Group: 33 **Members:** António Saraiva (52795) - 70h; Matilde Carvalho (54838) - 70h; Rita Rodrigues (54859) - 70h

Introduction

For our machine learning course project, we used the World Bank dataset as the source of data for our analysis. The World Bank dataset is a rich and diverse collection of data on a variety of economic and social indicators, covering over 200 countries and territories worldwide. The aim of this study was to develop a prediction model for forecasting country population, fertility rate, and life expectancy using three datasets covering the period up to 2016. In the following sections, we describe the specific aspects of the World Bank dataset used in our analysis, the machine learning algorithms implemented, and the evaluation and performance of the algorithms. We then present and discuss the results that we obtained, and conclude with a summary of our work.

Dataset information

The dataset on country population provides information on the number of people living in each country. It includes all residents of a country, regardless of their legal status or citizenship. The data is collected through official surveys and is reported at midyear. The dataset on fertility rate reflects the number of children that a woman would have if she lived to the end of her reproductive years and had children at the fertility rates for her specific age group in a given year. Life expectancy at birth dataset is a measure of the number of years a newborn infant is expected to live, if the mortality patterns at the time of its birth were to continue throughout its life.

As a first step, a Jupyter notebook was created and the three provided datasets were imported for analysis.

Data processing

This study was conducted under the assumption that it was December 2016, with the objective of forecasting the values for country population, fertility rate, and life expectancy for the years 2017 and 2018. We decided to focus on only the previous six years of data (2011-2016), although we had values since 1960, because of several reasons: the data for the most recent years is likely to be more accurate and up-to-date than data from further in the past; focusing on more recent data may allow to capture recent trends and patterns in the data that may not be evident when looking at a longer time period; limiting the analysis to a shorter time period may make it easier to manage and analyze the data, as the work is done with a smaller dataset.

Following the selection of data from the past six years, we conducted an analysis of the number of countries and regions in each dataset that had missing values for any of the years. We found that the number of countries and regions in this situation was relatively low, leading us to eliminate from all datasets those countries and regions that had at least one missing value in any year for any of the variables. This resulted in a reduction of the initial sample of 264 countries and regions to 243, which we considered acceptable for the purpose of the study.

Subsequently, we addressed the issue of data structure. We decided to focus on the rate of change (delta) between consecutive years rather than absolute values of the data, due to several reasons: the rate of change may provide a more accurate representation of trends or patterns in the data than the absolute values. The rate of change can be more meaningful in terms of understanding the dynamics of a system. For example, to analyze population data, the rate of change in the population size can be more informative than the absolute population size itself, as it can provide insight into how the population is growing or declining. Also, the rate of change can be easier to compare across different time periods or between different groups or regions. For example, to compare population growth rates between two different countries, the rate of change may be a more meaningful comparison than the absolute population sizes, as the countries may have very different starting population sizes.

We developed a function that, given a dataframe, would return another dataframe containing the rates of change. For each country, the resulting dataframe included two rows: one with the rates of change between 2011 and 2015, and another between 2012 and 2016.

An example of the resulting dataframe with the rates of change for the population dataset is represented in the following figure:

	D0	D1	D2	D3
0	524.0	610.0	608.0	546.0
1	610.0	608.0	546.0	481.0
2	988359.0	1034730.0	1026332.0	978474.0
3	1034730.0	1026332.0	978474.0	919538.0
4	877585.0	902190.0	922126.0	938839.0
...	...	...	...	...
481	769183.0	772175.0	751654.0	724248.0
482	435181.0	453273.0	467764.0	479613.0
483	453273.0	467764.0	479613.0	490803.0
484	324177.0	343680.0	357169.0	365776.0
485	343680.0	357169.0	365776.0	372911.0

Modelling

We are interested in predicting the country's population, fertility rate and life expectancy for 2017 based on the values of the variables for the previous six years. We approached this as a regression problem, as we are trying to predict a numerical outcome (the values for 2017) based on the predictor variables (the values in the previous six years).

The X variable was defined as the first three rates of change, while the Y variable was defined as the last. For instance, in the previously mentioned example, the X corresponded to the columns ‘D0’, ‘D1’, ‘D2’ and Y to ‘D3’.

For each variable we used train-test-split, a technique that divides the dataset into two subsets: a training set and a test set (which we considered to have a size of 25% of the total dataset). The training set is used to fit the model, while the test set is used to evaluate the model’s performance. This technique allows us to evaluate the model’s generalization ability, helps avoid overfitting and allows us to compare different models.

The models that we compared were Linear Regression, XGBRegressor and MLPRegressor.

