Predicting population growth in the Arab region using RapidMiner.

Student number: 22006181

June 2021

Contents

Introduction

Overview

Health is a multifaceted concept and collectively, population health can affect the economic and social development of a country by impacting human capital within the country. Some of the indicators of population health include life expectancy, death rates, healthy life years and infant mortality rates. In measuring the national quality of life, medical coverage, levels of physical activity, levels of unhealthy behaviours such as smoking or over-eating, and nutrition levels through consumption of fruits and vegetables are considered. These factors are linked to the national health expenditure through the nation's ability to provide access to health and physical improvement facilities as well as health information and support.

Population growth puts a strain on the economy by restricting the quantity of resources available due to an increased demand for these resources (Shneider et al., 2011). This includes health resources. Where population can be predicted, it allows a nation to make more efficient financial budgets. In many countries, especially European countries, health expenditure is a significant aspect of the budget to ensure access to health amenities and promote health coverage. This funding also goes into the development and implementation of policies to promote a better lifestyle for improved mental and physical health.

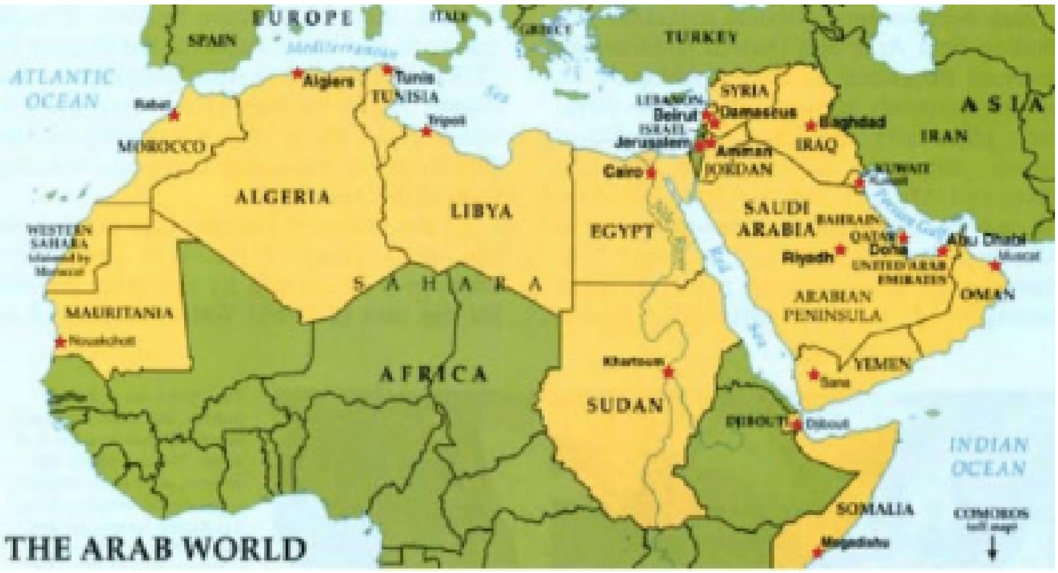


Figure 1. The map of the Arab region with the member countries highlighted in yellow.

The national economy for countries in the Arab region is remarkably diverse as it consists of very affluent countries considered to be developed and extremely poor, war-torn countries with a malnourished population and large number of refugees. For example, Qatar has GDP per capital in the world which is 260 times higher than the GDP per capita of countries like Yemen (Asfour & Jabbour, 2020). The Arab region is home to over 400 million individuals and political instability in the region has given rise to millions of refugees, thereby putting a strain on the infrastructures of countries like Jordan and Lebanon. The overall public health system in this region is weak and has received insufficient health funding. The government supplies a health budget of 2.8% of the GDP on average, compared to the global rate of 7.8% (Asfour & Jabbour, 2020).

For this reason, the Arab region was selected as the area of interest for this study. If population growth could be predicted it could potentially be useful in estimating the amount of money that should be allocated in the health budget for a better quality of life within the region, through localised means and international support.

Objectives

This study aims to estimate and predict population growth within the Arab region and investigate a probable link between regional health budgets and population growth.

Data source

This study uses a time series dataset from HealthStats under the World Bank data catalogue consisting of a collection of health, nutrition, and population statistics. The dataset consists of the key health, nutrition and population statistics gathered nationally and internationally, covering 217 economies between the years 1960 and 2020. It includes over 250 indicators of health financing, nutrition, communicable and non-communicable diseases, and population estimates.

The CRISP-DM methodology was referenced in building the predictive model for this research.

CRISP- DM Methodology

1) Data Understanding

Six datasets were extracted from the World Bank Data catalogue, one of these contained the main data used in this study in csv format and the other five datasets explained the many categories within the large dataset.

The main dataset contained both national and regional datasets with some blank categories for certain earlier years within some regional and national data.

Input Attributes

The original dataset consists of 110,852 examples, 0 special attributes and 65 regular attributes and missing data in almost every attribute.

The dataset was initially cleaned and transposed in Excel for analyses in RapidMiner. The dataset used only contains information relating to the Arab region.

Terminology

- CSV = Comma separated values
- Special attributes – Factors of interest for the study
- LR – Linear regression

2) Data Preparation

Pre-processing

The quality of the dataset retrieved from World bank was highly dependent on both the year of collection and the region with some years and regions being sparsely populated (Including the Arab region) while others had higher quality data.

Initial preparation of the csv file was performed in Excel. Data for all regions apart from the Arab region were removed, and attributes unrelated to the area of research were removed as well. The dataset was transposed in Excel to fit the required format for further analysis within RapidMiner.

The dataset was then imported into RapidMiner using the “Read CSV” operator and the Import Configuration Wizard. The “Parse numbers” operator was used to convert the examples appearing as nominal to numeric. “Replace Missing Values” operator was then introduced to replace all missing values in the dataset with “zero”. The missing values were changed to zero rather than estimated averages to avoid tampering with the results. The “Rename” operator was used to change the name of the column with the years from “Indicator name” to “Year” then the “Set Role” operator was introduced at this stage to identify the target attributes. This process was titled “Pre-processing”.

