**World Population Growth Rate Project**

**Introduction**

This document provides an overview of the project, its goals, and the technologies used.

**Overview of Project**

The project aims to build a predictive model for estimating population growth rates using historical population data from various countries spanning the years 1960 to 2017. The dataset contains population counts for each country in each year, identified by a unique country code. The project involves data preprocessing, model training, and evaluation to predict population growth rates.

**Project Goal**

The goal of the project is to develop a simple model using decision trees that can accurately predict the population growth rate for a given country in a specific year based on historical data. This model can be useful in understanding population trends and planning future strategies related to population growth and development.

**Key Features of Project**

- Calculating Population Growth Rate
- Splitting data into training and testing sets
- Training a Decision tree model
- Assessing model performance using RMSLE calculation

**Tools Used**

- Python
- Jupyter notebooks/VS code/Google Collab
- Pandas and Sklearn library

**What I did**

- To calculate the population growth rate for a specific country starting from the year 1961 I created a function that first accesses the population data for a specified country_code using the loc method. It then iterates over the population data, starting from 1961, and calculates the growth rate for each year. For each country and year, the function creates a list containing the year and the corresponding growth rate (rounded to 5 decimal places for accuracy). These lists are then appended to a list of growth rates for the entire period and then finally converted into a 2-dimensional NumPy array.

- To divide the 2-dimensional NumPy array containing my population growth rate data into training and testing sets, I created a function that first extracts the years and growth rates from the 2-D array. It then creates two separate arrays for even and odd years, along with their corresponding growth rates by iterating over the years and growth rates and checking if the year is even or odd using the modulus operator (%). Once the data is divided into even and odd sets, the function returns two tuples: (X_train, y_train) and (X_test, y_test), where (X_train, y_train) contains the features and response variables for the training set (even years), and (X_test, y_test) contains the features and response variables for the testing set (odd years).

- For the third challenge of the project, I implemented a function to fit a model using the DecisionTreeRegressor class from the Sklearn library to the training data. The function creates a decision tree model with a specified maximum depth (MaxDepth) and fits the model to the training data (X_train, y_train) using the fit method. Finally, the trained model is returned as the output of my function.

- Finally, to assess model performance a function was created that uses the trained model to predict the population growth rates for the testing data (X_test) using the predict method. The predicted growth rates (y_pred) together with our testing data for the response (y-pred) are then used to calculate the Root Mean Squared Logarithmic Error (RMSLE) using the formula:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[log(1 + p_i) - log(1 + y_i)\right]^{2}}$$

## Conclusion

This project demonstrates the application of machine learning techniques to population data and highlights the potential for using predictive modeling to understand and forecast population trends.