Linear Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It is often used to make predictions about a continuous numerical outcome based on one or more predictor variables. That’s why we considered it would be a good model to predict the country’s population and the other variables for a given year based on the previous years. Also, it is a simple model, which makes it easy to implement and it is relatively fast to fit. XGBoost (eXtreme Gradient Boosting) is a powerful and popular machine learning algorithm that is often used for regression and classification tasks. It is an implementation of gradient boosting, which is a technique that involves combining a series of weak models to create a strong, accurate model. We decided to use XGBRegressor in our study because it is able to model complex relationships between the predictor variables and the outcome variable. It is also relatively fast to train, even with large datasets, and can often perform well on a wide range of regression tasks. MLPRegressor is a type of artificial neural network (ANN) that is specifically designed for regression tasks. ANNs are machine learning models that are inspired by the structure and function of the human brain, and are composed of multiple interconnected layers of artificial neurons. MLPRegressor is a type of feedforward ANN that is particularly useful for predicting numerical outcomes based on one or more predictor variables which is what we want to accomplish. It is capable of modeling complex non-linear relationships between the predictor variables and the outcome variable, and has been shown to achieve high levels of accuracy on a wide range of regression tasks.

For XGBRegressor and MLPRegressor, GridSearchCV was used. It is a method for automating the process of hyperparameter tuning by searching a grid of possible hyperparameter combinations and selecting the combination that performs the best on the training data. It does this by fitting the model using each combination of hyperparameters, evaluating the model’s performance using cross-validation, and selecting the combination that achieves the highest performance. The scoring metric used in GridSearchCV was the negative root mean squared error, as the lower this value is, the better the model performs, and GridSearchCV searches for the highest value of the metric.

```
xgb_params = {
    "learning_rate": (0.05, 0.10, 0.15),
    "max_depth": [ 3, 4, 5, 6, 8],
    "min_child_weight": [ 1, 3, 5, 7],
    "gamma":[ 0.0, 0.1, 0.2],
    "colsample_bytree":[ 0.3, 0.4]
}
```

```
mlp_params = {
    "activation":['identity', 'relu'],
    "solver":['lbfgs', 'sgd', 'adam'],
    "hidden_layer_sizes": [(10,),(50,),(100,)],
    "alpha": [0.0001, 0.01, 0.1, 1]
}
```

The grid on the left corresponds to XGBRegressor and the one on the right to MLPRegressor.

Root mean squared error (RMSE) is a common metric used to evaluate the performance of regression models. It is defined as the square root of the mean of the squared differences between the predicted values and the actual values. A smaller RMSE value indicates a better fit. Other metrics were used to evaluate the models: RVE, Maximum Error and Mean Absolute Error. The Ratio of the Variance Explained (RVE) is a measure of the proportion of the variation in the outcome variable that can be explained by the predictor variables, and is calculated as the squared correlations between the predicted values and the actual values divided by the variance of the actual values. A higher RVE value indicates that the model is able to explain a larger proportion of the variation in the outcome variable. Maximum Error is a measure of the largest difference between the predicted values and the actual values. A smaller Maximum Error value indicates that the model is able to make more accurate predictions. Mean Absolute Error (MAE) is a measure of the average difference between the predicted values and the actual values. A smaller MAE value indicates that the model is able to make more accurate predictions.
The RMSE was identified as the primary metric for evaluating the performance of the previously mentioned models, as it is a widely used and well-established measure of prediction error in the field of regression analysis. The model that achieved the lowest RMSE was considered the best, as this indicates the lowest level of prediction error. Additionally, the other metrics evaluated (RVE, Maximum Error and MAE) also demonstrated good performance for the best model, providing further support for the selection of the model with the lowest RMSE as the optimal model.

Results

The best model obtained to predict the country's population was MLPRegressor with alpha equals to 1 and solver 'lbfgs'. The metrics obtained were:

```
The RVE is:  0.9999575698032417
The rmse is:  76674.61642138011
The Maximum Error is is:  440792.84928962216
The Mean Absolute Error is:  33563.24583469129
```

The best model obtained to predict fertility rate was XGBRegressor with the metrics:

```
The RVE is:  0.5921211617285371
The rmse is:  0.02334243043651772
The Maximum Error is is:  0.14704733207821863
The Mean Absolute Error is:  0.01437573699485383
```

