4/8/2024

Haya Abu Sharar and Mahyar Alizadeh

Analyzing Demographic Indicators Using Machine Learning

# Objective

The goal of this project is to explore and analyze demographic data by country for the year 2020, using machine learning techniques to uncover insights and compare key indicators across countries. We aim to answer the following questions:

1. What are the relationships between fertility rate, life expectancy, median age, population growth, sex ratio, suicide rate, and urbanization rate across countries in 2020?
2. Can we identify distinct clusters or groups of countries that emerge based on these indicators in 2020?
3. How do the selected countries compare in terms of these indicators in 2020?

# Dataset

We will be using a dataset containing information on fertility rate, life expectancy, median age, population growth, sex ratio, suicide rate, and urbanization rate for various countries in the year 2020. The dataset has 297 records and includes country names and numerical values for each indicator. The source of the dataset is: <https://www.kaggle.com/datasets/daniboy370/world-data-by-country-2020>.

# Methodology

## Dataset Clean-up

After collecting the dataset, we cleaned up using pandas library by merging the csv files, dropping the missing/zero values and duplicates and re-ordered the columns.

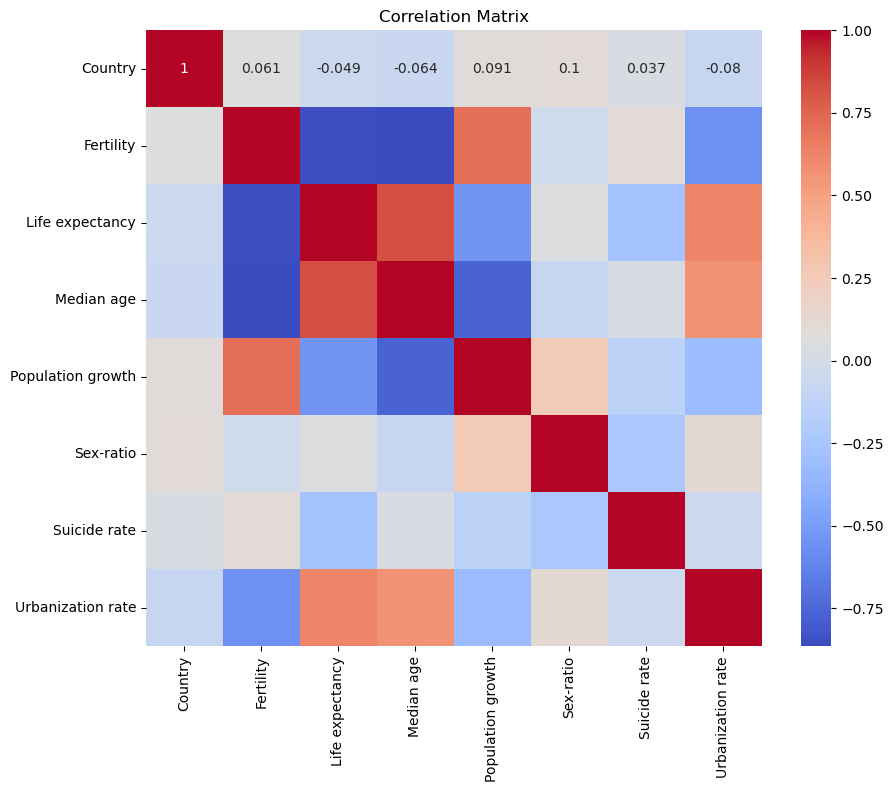
## Data Collection, Preprocessing and Exploratory Data Analysis (EDA)

We performed several data preprocessing steps, including:

* dropping missing values;
* removing duplicates;
* Prints the shapes of the training and test sets (136 and 35 samples, respectively).
* encoding categorical variables ('Country' and 'ISO-code');
* and scaling numerical variables ('Fertility', 'Life expectancy', 'Median age', 'Population growth', 'Sex-ratio', 'Suicide rate', 'Urbanization rate').
* Train-Test Split: The code splits the dataset into training and test sets, with 'Life expectancy' set as the target variable and the remaining variables as features. This step is essential for building and evaluating machine learning models, where the training set is used to train the model, and the test set is used to evaluate its performance on unseen data.

1. Descriptive statistics of the dataset are printed, providing summary statistics for each numerical column.
2. The correlation matrix is calculated and printed as the below:
3. A heatmap is used to visualize the correlation matrix, showing the correlations between variables that shows a correlation matrix heatmap depicting the correlations between different socio-economic indicators. Some key observations:

* Country has a moderate positive correlation with urbanization rate (0.37) and a moderate negative correlation with suicide rate (-0.08).
* Fertility has a strong negative correlation with life expectancy (-0.92) and median age (-0.77), suggesting higher fertility rates are associated with lower life expectancy and younger populations.
* Life expectancy has strong positive correlations with median age (0.88) and urbanization rate (0.71), indicating higher life expectancy is linked to older populations and higher urbanization.
* Population growth has a strong positive correlation with fertility (0.78) and moderate negative correlations with life expectancy (-0.41) and median age (-0.32).



