Machine Learning Project

Oral presentation: May, 28 and 31 2024

Dataset

The data is taken from the KAGGLE competition website; it is the "Global Data on Sustainable Energy" dataset (2000-2020), available here :
https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy.
The dataset comprises 3649 observations and 21 variables, representing various characteristics related to the energy consumption and geography of 176 countries worldwide over the years 2000 to 2020.

The variables are as follows :

- **Entity** : The name of the country or region for which the data is reported.
- **Year** : The year for which the data is reported, ranging from 2000 to 2020.
- **Access to electricity (% of population)** : The percentage of population with access to electricity.
- **Access to clean fuels for cooking (% of population)** : The percentage of the population with primary reliance on clean fuels.
- **Renewable-electricity-generating-capacity-per-capita** : Installed Renewable energy capacity per person
- **Financial flows to developing countries (US Dollars)** : Aid and assistance from developed countries for clean energy projects.
- **Renewable energy share in total final energy consumption (%)** : Percentage of renewable energy in final energy consumption.
- **Electricity from fossil fuels (TWh)** : Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours.
- **Electricity from nuclear (TWh)** : Electricity generated from nuclear power in terawatt-hours.
- **Electricity from renewables (TWh)** : Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours.
- **Low-carbon electricity (% electricity)** : Percentage of electricity from low-carbon sources (nuclear and renewables).
- **Primary energy consumption per capita (kWh/person)** : Energy consumption per person in kilowatt-hours.
- **Energy intensity level of primary energy (MJ/2011 PPP GDP)** : Energy use per unit of GDP at purchasing power parity.
- **Value-co2-emissions (metric tons per capita)** : Carbon dioxide emissions per person in metric tons.
- **Renewables (% equivalent primary energy)** : Equivalent primary energy that is derived from renewable sources.
- **GDP growth (annual %)** : Annual GDP growth rate based on constant local currency.
- **GDP per capita** : Gross domestic product per person.
- **Density (P/Km2)** : Population density in persons per square kilometer.
- **Land Area (Km2)** : Total land area in square kilometers.
- **Latitude** : Latitude of the country's centroid in decimal degrees.
- **Longitude** : Longitude of the country's centroid in decimal degrees.

The objective is to predict the variable **Value-co2-emissions** from the other variables.
Please note : As the dataset contains many missing values, a preliminary exploratory study is more than ever necessary to familiarize yourself with the data and prepare them for the modeling phase.

Questions asked

Exploratory data analysis

The first step is to explore the different variables, an essential preliminary to the analysis. Below are a few basic questions. You can complete the analysis according to your own ideas.

1. Start by checking the nature of the different variables and their encoding. Convert the variable **Year** into a qualitative variable. N.B. Curiously, the variable **Density (P/Km2)** is not considered a numerical variable. Convert it into a numerical variable, taking care not to transform decimal numbers into `NA`. For example, in R, you can use the formula: `as.numeric(gsub(",","",data$Density.n.P.Km2.))`, where `data` represents the dataset used.
2. Determine the rate of missing values for each variable.
   For this project, we propose to remove the variables with a very high rate of missing data: **Renewable-electricity-generating-capacity-per-capita**, **Financial flows to developing countries (US Dollars)** and **Renewables (% equivalent primary energy)**.
3. For the rest of the study, you will create a dataset containing only those individuals with no missing values. This will leave 2868 observations.
4. Start your exploration with a unidimensional descriptive analysis of the data. Do you think transformations of quantitative variables are relevant?
5. Visualize the great heterogeneity of CO_2 emissions between countries. Which 5 countries emit the most CO_2?
6. Continue with a multidimensional descriptive analysis. Use visualization techniques: e.g. scatterplots, correlation graphs... Analyze dependencies between quantitative variables.
7. Perform a principal component analysis of quantitative variables and interpret the results.
8. Visualize the possible dependency between the variable **Year** and the variable to be predicted.

Modelisation

We now consider the problem of predicting the variable **Value-co2-emissions** from the other variables from a machine learning point of view, i.e. by focusing on model performance. The aim is to determine the best performance we can expect, and the models that achieve it. Here are a few questions to guide you.

1. Divide the dataset with no missing data into a training sample and a test sample. Take a percentage of 20% for the test sample. Why is this step necessary when we're focusing on algorithm performance?
2. Compare the performance of a linear regression model with/without variable selection, with/without penalization, an SVM, an optimal tree, a random forest, boosting, and neural networks. Justify your choices (e.g. kernel for SVM), and carefully adjust parameters (by cross-validation). Interpret the results and quantify the potential improvement provided by nonlinear models.
3. Compare the different optimized models on your test sample. Which models perform best? How accurate are they?
4. Interpretation and feedback on data analysis: are your results consistent with the exploratory data analysis, for example in terms of the importance of variables?
5. In a second step, you can use an algorithm to fill in the missing values and repeat the modeling (for the algorithms that have proved most successful) with the completed dataset.

Terms and evaluation

You will complete the project in groups of 4 students. Assessment will be based on an oral presentation and two Jupyter notebooks (one in R and one in Python).

Assignment due : As a deliverable, each group will submit on Moodle :

- **no later than May 27 at 6 p.m.**, a zip file containing the two Jupyter notebooks (R and Python),
- **no later than 6 p.m. the day before the defense**, the slides of the presentation **in pdf format**.

Oral defense on May 28 and 31, 2024 : 20 minutes for presentation, followed by 5-10 minutes for questions. The presentation should include an introduction presenting the data and all the transformations you've performed, a brief description of the algorithms used (making it clear which hyperparameters you've optimized and how), an interpretation of the results, and a conclusion. Questions may relate to your code (so remember to open your notebooks and compile them before the presentation).

Evaluation criteria : The evaluation will take into account the quality of the oral presentation (clarity, argumentation, interpretation of results, etc.), the coherence of the study, the quality of the notebook presentation (don't forget to comment on your code), and the interpretation of results (graphs, etc.).