3) Modelling

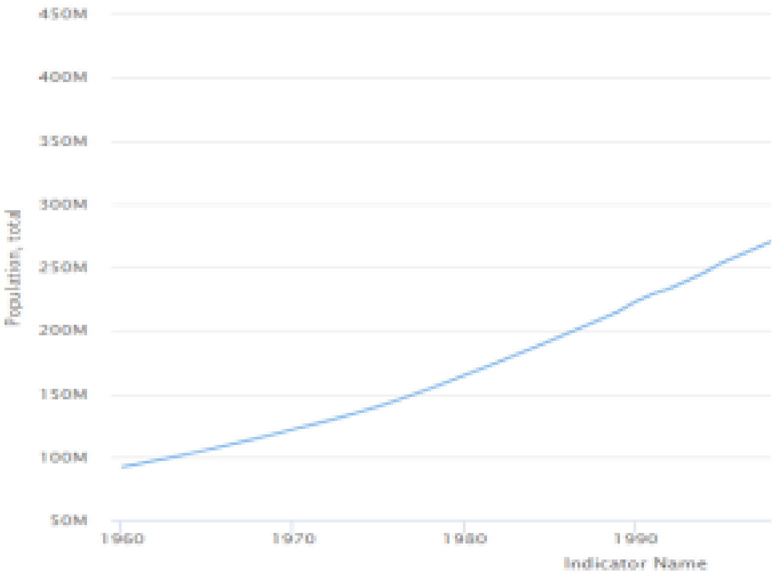
This study explores the most accurate predictive model to investigate population growth in the Arab region. Four models will be developed in RapidMiner to provide visual analyses to answer the following questions:

1. Which predictive model would be best for this data?
2. Is there a relationship between health expenditure and population growth?
3. Can health expenditure trends be used to predict population growth in the Arab region?
4. Can existing trends be used to predict population growth in the next 10 years?

Selection of model

In considering what model would most accurately investigate the relationship between health expenditure and population growth, the data was visualized in RapidMiner.

A



Year

B

Year

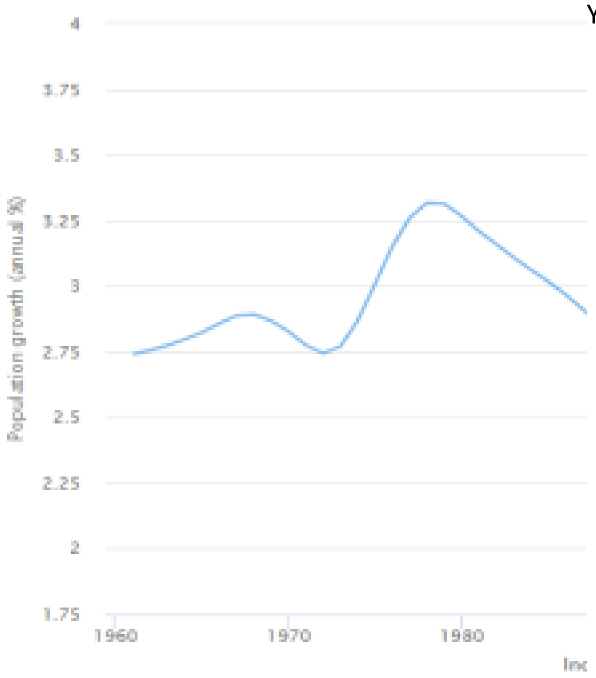


Figure 2. A) The trend of total population in the Arab region and **B)** the trend of popoulation growth over the years 1960 and 2020 in the Arab region.

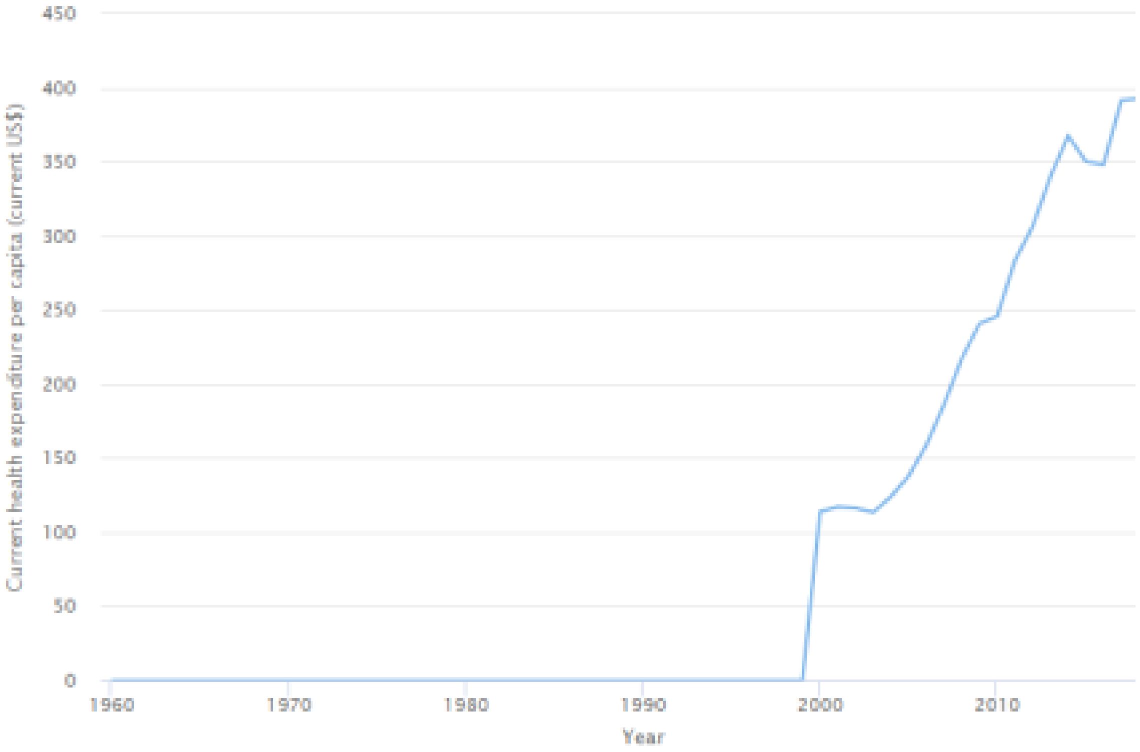


Figure 3. The trend of health expenditure in the Arab region between 2000 and 2020. There was no data available for the years earlier than that.