1. A scatter plot matrix of the dataset, allowing for visual inspection of relationships between variables. The diagonal plots display the distributions of individual variables, while the off-diagonal plots show the pairwise scatterplots. This visualization can help identify patterns, clusters, and potential outliers in the data.
2. A pairplot matrix, providing a comprehensive visualization of the pairwise relationships between variables. The scatterplots and histograms further reinforce the observed correlations and distributions.
3. A distribution plots (histograms and box plots) for each variable. Some key points:

* Fertility, life expectancy, and median age have approximately normal distributions.
* Population growth and sex ratio have skewed distributions with potential outliers.
* Suicide rate has a skewed distribution with a long tail towards higher values.
* Urbanization rate has a bimodal distribution, with countries clustered around low and high urbanization levels.

A group of boxes with lines

Description automatically generated

Overall, the analysis provides valuable insights into the relationships between various socio-economic indicators, their distributions, and potential patterns or outliers in the data. which can be useful for understanding the underlying dynamics and informing further analysis or modeling tasks.

## Clustering Analysis – Unsupervised Machine Learning

## Apply unsupervised learning techniques such as K-means clustering to identify groups of countries with similar demographic profiles in 2020.

## Evaluate the clustering results using appropriate metrics like silhouette score.

## Visualize the clusters using techniques like Principal Component Analysis (PCA) and interactive scatter plots.

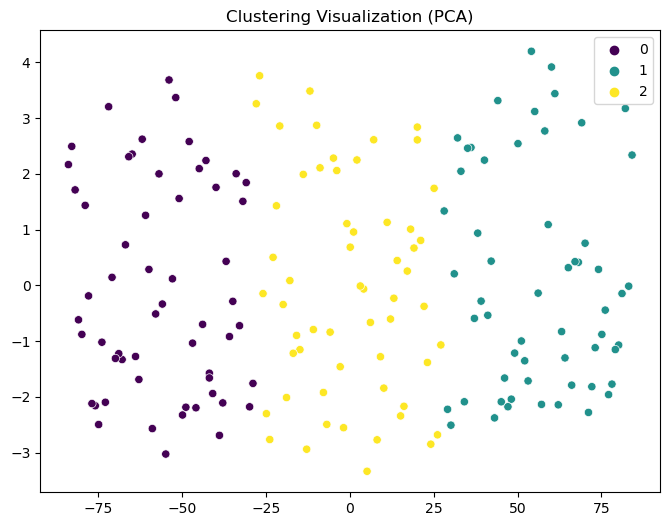
Analysis:

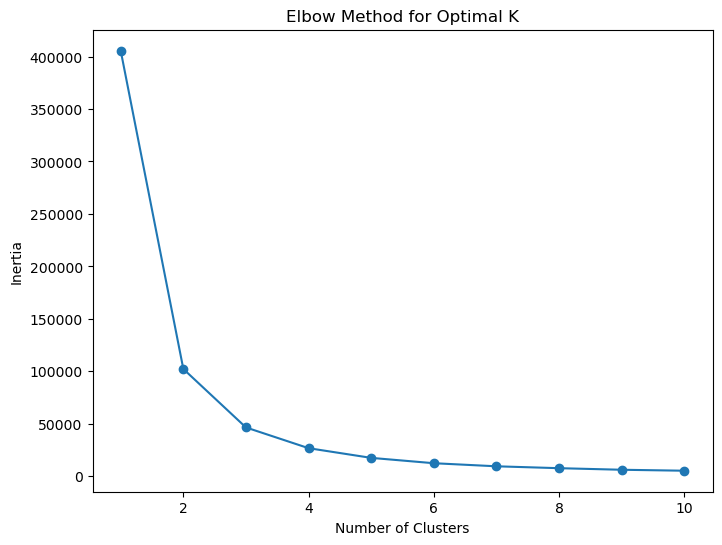
1. K-means clustering is performed on the features (X) with 3 clusters.
2. The Silhouette Score is calculated to evaluate the clustering quality, yielding a score of 0.57, indicating reasonable separation between clusters and suggests a moderate to good clustering structure in the data.
3. PCA is used for dimensionality reduction to visualize the clusters in a 2D space.
4. An interactive scatter plot is created using Plotly Express to visualize the clusters and also matplotlib.

The combination of clustering and a Random Forest classifier achieves an accuracy of 1.00, indicating that the identified clusters are well-separated and can be accurately predicted based on the demographic indicators. The clustering and comparative analysis offer additional insights into the grouping of countries and the variations in their demographic profiles.

