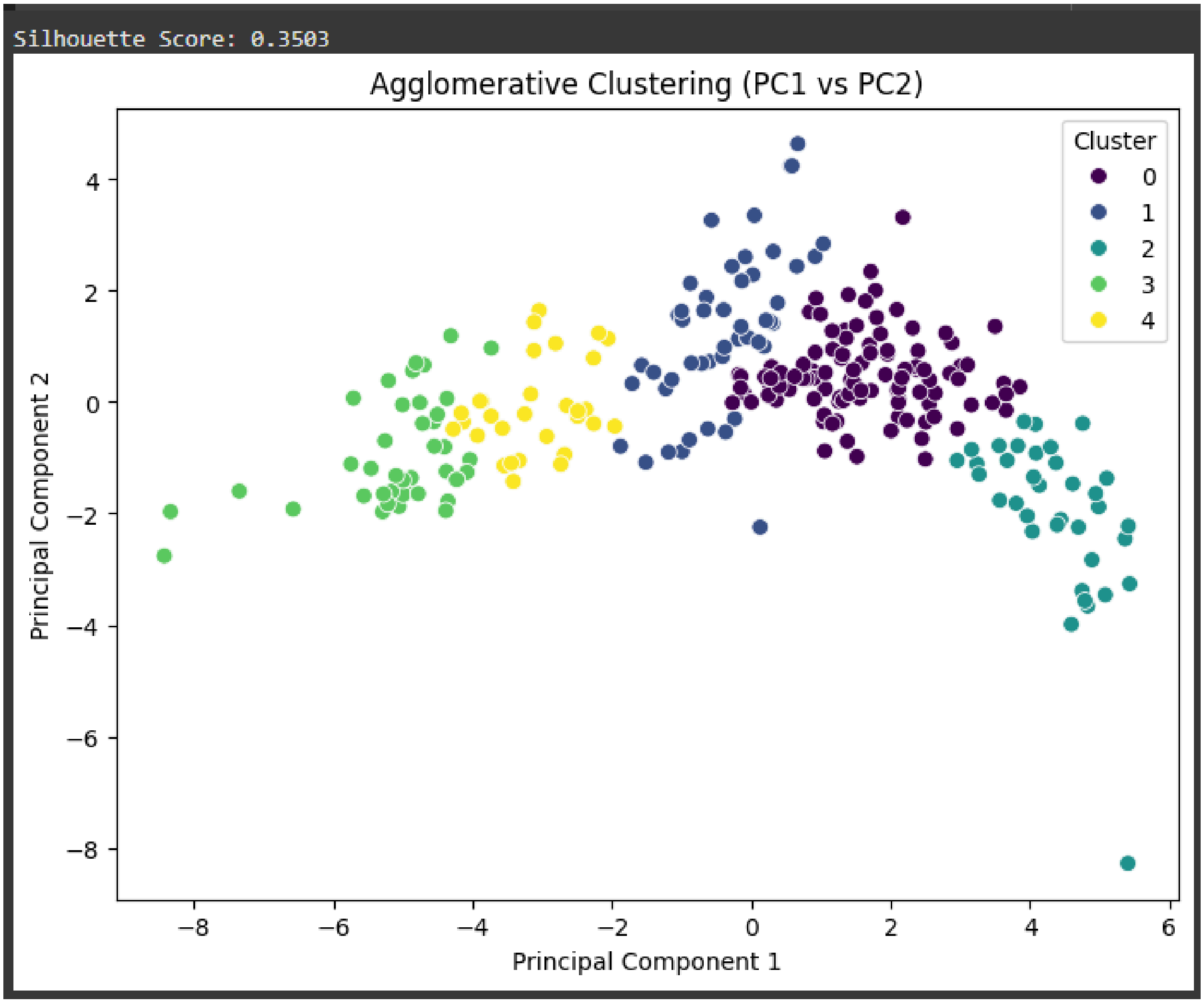


Methodology

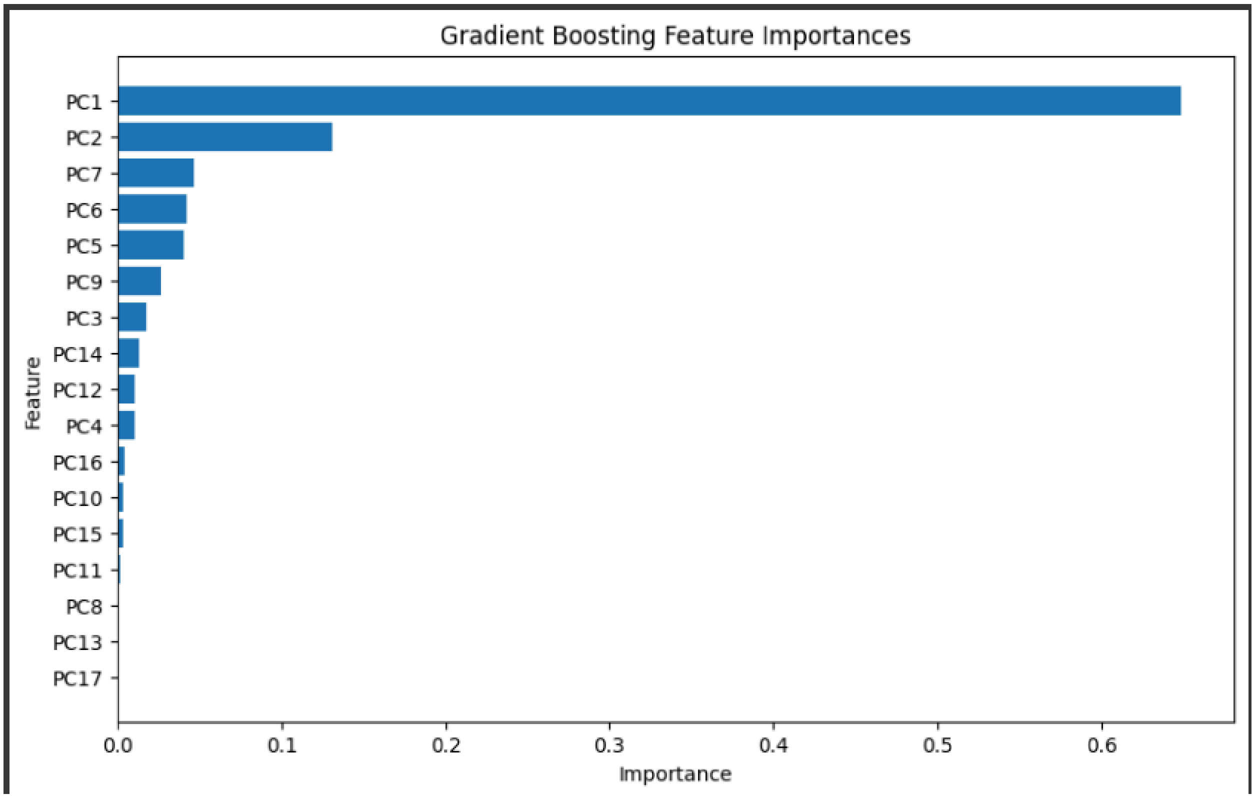
Approach 1

Our first approach implemented Principal Component Analysis (PCA) to reduce the dimensionality of our dataset while retaining the most informative variables. We ended up with 17 principal components in the data.

Next, we performed agglomerative clustering with the principal components we found. From clustering, we found that there are. 5 clusters in the data which correspond to labels that can be used in our index.



Initially, we intended to use Random Forest to analyze feature importance to further refine the indicators for building the index, but this approach told us that each of the principal 17 components were equally important, so we scrapped that idea. As an alternative, we implemented XGBoost for feature selection, which provided better performance in identifying the most relevant indicators.



Gradient Boosting told us that principal component 1 was most important

Analysis

Upon applying our two approaches, we compared the resulting indices across different countries to identify trends and disparities. Some key insights from our analysis include. Countries with high electricity access and education rates consistently scored lower on the poverty index, indicating a strong correlation between these factors and economic well-being. Healthcare availability emerged as a critical determinant, with countries lacking adequate healthcare infrastructure showing significantly higher poverty scores.Temporal analysis revealed that certain countries have improved their conditions over the years, while others have stagnated or worsened, highlighting the need for targeted interventions.

Policy Recommendations

Based on our findings, we propose the following policy recommendations to effectively address poverty at a structural level:

1. Countries with higher literacy rates and access to quality education tend to have lower poverty levels. Governments should prioritize free and mandatory education, vocational training, and digital literacy programs.
2. Ensuring widespread access to electricity, clean water, and sanitation can significantly improve living conditions and economic productivity.
3. Supporting small businesses, providing microfinance opportunities, and ensuring fair labor policies can uplift economically vulnerable populations.