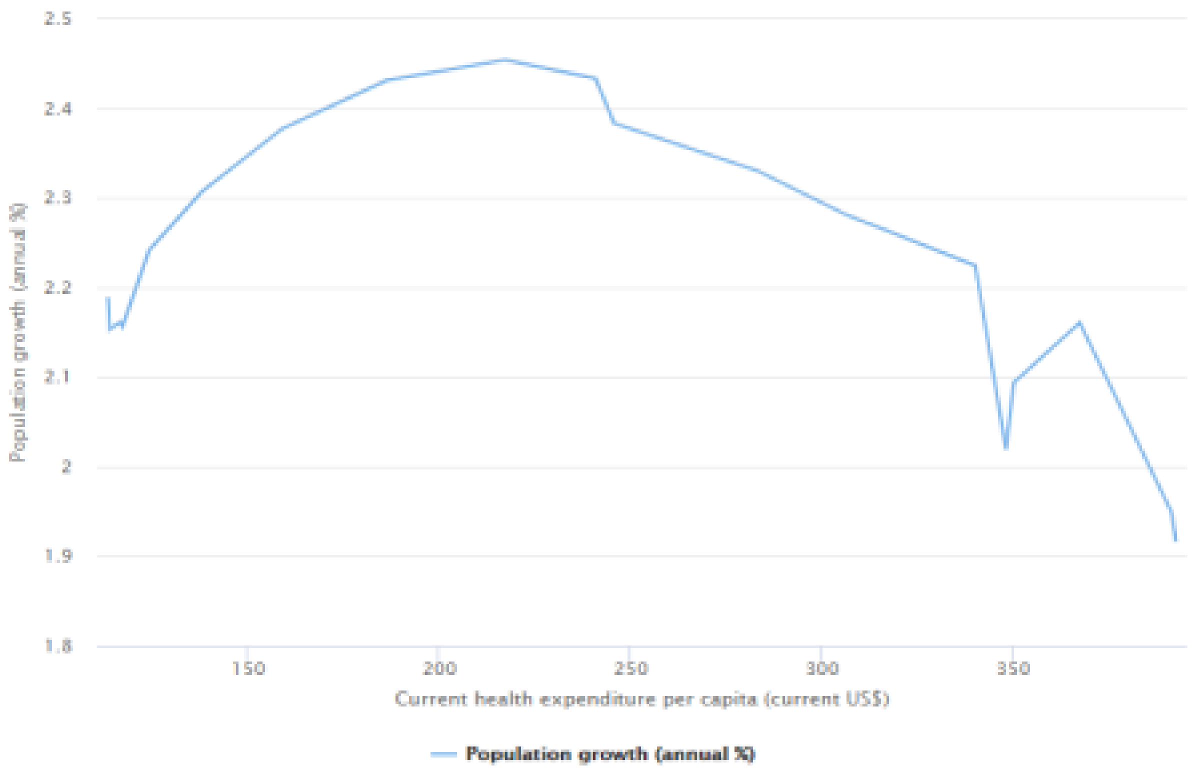


Figure 4. Population growth with health expenditure in the Arab region between 2000 and 2020.

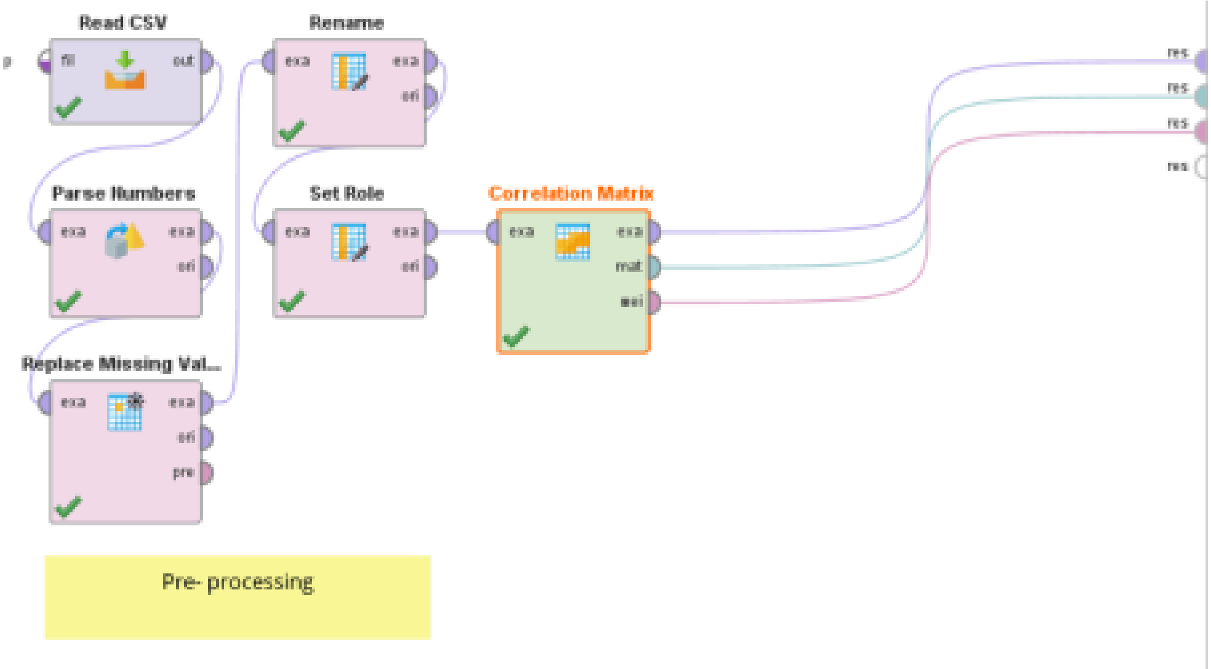
Based on the visualization and the questions the study aims to address, the following models were developed.