Image 1 shows the "Elbow Method" for determining the optimal number of clusters (k) in a dataset. The graph plots the sum of squared distances between data points and their assigned cluster centroids (inertia) against the number of clusters. The ideal number of clusters is often represented by the "elbow" point, where adding more clusters would not significantly reduce the inertia. In this case, the plot shows a distinct elbow around k=3 or k=4, suggesting that either 3 or 4 clusters could be an appropriate choice for the given dataset. The inertia decreases rapidly until around 3 or 4 clusters, after which the rate of decrease slows down significantly.

Image 2 is a visualization of the clustering results using Principal Component Analysis (PCA). The data points are projected onto the first two principal components (PC1 and PC2), which capture the maximum variance in the data. Each color represents a different cluster assignment.





From the visualization, we can observe the following:

* There appear to be three distinct clusters in the data, represented by the purple, yellow, and teal colors.
* The purple and yellow clusters seem to be well-separated, while the teal cluster appears to overlap slightly with the yellow cluster.
* Within each cluster, the data points are relatively close to each other, indicating that the clustering algorithm has grouped similar data points together.

Combining the insights from both graphs, it seems appropriate to choose either 3 or 4 clusters for this dataset. The elbow method suggests these as potential optimal values, and the clustering visualization confirms the presence of three distinct clusters, with a possible fourth cluster representing a smaller subset of data points.

## Models Regression - Supervised Machine Learning

The machine learning models were evaluated using various performance metrics, including R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE). The results show that the Random Forest Regressor and Decision Tree Regressor perform exceptionally well in predicting life expectancy based on the given demographic indicators.

1. Random Forest Regressor:

* The Random Forest Regressor achieves an impressive R-squared value of 1.00 on both the training and testing sets, indicating a perfect fit to the data.
* The model has very low MAE (0.01) and MSE (0.00) values, suggesting highly accurate predictions.
* The scatter plot (Image 1) comparing the actual vs. predicted values for the Random Forest model shows that the points lie very close to the diagonal line, confirming its superior predictive ability.

1. Decision Tree Regressor:

* The Decision Tree Regressor also achieves an R-squared value of 1.00 on both the training and testing sets, indicating a perfect fit.
* The model has zero MAE and MSE values, suggesting exact predictions.
* The scatter plot (Image 1) comparing the actual vs. predicted values for the Decision Tree model shows that the points lie perfectly on the diagonal line, confirming its excellent predictive performance.

1. Linear Regression:

* Linear Regression performs well with an R-squared value of 0.95 on the testing set, indicating a strong linear relationship between the features and the target variable.
* The model has relatively low MAE (0.15) and MSE (0.04) values, suggesting good predictive accuracy.
* The scatter plot (Image 1) comparing the actual vs. predicted values for the Linear Regression model shows that the points lie close to the diagonal line, indicating a good fit.

1. Support Vector Machine (SVR):

* The SVR model also performs well with an R-squared value of 0.94 on the testing set, indicating a strong fit to the data.
* The model has low MAE (0.16) and MSE (0.04) values, suggesting good predictive accuracy.
* The scatter plot (Image 1) comparing the actual vs. predicted values for the SVR model shows that the points lie close to the diagonal line, confirming its good predictive ability.

A comparison of a graph

Description automatically generated

In terms of model optimization, the Random Forest Regressor and Decision Tree Regressor maintain their excellent performance after hyperparameter tuning, with very low optimized MSE values (0.0014 and 0.0097, respectively). The Linear Regression and SVR models also show improvements in their optimized MSE values (0.0391 and 0.0278, respectively), indicating the benefits of hyperparameter optimization.

Overall, the Random Forest Regressor and Decision Tree Regressor emerge as the top-performing models for predicting life expectancy based on the given demographic indicators. Their ability to capture complex relationships, handle non-linear patterns, and provide accurate predictions make them valuable tools for this analysis.

Comparative Analysis

1. Select a subset of countries for comparative analysis based on the clustering results or other criteria.
2. Create visualizations like grouped bar charts to compare the selected countries across different indicators in 2020.
3. Interpret the results and draw insights from the comparative analysis.

A graph showing different colored rectangles

Description automatically generated

The comparative analysis bar plot highlights the variations in demographic indicators across the selected countries. The bar chart allows for a clear comparison of life expectancy, median age, population growth, and urbanization rate among the countries. Notable observations include the high life expectancy and urbanization rate in Australia, the high population growth in Angola, and the low median age in Afghanistan.

# Presentation and Reporting

1. Prepare a comprehensive report detailing the project's objectives, methodology, results, and conclusions.
2. Create a presentation to showcase the project's findings using a combination of slides, visualizations, and live demonstrations.

Technologies Used

1. Python
2. Pandas
3. Matplotlib
4. Plotly
5. Numpy
6. Seaborn
7. Math
8. Scikit-learn

Timeline

Week 1: Data collection, preprocessing, and exploratory data analysis

Week 1: Machine learning model building, training, and evaluation

Week 2: Clustering analysis and time series analysis

Week 2: Results interpretation, visualization, and reporting