PovertyEmbedding: Utilizing Latent Embeddings of Wikipedia Articles to Predict Poverty

Team Members:
Bhuvan Hebbar (PES2UG23CS129)
Bikram Dutta (PES2UG23CS132)

---

1. Problem Statement

PovertyEmbedding aims to predict country-level poverty rates by leveraging **semantic information extracted from Wikipedia country summaries**.
While additional economic indicators such as GDP per capita and multidimensional poverty metrics (MPI) are incorporated to enhance the model's precision, the **core novelty** of this project lies in the **use of latent textual embeddings** generated from Wikipedia data.

These embeddings capture a country's socio-economic, historical, and geographic context, allowing the model to interpret qualitative information and combine it with quantitative metrics for improved poverty prediction.

---

2. Project Implementation and Methodology

The methodology was structured into three main phases:

1. **Data Engineering**
2. **Exploratory Data Analysis (EDA) & Feature Selection**
3. **Model Training & Evaluation**

This pipeline integrates natural-language-based semantic embeddings with structured socio-economic data.

2.1 Data Engineering

A. Wikipedia Country Summary Collection

- **Implementation:**
  Using `pycountry` and `wikipedia`, summaries (≤800 characters) were collected for all valid countries.
  Errors such as `PageError` and `DisambiguationError` were handled gracefully.
  The dataset was saved as `wikipedia_countries.csv`.

B. Wikipedia Country Embedding Generation

- **Implementation:**
  The script encoded summaries using **SentenceTransformer (`all-MiniLM-L6-v2`)**, creating **384-dimensional latent embeddings** representing each country's semantic context.
  These were stored in `wikipedia_country_embeddings.csv`.

C. Integration with Economic and Poverty Data

Wikipedia embeddings were merged with:

- **World Bank Poverty Rate (2018)** — target variable
- **GDP per Capita** — economic baseline feature
- **Multidimensional Poverty Index (MPI)** — supplementary indicators

All merges used inner joins to maintain consistency, yielding 67–96 usable country entries.

---

3. Data Insights and Feature Selection

Why Wikipedia Embeddings Are Central

- Wikipedia summaries embed **social, political, and economic cues** unavailable in numeric datasets.
- Embedding dimensions such as 161 (corr = 0.544) and 113 (corr = 0.506) correlated with MPI poverty rates — showing that purely textual latent features carry significant

predictive power.

- These embeddings formed the **main input feature space**, supported by GDP and MPI values for grounding.

Feature Selection

- Standardized all numerical features using **StandardScaler**.
- Retained top **73 features** (embeddings + numeric indicators) based on correlation with the target.
- Split into **train (80%)** and **test (20%)** datasets (`random_state = 42`).

---

4. Model Training and Evaluation

4.1 Randomized Search Results

Model	Best Parameters	Best Cross-Validation R^2
Linear Regression	{}	0.5621
Ridge Regression	{'alpha': 14.0494}	0.8227
ElasticNet	{'alpha': 0.3541, 'l1_ratio': 0.1079}	0.7606
SVR	{'C': 9.3187, 'epsilon': 0.0542, 'kernel': 'linear'}	0.8042

4.2 Final Test Performance

Model	Test R^2	Test MAE
Ridge Regression	**0.8824**	**0.0297**
ElasticNet	0.8000	0.0410
SVR	0.7862	0.0375
Linear Regression	−0.5434	0.0989

Best Model:
Ridge Regression (α = 14.05)
R^2 = 0.8824, MAE = 0.0297

Interpretation: The model explains 88% of variance in MPI-derived poverty rates with an average prediction error of just 0.03.

---

5. Error Analysis

- **Residuals:** Small and centered around 0 — indicates low bias.
- **Residual vs Predicted Plot:** Random scatter shows no systematic over- or under-prediction.
- **Max Residual:** 0.0325 (Burkina Faso) → indicates strong accuracy across all samples.

---

6. Project Alignment Verification

The final notebook (`ml_mini_project_129_132.ipynb`) fully satisfies the objective:

> **"Utilizing Latent Embeddings of Wikipedia Articles to Predict Poverty."**

Evidence of alignment:

- Wikipedia text was the **primary data source**, forming the **main latent feature set** through 384-dimensional embeddings.
- GDP and MPI were **secondary** structured features used to stabilize and contextualize predictions.
- Ridge Regression achieved strong performance, demonstrating that **semantic signals from text embeddings** can effectively model socio-economic phenomena such as poverty.

Conclusion:
The project's core achievement lies in transforming qualitative Wikipedia narratives into quantitative representations that accurately predict poverty — showing that **text-driven embeddings** can serve as powerful proxies for complex socio-economic patterns.