Correlation Matrix

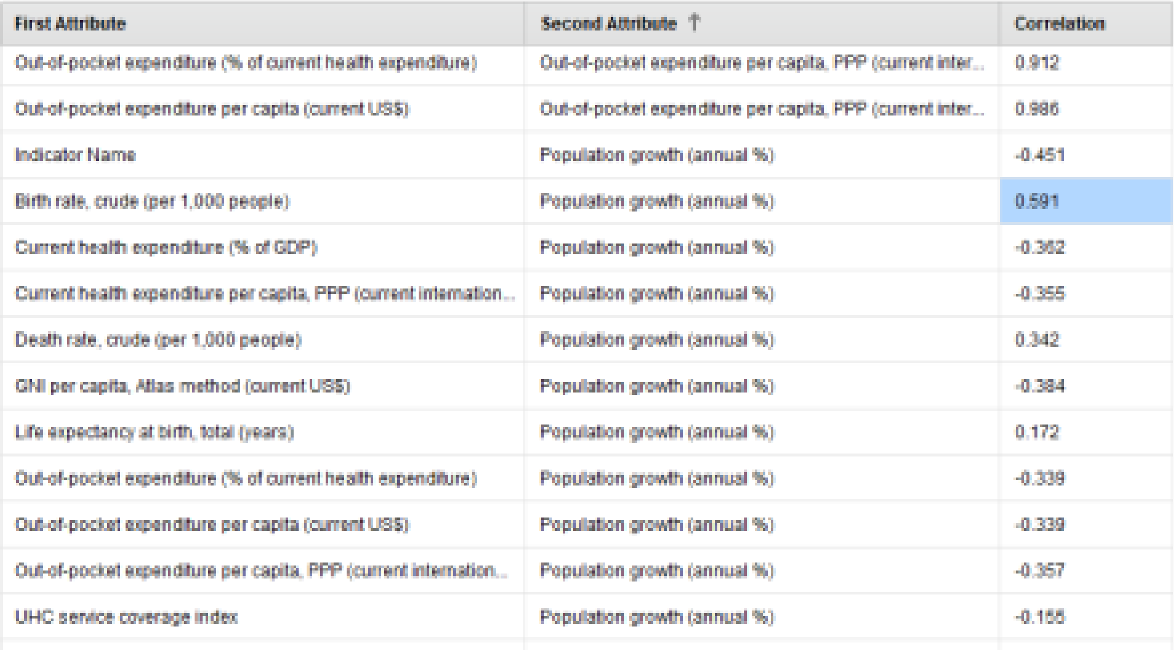
This operator determines correlation between all the attributes present in the dataset and produces a weights vector based on these correlations. It is used to show whether and how strongly pairs of attributes are related. A correlation is a number between -1 and +1 to indicate the degree of association between two attributes. Correlation analysis was used to investigate a relationship between health expenditure and population growth.

Process:

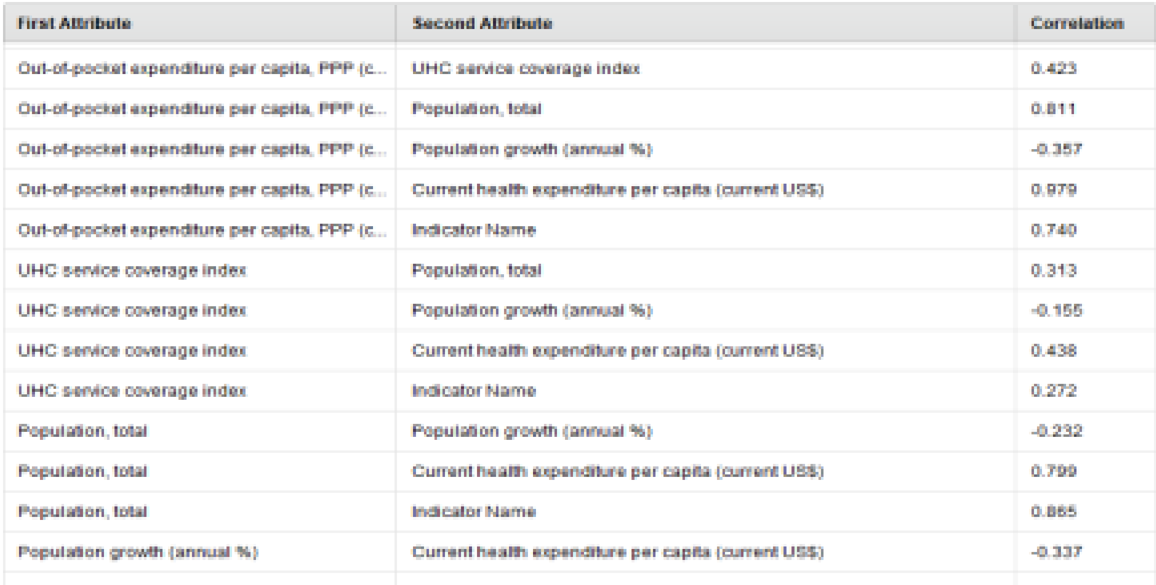
The process started with the Pre-processing system described above. The “Correlation Matrix” operator was then introduced.



Output:



First Attribute	Second Attribute	Correlation
Out-of-pocket expenditure (% of current health expenditure)	Out-of-pocket expenditure per capita, PPP (current inter...	0.912
Out-of-pocket expenditure per capita (current US$)	Out-of-pocket expenditure per capita, PPP (current inter...	0.986
Indicator Name	Population growth (annual %)	-0.451
Birth rate, crude (per 1,000 people)	Population growth (annual %)	0.591
Current health expenditure (% of GDP)	Population growth (annual %)	-0.362
Current health expenditure per capita, PPP (current internation...	Population growth (annual %)	-0.355
Death rate, crude (per 1,000 people)	Population growth (annual %)	0.342
GNI per capita, Atlas method (current US$)	Population growth (annual %)	-0.384
Life expectancy at birth, total (years)	Population growth (annual %)	0.172
Out-of-pocket expenditure (% of current health expenditure)	Population growth (annual %)	-0.339
Out-of-pocket expenditure per capita (current US$)	Population growth (annual %)	-0.339
Out-of-pocket expenditure per capita, PPP (current internation...	Population growth (annual %)	-0.357
UHC service coverage index	Population growth (annual %)	-0.155



