# Integrating Open Data for GDP Prediction Leveraging ALITE and Machine Learning

## Team Information

* Madhu Mausam Thapa, [u1463636@utah.edu](mailto:u1463636@utah.edu), u1463636
* Sanjay Luitel, [u1559592@utah.edu](mailto:u1559592@utah.edu), u1559592
* Sushil Rijal, [u1323213@utah.edu](mailto:u1323213@utah.edu), u1323213

## Progress Till Date

### Data Acquisition

1. Extracted Public, Private and Public-Private-Partnership Capital stock (Primary feature) and GDP data (Target Variable) of 194 available countries from 1960 to 2019: Source: [IMF](https://infrastructuregovern.imf.org/content/dam/PIMA/Knowledge-Hub/dataset/IMFInvestmentandCapitalStockDataset2021.xlsx) .
2. Extracted Human Capital data (feature variable) like ( [Working Population](https://data.worldbank.org/indicator/SP.POP.1564.TO), [Unemployment,](https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS) Labor Force) from 1960 to 2063. Source: World Bank
3. Extracted Human Development Index. Source: [UNDP](https://hdr.undp.org/data-center/documentation-and-downloads)
4. Extracted Productivity and Technology Data (Energy use, Patent applications, R & D expenditure, Researchers in R & D). Source: World Bank
5. Demographics (Age- Dependency Ratio, Total Population). Source: World Bank
6. Productivity and other macro-economic indicators ([Penn World Table](https://pwt-data-tool.streamlit.app/))

### Data Wrangling

1. Data transformation:
   1. Converted the tables for wide to long format.
2. Imputation Strategy for datasets:
   1. Grouped by country and series.
   2. Interpolated values using linear method.
   3. Forward filled and backward filled to catch the edge NaNs.
   4. Remaining NaNs imputed using near zero values; as we are planning to use log transform later on.
3. Labor Force (1960- 1989)

We estimated the historical labor force using known working age population data.

1. First merged labor force and population dataset after calculating working age population from population dataset.
2. Trained a linear regression model using 1990-2023 data
3. Predicted labor force for 1960-1989
4. All data were processed with column standardized, missing values imputed, using custom linear regression or grouping and linear interpolation.

### Exploratory Data Analysis:

* Augmented Dickey-Fuller (ADF) and kwiatkowski-Phillips-Schimidt-Shin (KPSS) tests were conducted to test stationarity of GDP and used collected exploratory variables. Both ADF and KPSS tests agree that selected variables are stationary at 5% significance level. Since stationarity is confirmed, it is appropriate to proceed with regression analysis.
* A pairwise correlation test was conducted to examine the linear relationship between GDP and its explanatory variables. The results indicate that energy use and the number of researchers in R&D exhibit a weak positive correlation with GDP. Public-private investment shows a moderate positive correlation, while other variables demonstrate a strong positive linear relationship with GDP
* The Variance Inflation Factor (VIF) was calculated to assess multicollinearity among the explanatory variables. The results indicated that Public-Private Partnership (PPP) investment had a VIF greater than 10, which is considered unacceptable for regression analysis. Therefore, this variable was removed. After dropping the variable, the VIF test was conducted again, confirming that the multicollinearity criteria were now satisfied for further analysis.

### Analysis Methodology:

* Used Cobb Douglas equation to get the base model.
* Transformed and prepared data for linear regression analysis.
* Split the data into train and test (70/30)
* Trained the Linear Regression model (Multivariate) with K Cross-Validation, and hyperparameter tuning
* Trained the Random Forest model with K Cross-Validation, and hyperparameter tuning
* Conducted Feature importance analysis to assess the contribution of Capital stocks and other factors on GDP prediction.

## Remaining Features to be Implemented

#### Data:

1. Integrated datasets using Alite
2. Data Completeness: Evaluating the Added Value of Data Integration using ALITE
   1. Measure the percentage of missing values in manually-integrated data vs ALITE-integrated data
   2. Compute Pearson correlation for manually-integrated data vs. ALITE–integrated data
3. Scalability: Assessing ALITE’s Automation in Multi-Source Data Preparation
   1. Measure integration time manually vs. using ALITE.
   2. Compare manual operations required in traditional vs. alite integration.

#### Models:

* + - 1. Work with the Linear Regression and Random Forest model with data from Alite Pipeline
      2. Work on models like LSTM if time permits.

#### Comparison between Non-Alite and Alite pipeline:

With the models from two pipelines, compare how the results vary.