World Population Analysis Project Report

Introduction

This project analyzes global population trends using historical data, with a goal to predict future population growth. By exploring population patterns, I identified factors that influence changes and utilizes Machine Learning for predictive insights. Understanding population trends is crucial for planning resources, healthcare, and infrastructure. This analysis provides insights into demographic shifts and their potential impact on a global scale.

Data Collection

The dataset used in this project is sourced from the United Nations World Population Prospects The initial dataset had a shape of **(234, 17)** with key features as follows:

- Country/Territory: Regions being analyzed.
- Continent: Geographical classification.
- Density: Population density.
- Growth Rate: Annual population growth.
- Rank: Population ranking globally.
- World Population %: Share of global population.
- Population figures: From 1970 to 2022

Data Preprocessing

1. Loading and Initial Data Analysis

- The dataset was loaded and analyzed the data types of each feature.
- Checked for missing values and confirmed there were no missing values in the dataset.
- Duplicate rows were checked and none were found.

2. Statistical Summary

- A summary of statistics revealed the distribution, mean, and spread of variables, providing valuable insights into central tendencies and ranges of key variables.

3. Feature Selection

- After an initial review, three columns were dropped: **Rank**, **CCA3**, and **Capital**. These columns were not expected to provide significant insights for EDA or model-building and were removed.
- After finding insights through EDA, the **Continent** column was dropped as it is not necessary for the model building process.

4. Feature Engineering

- **Population Change (1970-2022)**: This variable measures the absolute change in population from 1970 to 2022.
- **Average Annual Growth Rate**: Calculated as the average population growth per year over 52 years (from 1970 to 2022). This value indicates the yearly increase relative to the population in 1970.

5. Handling Categorical Variables

- The **Country/Territory** feature is a categorical variable. To prepare it for machine learning, label encoding was applied, converting each country/territory into a numerical format suitable for model training.

6. Scaling

- To improve the model's performance, all numerical data were scaled using **StandardScaler**. Standardization adjusts values to a common scale with a mean of 0 and standard deviation of 1, which is particularly important when features vary widely in range, as it helps the model converge more effectively and improves interpretability.

Exploratory Data Analysis (EDA)

1. Correlation Analysis

- A **correlation heatmap** was created to examine the relationships between numerical features. The heatmap helped identify the strength and direction of correlations, which is crucial for selecting and understanding features that may impact the model performance.

2. Population by Continent (2022)

- The data was grouped by continent to analyze the total population in 2022. This analysis revealed that **Asia** had the highest population, followed by **Africa**, **Europe**, **North America**, **South America**, and **Oceania**.

3. Top 10 Most Populated Countries (2022)

- The top 10 countries with the largest populations were identified. **China** and **India** had significantly higher populations compared to other countries, followed by the **United States**, **Indonesia**, **Pakistan**, and **Nigeria**.

4. Density per Continent

- Another grouping analysis calculated the average population density per square kilometer for each continent. The bar plot revealed that **Asia** has the highest density, followed by **Europe** and **North America**

5. Population Growth Trends Over the Years

- The plot reveals a global trend of declining population growth rates across all continents, with Africa maintaining the highest but decreasing growth, and Asia and South America experiencing sharp declines, especially post-2000; Europe and Oceania started with low growth rates and continued to decline gradually, indicating demographic shifts and a global movement toward population stabilization.

Model Building

1. **Model Selection**: The **Random Forest Regressor** was chosen due to its ability to handle complex, non-linear relationships and provide high accuracy through an ensemble of decision trees. Its robustness to overfitting and effectiveness in dealing with large datasets made it suitable for predicting population figures based on historical data.
2. **Data Splitting and Training**: The data was split into 80% training and 20% testing sets, and the Random Forest model was then trained on the training data.

Model Evaluation

- Mean Absolute Error (MAE): 0.0255
- Mean Squared Error (MSE): 0.0074
- R² Score: 0.9581
- Model Used for Cross-Validation: `cross_val_score`
- Average R² Score: 0.9314

Results and Discussion

- The **MAE** indicates that, on average, the model's predictions are approximately 2.55% off from the actual population values. This low MAE suggests that the model provides reasonably accurate predictions, making it reliable for forecasting population changes.
- The **MSE**, being relatively low, emphasizes that the model minimizes larger errors effectively. This metric is especially useful in understanding the model's sensitivity to outliers. A low MSE implies that the model performs well across various population sizes without significant deviations.
- The **R^2** score indicates that approximately 95.81% of the variance in population data can be explained by the independent variables included in the model. This high value reflects an excellent fit, suggesting that the model captures the underlying patterns in the data effectively.
- The average R^2 score from **cross-validation** further reinforces the model's robustness. It indicates that, even when evaluated on different subsets of the data, the model consistently explains a significant portion of the variance (approximately 93.14%). This consistency suggests that the model is not overfitting and can generalize well to unseen data.

Conclusion

The World Population Analysis Project effectively utilized historical data to predict future population trends, employing a Random Forest Regressor model. The evaluation metrics—Mean Absolute Error of 0.0255, Mean Squared Error of 0.0074, and an R^2 score of 0.9581—demonstrate the model's strong predictive accuracy and its ability to explain a significant portion of variance in population data. With an average R^2 score of 0.9314 from cross-validation, the model showcases its robustness and generalizability across different data subsets. Overall, the findings highlight crucial insights into demographic shifts and trends, providing valuable information for resource planning, policy-making, and understanding the factors influencing global population dynamics.