First Attribute	Second Attribute	Correlation
Out-of-pocket expenditure per capita, PPP (c...	UHC service coverage index	0.423
Out-of-pocket expenditure per capita, PPP (c...	Population, total	0.811
Out-of-pocket expenditure per capita, PPP (c...	Population growth (annual %)	-0.357
Out-of-pocket expenditure per capita, PPP (c...	Current health expenditure per capita (current US$)	0.979
Out-of-pocket expenditure per capita, PPP (c...	Indicator Name	0.740
UHC service coverage index	Population, total	0.313
UHC service coverage index	Population growth (annual %)	-0.155
UHC service coverage index	Current health expenditure per capita (current US$)	0.438
UHC service coverage index	Indicator Name	0.272
Population, total	Population growth (annual %)	-0.232
Population, total	Current health expenditure per capita (current US$)	0.799
Population, total	Indicator Name	0.865
Population growth (annual %)	Current health expenditure per capita (current US$)	-0.337

For health expenditure and total population, there was a strong positive correlation indicating that an increase in health expenditure is related to an increase in total population. Whereas for health expenditure and population growth, there was a weak negative correlation, meaning that although it might appear to be the case, it cannot statistically be deduced that an increase in health expenditure correlates to a decline in population growth.

To predict population growth using the trend of health expenditure the following prediction models were developed:

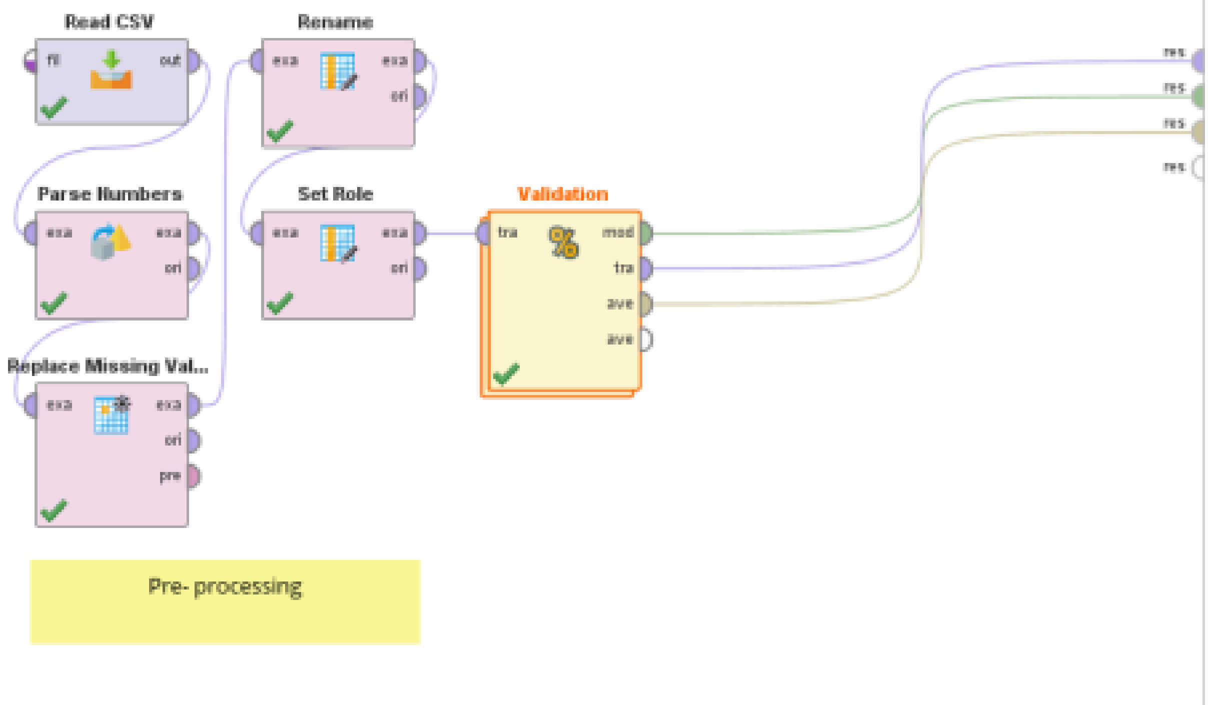
Linear regression model

According to a study by Ognjanovski (2018), population growth can be predicted using linear regression. Regression can be used to identify the strength of the relationship between a dependent variable and an independent variable. It can also be used for predicting a continuous variable.

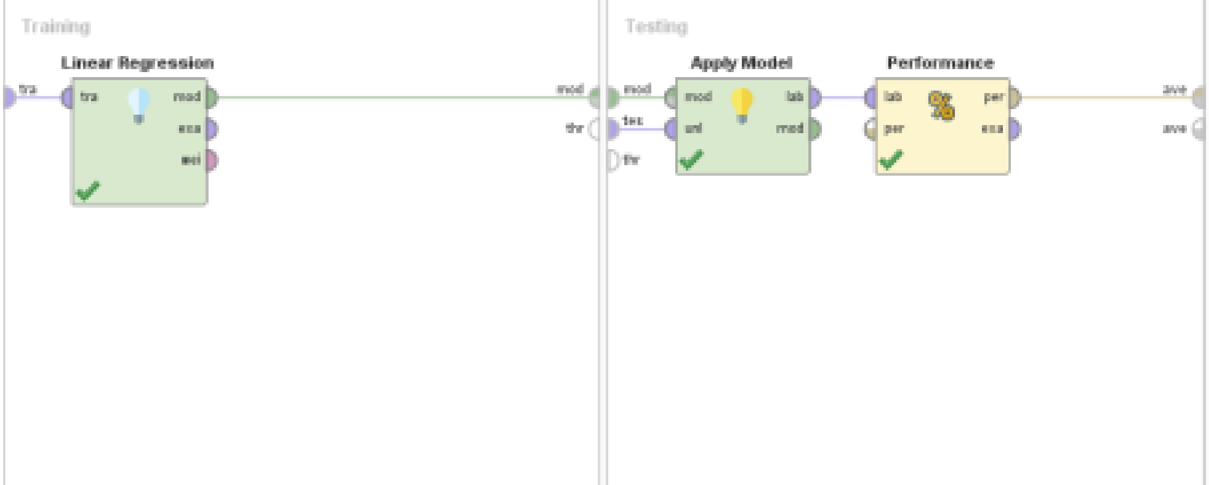
One of the assumptions for multiple linear regression is that attributes should be highly correlated, that is, have a correlation co-efficient of 0.6 or above. Based on this, the only attribute that could be considered for the prediction of population growth would be birth rate which had a COR of 0.59. For the purposes of exploring the use of multiple regression within RapidMiner, the regression was performed with all the outlined attributes, purely as an exercise.

