DESCRIPTION OF THE PROJECT 3

The aim of this project is to find a model that is able to forecast electricity consumption of Senegal (West Africa) up to 2045 and to compare electricity demand's pattern with other environmental and social indicators in order to find possible correlations between each other.

First of all, data have been collected from the World Bank Group database, an open-source archive which contains a wide variety of data and statistics for all world's countries. For each country, World Bank provides a CSV file containing 1443 indicators describing the region under an economic point of view (GDP, foreign investments, labor force, etc.), an environmental outlook (deforestation, CO2 emissions, pollution…) plus giving information regarding the energy mix and human development indexes. Generally, the temporal range of data acquisition goes from 1961 to 2019, but this condition is not valid for all the indicators, so it is necessary to properly clean and re-order data.

Data cleaning and preparation

From the World Bank dataframe (API_SEN_DS2_en_csv_v2_2462095), 49 indicators have been selected, choosing the most relevant and representative for the purposes of the analysis and then saved in a separate file named "Data_selection". As second step, the Data_selection dataframe has been ordered sorting the columns in alphabetic order and the year's column is set to index. Then, for an easier management of the data, the file has been split in three different datafiles diving the indicators related to energy, those related to environmental issues and the ones describing financial and social topics and named "WorldBank_energy", "WorldBank_environment" and "WorldBank_social" respectively.
Since most of the columns had no values at the beginning of the time serie, WorldBank_energy and WorldBank_social have been made starting from 1971 instead of 1961, while WorldBank_environment have been left starting from 1961. Then, some of the columns of the dataframes have been converted into more readable units and later renamed:

- "Electricity consumption (kWh)" turned into "Electricity consumption (GWh)"
- "Foreign direct investment, net inflows (BoP, current US$)" turned into "Foreign direct investment, net inflows (BoP, current million US$)"
- "GDP (current US$)" turned into "GDP (current million US$)"
- "Population, total" turned into "Population, total (million people)"

Due to the fact that there are some gaps in WorldBank_social, an interpolation is required so that the blank cells are filled with linearly approximated values. Finally, the three datasets are saved as new CSV files.

Data display and interpretation

In this section of the project several plots are displayed, to see the behaviour of the analysed indicators under a temporal scale. At a first glance, it can be seen that the indicators show a general upward tendency, so an increase of GDP, population and life expectancy which determine higher primary energy and electricity consumptions, causing an increase of CO2 emissions and pollution. Senegal, in fact, similarly to most of Sub-Saharian countries, is strongly pushing on the development of its society: the economy is sharply growing together with the population and its wellbeing so this seeks very strong energy requirements.
Looking at the graph describing the access to electricity, it is worth noticing that urban population is highly favoured with respect to the rural context, but they are both increasing with the same growth rate reaching in 2019 almost 100% and 50% of population for urban and rural respectively.

Many indicators (such as electricity consumption, population, GDP, cropland and internet usage) show the classic "Hokey stick" growth, so this suggests a strong correlation among them. Therefore, after data

display, it is interesting to plot the variables in scatter plots to understand the degree of correlation between the variables and electricity consumption. In addition to the scatter plots, the correlation coefficients between the variables are calculated, using Pearson, Spearman and Kendall methods.
As a confirmation of the previous data display, the indicator which are mostly related to electricity consumption are the population, GDP, CO2 emissions and deforestation, with Spearman coefficients always greater than 92%.
Regarding the correlation electricity consumption and school enrolment, the results provided by the scatter plot are rather interesting: it can be seen that school enrolment sharply grows between 100 and 130 kWh/capita, then the growth gets milder for higher values of electricity consumption. Thus, we can define a "critical consumption" (at around 128 kWh/capita), which is the minimum energy required to allow a great percentage of children enrolled to primary school. In addition to this, a higher electricity consumption guarantees better gender parity: at low consumptions male children are more likely to enrol to school with respect to females, while after the critical point the percentage is more homogeneous.

Feature selection

In order to forecast electricity consumption, it is firstly necessary to determine the most relevant features through Feature Selection. Here, we added electricity consumption plus electricity consumption -1 (the consumption shifted of one position) to every dataframe, in order to find the features which are mostly related to electricity in the three files. In this case, the Filter method with kBest has been adopted, both using the f-test ANOVA and mutual info regression and selecting k=3.
Here the results of the implementation of the two regression for the three datafiles, selecting those which gained the greatest score:

1) f-test ANOVA
   WorldBank_energy:
   I) electricity consumption -1
   II) Electricity production from hydroelectric sources (% of total)
   III) Energy intensity level of primary energy (MJ/$2011 PPP GDP)

   WorldBank_environment:
   I) electricity consumption -1
   II) year
   III) Forest area

   WorldBank_social:
   I) electricity consumption -1
   II) School enrollment, primary (% gross)
   III) School enrollment, primary, female (% net)

2) Mutual info regression
   WorldBank_energy:
   I) electricity consumption-1
   II) Year
   III) Electricity production from renewable sources, excluding hydroelectric (kWh)

   WorldBank_environment:
   I) electricity consumption-1
   II) Year
   III) Permanent cropland (% of land area)

   WorldBank_social:

I) population
II) rural population
III) electricity consumption-1

Finally, the features that will be chosen for modelling energy consumption will be:

A) electricity-1
B) year
C) population (million people)
D) Permanent cropland (% of land area)
E) Energy intensity level of primary energy (MJ/$2011 PPP GDP)
F) school enrolment, primary (% gross)
G) Electricity production from hydroelectric sources (% of total)
H) Electricity production from renewable sources, excluding hydroelectric (kWh)
I) Forest area (% of land area)

Modeling and forecast

In this section of the project we implemented several ways to forecast the electricity consumption pattern of Senegal for the period 2015-2045. As a first step, we created a new dataframe containing the features chosen during Feature Selection part plus the electricity consumption, so that is possible to process the data and find the best model. Data are randomly split in training and test data, where y_train is the vector which corresponds to electricity consumption and X_train to the features that will be used to fit the model; they contain 75% of the dataset. X_test and y_test instead are the vectors used to test the model, using the remaining 25% of data. At this point it is possible to try the different models and evaluate the error between the prediction and the true data (y_test). It is possible to see that the model that provides the best results is linear regression, followed by gradient boosting and bootstrapping. Usually, linear regression is the least accurate method among all the types of regression, but in this case electricity consumption's pattern such as the selected features follow a very smooth trend, without excessive fluctuations, so a polynomial function can properly fit such data.
Now it is possible to extrapolate the main variable and its features. Due to the aforementioned conditions, a polynomial can properly describe the variables, so the python "polyfit" and "polyval" commands have been used to determine the coefficients of the function and to fit the data. In the plots we can see the orange line that fits the true data and estimates the future trend up to 2045: the features have been fitted by a second order polynomial, while for electricity consumption a fourth order polynomial has been chosen.
Aside from those methods another method was to assume a yearly relative increase of 4%.
At this point we can validate the previous estimations by using linear regression and gradient boosting and calculating the error between the polyval model and the regression model. Therefore, a new dataframe with the estimated features from 1971 to 2044 has been created and randomly split in training and test data, then the regression methods are repeated as in the previous part. To summarise, we can notice that with a fourth order polynomial in the polyval function we obtain results which are rather realistic (the polyval curve fits very well the data from 1971 to 2015) and the validation results with gradient boosting and linear regression shows errors of 9 – 15% respectively.