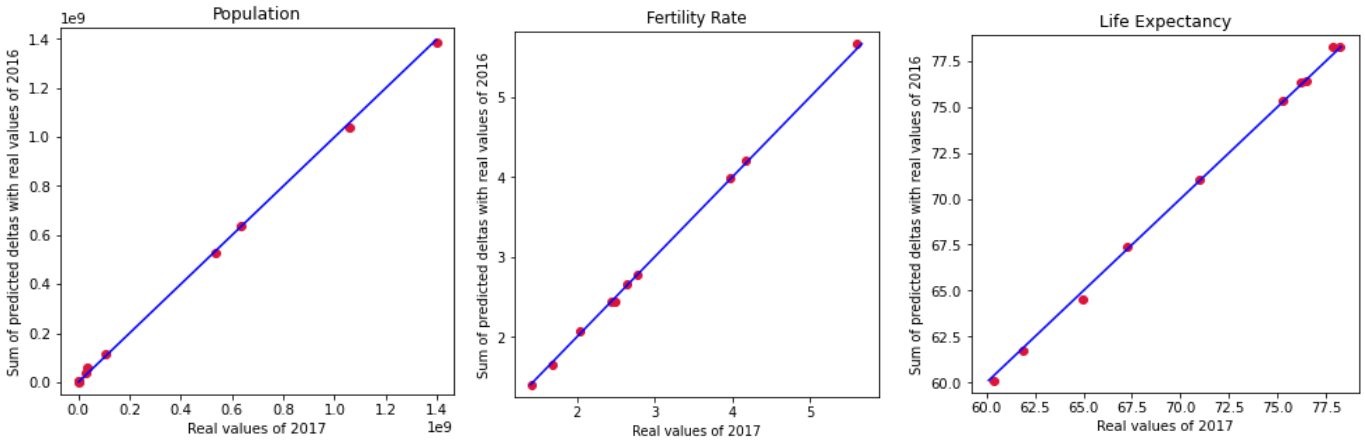
And the best model obtained to predict life expectancy was MLPRegressor with alpha 0.1, hidden layer sizes (10,) and solver 'lbfgs'. The metrics obtained were:

```
The RVE is:  0.687975401021323
The rmse is:  0.11343076646615576
The Maximum Error is is:  0.7571266595461273
The Mean Absolute Error is:  0.05062444892671546
```

To evaluate the models, the real values for country population, fertility rate, and life expectancy for the years 2017 and 2018 were obtained from the World Bank API and compared to the predictions made by the models for a sample of 10 randomly selected countries.
In order to make predictions for the year 2017 using the regression models, three dataframes were created containing the rates of change for each variable from 2013 to 2016. This data was then used as input to the models, which were trained to make predictions based on the rates of change from the previous six years.
As an example, the following table shows the results for fertility rate.

Country Name	Predicted Deltas	Sum of predicted deltas with real values of 2016	Real values of 2017
Angola	-0.016391	5.669609	5.600000
China	-0.020925	1.654075	1.683000
Fiji	-0.019958	2.777042	2.788000
Croatia	-0.024315	1.405685	1.420000
IDA blend	-0.032732	4.200518	4.175809
IDA only	-0.032663	3.992640	3.968618
Latin America & Caribbean	0.000382	2.065797	2.046015
Morocco	-0.052771	2.436229	2.451000
Panama	-0.064126	2.449874	2.487000
Philippines	-0.050960	2.667040	2.640000

In order to visualize the relationship between the predicted values and the actual values, prediction-truth plots were made.



To make predictions for the year 2018, three dataframes were created that included the rates of change for each variable from 2014 to 2016, as well as the predicted rates for 2017, which were obtained previously. This data was then used as input to the models.
The results for life expectancy are in the following table.

Country Name	Predicted Deltas	Sum of predicted deltas with predicted values of 2017	Real values of 2018
Angola	0.279738	60.402630	60.782000
China	0.244896	76.614262	76.704000
Fiji	0.165355	67.531485	67.341000
Croatia	0.250706	78.498751	78.070732
IDA blend	0.267383	62.001590	62.161867
IDA only	0.208246	64.728481	65.305310
Latin America & Caribbean	0.258284	75.546775	75.440849
Morocco	0.306270	76.606236	76.453000
Panama	0.281082	78.528554	78.329000
Philippines	0.257256	71.298315	71.095000

Discussion

Overall, the MLPRegressor was found to be the most effective model for these tasks, achieving the lowest levels of RMSE. This suggests that the model was able to effectively capture the underlying patterns in the data and make accurate predictions.
The prediction-truth plots also provided visual evidence of the high accuracy of the predictions made by the models. These plots showed that the predicted values were closely aligned with the actual values, indicating that the models were able to effectively capture the relationship between the predictor variables and the outcome variables.

Conclusion

In conclusion, this work explored the use of machine learning techniques for predicting country population, fertility rate, and life expectancy. A variety of regression models were trained using data from the World Bank, and their performance was evaluated using a range of metrics. Overall, the results suggest that machine learning has the potential to be useful for forecasting and decision-making in the field of population and health.