Process

The Process starts with the Pre-processing defined earlier and the “Linear Regression” operator was introduced. The “Split validation” operator was introduced following the Pre-processing described. Within the Validation operator, the “linear regression” operator is introduced inside the training section of the validation operator. The “Apply Model” operator is then introduced in the testing section of the Validation operator and connected to the model and testing data connections. “Performance Regression” operator is then introduced in the testing section of the validation operator as it is the most suitable performance checker for this analysis.

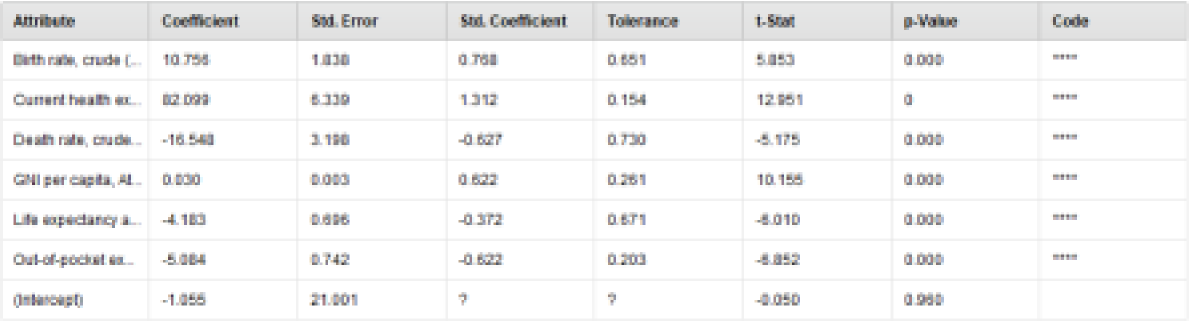


Within the “Split Validation” operator:



Output:

The performance operator resulted in a root mean squared error and prediction average were high indicating that another regression technique would be required to determine what model would perform better.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Birth rate, crude (...	10.756	1.838	0.768	0.651	5.853	0.000	****
Current health ex...	82.099	6.339	1.312	0.154	12.951	0	****
Death rate, crude...	-16.548	3.198	-0.627	0.730	-5.175	0.000	****
GNI per capita, Al...	0.030	0.003	0.622	0.261	10.155	0.000	****
Life expectancy a...	-4.183	0.696	-0.372	0.671	-6.010	0.000	****
Out-of-pocket ex...	-5.084	0.742	-0.622	0.203	-6.852	0.000	****
(Intercept)	-1.055	21.001	?	?	-0.050	0.960	

All these attributes are strong linear indicators of what population growth would be based on the star codes. While the graph shows that all these attributes are strong candidates to be included in the LR model, we recognise that all the attributes, but

Birth rate failed the correlation co-efficient test and therefore, they cannot be used to determine population growth.

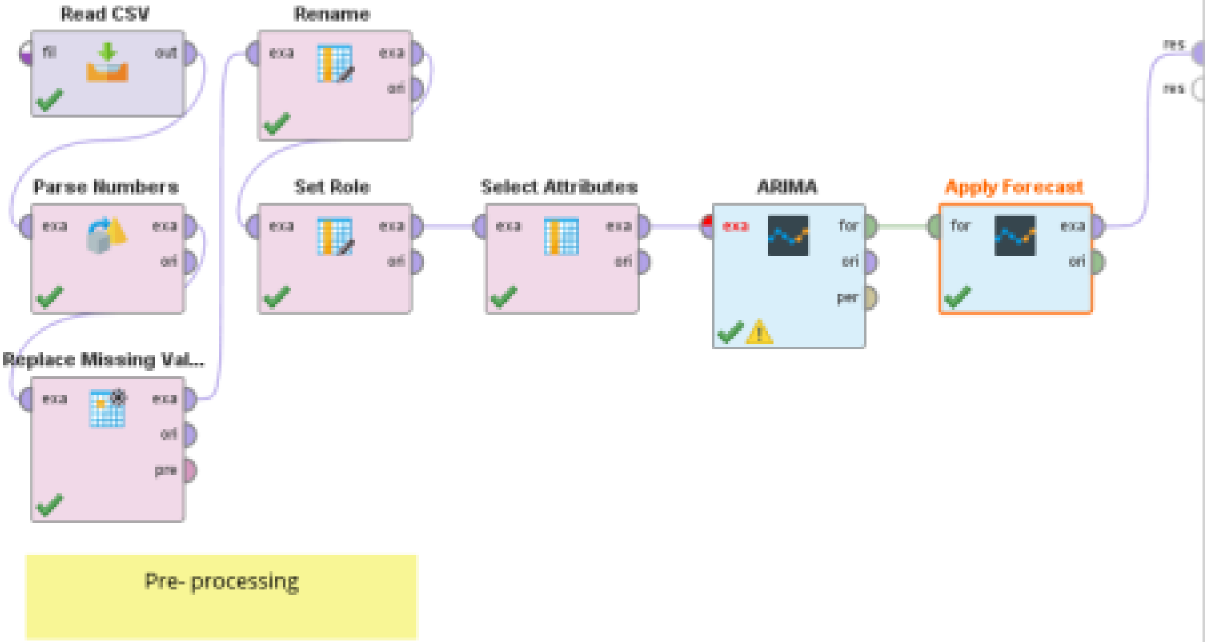
Therefore, to predict population growth, birth rate would be inputted in the *y=mx+c* formulas where, y = pop growth, m = coefficient, x = birth rate, and c = intercept. To create the model. Ideally, x would've have been health expenditure, but it failed the test for linear regression.

