

Data Collection and Preprocessing Phase

Date	16 June 2025
Team ID	SWTID1749653449
Project Title	Economic Growth: A Machine Learning Approach to GDP per Capita Prediction
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report summarizes the identified issues within the selected dataset sourced from **Kaggle and UCI repositories**. The report highlights severity levels of the problems and outlines plans for resolution. It plays a vital role in ensuring accurate model training and trustworthy predictions by addressing data inconsistencies, missing values, and formatting issues..

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Missing values in Net migration, GDP (\$ per capita), Birthrate, Deathrate	Moderate	Use mean/median imputation or regression models
Kaggle Dataset	Missing values in Literacy (%), Phones (per 1000), Arable (%), Crops (%)	Moderate	Impute with mean or create categorical buckets
Kaggle Dataset	Missing values in Climate, Agriculture, Industry, Service	High	Use domain-informed estimations or drop rows

Kaggle Dataset	Categorical fields present (e.g., Region)	Low	Apply one-hot or label encoding
UCI Dataset	Mixed formats and inconsistent units in economic indicators	Moderate	Standardize units and normalize data
UCI Dataset	Outliers in GDP and population values	High	Apply log transformation and outlier treatment