Data Collection and Preprocessing Phase

Date	16 June 2024
Team ID	SWTID1749653449
Project Title	Economic Growth: A Machine Learning Approach to GDP per Capita Prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This machine learning project aims to predict GDP per capita using a diverse set of socioeconomic indicators such as population, literacy rate, net migration, industrial share, and more. The goal is to build a robust regression model that can assist economists and policymakers in forecasting economic performance at a country level.
Data Collection Plan	● Search for datasets related to global economic and demographic indicators. ● Prioritize datasets with recent, complete, and regionally diverse data. ● Ensure the inclusion of relevant variables like GDP, population, education, industry distribution, etc. ● Perform initial validation for data quality and completeness.

Raw Data Sources Identified	The raw data sources for this project included datasets obtained from Kaggle and UCI, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as gender, marital status, income, and loan-related details for machine learning analysis.

Source Name	Description	Location / URL	Format	Size	Access Permissions
Kaggle Dataset	Dataset includes global economic indicators such as GDP per capita, literacy rate, population, climate, etc.	https://www.kaggle.com/datasets/fernandol/countries-of-the-world	CSV	38.3 kB	Public
UCI Dataset	Contains detailed country-level demographic and economic features, suitable for GDP modeling and clustering analysis.	UCI - Country Data	CSV	30 kB	Public