Time series analysis model

This process was used to analyse existing trends within the dataset and predict the next values for the time series.

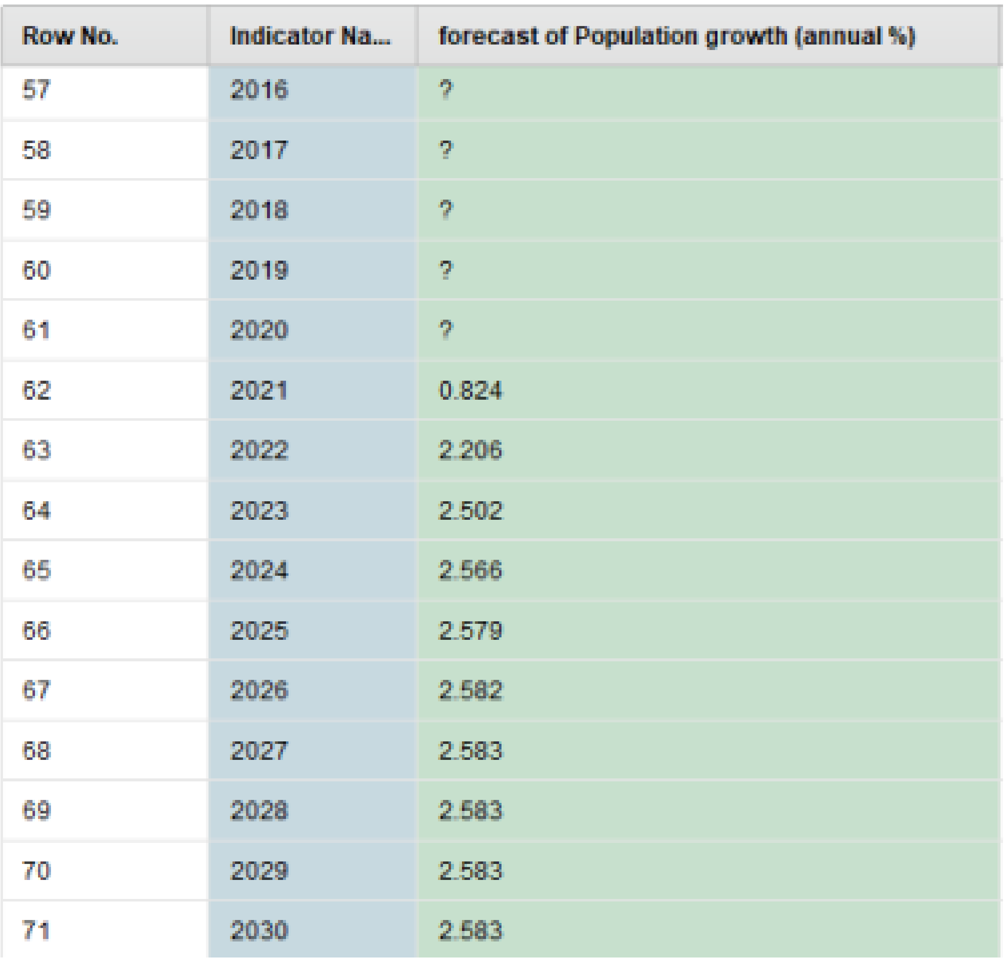
Process:

The "Apply Forecast" operator is key for this analysis and allows the length of the forecast to be defined. The process begins with the pre-processing defined earlier. The "Select Attributes" operator was used to select population growth as the target attribute for prediction. The "ARIMA" operator was then introduced for forecasting based on a moving window. Population growth was set as indices and year as indices attribute. The "Apply forecast" operator was then added in set horizon as 10:

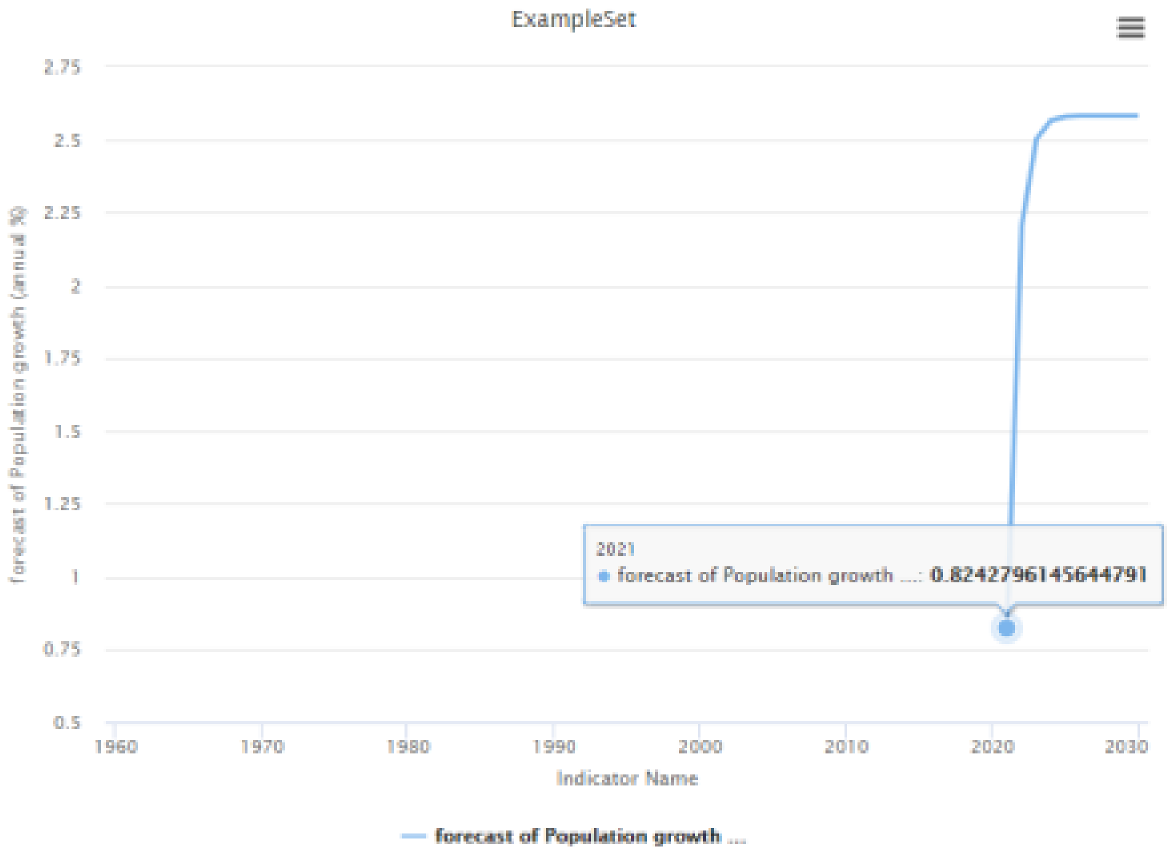


Output:

The root mean error for this analysis was 0.689 which is low, likewise the absolute error was low. This indicates that this model performed well for the predictive analysis.



Row No.	Indicator Na...	forecast of Population growth (annual %)
57	2016	?
58	2017	?
59	2018	?
60	2019	?
61	2020	?
62	2021	0.824
63	2022	2.206
64	2023	2.502
65	2024	2.566
66	2025	2.579
67	2026	2.582
68	2027	2.583
69	2028	2.583
70	2029	2.583
71	2030	2.583

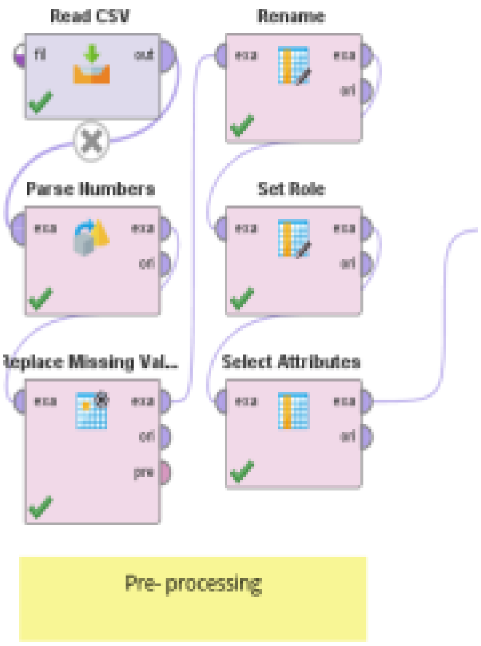


Decision tree model

The decision tree can be used in guiding decisions through classification of the data. It can be used in regression, here it separates values to reduce errors. In this study, we use it what amount of health expenditure results in a lower population growth rate per year.

Process:

The Pre-processing described was used to initiate the process then the “Select Attributes” operator was introduced to select the attributes to be considered in this process. “Split data” was used introduced following the standard 70/30 rule to partition the data into training and testing. The training set was connected to the “Decision tree” operator while the training set was connected to the “Apply Model” operator. The “Performance regression” operator was then used to evaluate the accuracy of the model.



Output:

The result shows that the decision tree is not a good model to use for this time series dataset. The decision tree model is mainly used for classification models and therefore does not fit in for predicting numerical attribute like this case.

4) Evaluation

At first glance, the visualization of our data shown in figures 2a – 4 gave the appearance of an increased total population of the years in the Arab region along with increased health expenditure. This makes sense because as the population increased, the financial burden increases as well. It also appeared that the population growth rate decreased over the years as health expenditure was increased. The correlation matrix model showed that regardless of this visualization, there was no actual relationship between population growth and health expenditure. As a result, the linear regression model could not be used to predict population growth using health expenditure. However, we found birth rate to show a relationship with population growth. This makes sense as a change in birth rate would alter population growth rate. It can be deduced that birth rate could be used in predicting population growth for health budgeting.

The modelling showed that time series analysis is the best model for this data as population growth is a factor of time as it performs better than regression analysis in forecasting time series data. Using the time series analysis, we were able to predict population growth in the Arab region for the next 10 years. The study was able to successfully answer the questions proposed.

Challenges

Missing data

The dataset contained missing data in most of the attributes including the special attributes. The Replace Missing Values operator was used to replace the missing values with zero rather than using an average value to avoid interfering with the predictions, however, the missing values being replaced by zero can cause bias. This is probably the reason why there was a drop in the prediction of population growth in 2021 as 2020 was missing and must be the reason for the spike in 2022.

Outliers

Outliers were detected using the Detect Outlier (LOF) operator. However, the outliers were included in this research. The Arab region is known to have a lot of political instability which could have contributed to the vast number of outliers in the data for certain time periods.

Future work

In depth analysis would need to be done to confirm that Birth rate is a suitable method for predicting population growth. If Birth rate can be predicted using the time series data, it would be possible to make better financial planning within the region for improved population health.

References

- Asfour, O. & Jabbour, S. (2020) “Public health systems in the Arab World: The COVID-19 Conundrum and the way forward”, Al Sharq Strategic Research, Last viewed: 26/06/2021 (online) Available at:

https://research.sharqforum.org/2020/07/10/public-health-systems-in-the arab-world-the-covid-19-conundrum-and-the-way-forward/

- Ognjanovski, G. (2018) “Predict Population Growth Using Linear Regression — Machine Learning Easy and Fun”, Analytics Vidhya. Last accessed: 16/06/2021 (online) Available at: https://medium.com/analytics-vidhya/predict population-growth-using-linear-regression-machine-learning-d555b1ff8f38
- Schneider, U. A., Havlik, P., Schmid, E., Valin, H., Mosnier, A., Obersteiner, M., Böttcher, H., Skalský, R., Balkovič, J., Sauera, T. & Fritz, S. (2011) “Impacts of population growth, economic development, and technical change on global food production and consumption”, *Agricultural Systems,* 104 (2): 204-215.
- World Bank Group (2021) Dataset of Health, Nutrition and Population. Last updated: April 27, 2021. Accessed: 30 May 2021 (online) Available at: https://datacatalog.worldbank.org/dataset/health-nutrition-and-population statistics