INFORME PROGRÉS I

Santiago Pérez Lete

Abstract– The study will consist of finding a dataset, cleaning it and adapting it to my needs. After this I will choose the independent variables and the dependent variable that we want to use in our models. Finally, I will analyse the different statistical prediction models, using the programming language "R", on the same variables of a dataset and see which one best fits the variables we choose and how the results change between each model.

Keywords– Statistical analysis, data cleaning, Python, RStudio, prediction model, multiple regression, binomial, poisson, gaussian, gamma

7.5pt

1 MOTIVATION

LAST year when we were in the third year. We took, among others, two subjects, Statistical Analysis and Machine Learning. At first I thought they had nothing to do with each other, but once we started the Machine Learning theory classes (having already taken Statistical Analysis), I realised that the basis of Machine Learning is statistics and that everything we had studied in statistical analysis was used by Machine Learning to generate its models. According to the blog Máxima Formación4, Machine Learning is a branch of statistics. Therefore, applying the wrong algorithm, not understanding the biases or limitations of an algorithm and not correctly interpreting the output are huge problems in the field of Machine Learning, or what amounts to the same thing, not having prior knowledge of statistics. In this blog, they also state that every step in a Machine Learning project requires the use of a statistical method. And because of all this and because I really liked the subject of Statistical Analysis, I find it very interesting to know and learn about these different statistical prediction models.

2 PROJECT OBJECTIVES

The objective of the TFG is to find a dataset, clean it to suit my needs and perform a statistical study of these variables using different statistical models that allow me to see how each one works and which one best fits the data I find through different analyses and metrics.

• E-mail de contacte: santipl2001@hotmail.com
• Treball tutoritzat per: Walter Andrés Ortíz Vargas (Facultad de Matemáticas UAB)
• Curs 2022/23

3 METHODOLOGY

The work methodology that I will follow will be to set a fixed schedule every week to dedicate to the TFG and thus achieve a good constancy. As for how to achieve the objectives I have decided that what has to do with data cleaning will be done using the programming language "Python" and for the part of the statistical models I will use "R Studio", this is because I am more familiar to do "data cleaning" with "Python" and for statistical issues with "R Studio". The study investigates how independent variables in each country affect the life expectancy of each country. With the aim of creating a model in which life expectancy can be predicted as a function of these variables, where different methods have been applied. Fig1 shows a summary of all the steps that were necessary to reach the objective of the study.

4 PLANNING

The planning and development of the work is set out in the following table Fig2

5 DATA DESCRIPTION AND VARIABLES

This part of the paper describes some characteristics of the countries investigated. The following subsections show the different variables chosen and an exploratory analysis of them and the relationship between them.

5.1 Data acquisition

The data used in the paper comes from The World Factbook, produced for US policymakers and coordinated throughout the US Intelligence Community, presents the basic realities about the world in which we live. These facts are shared with the people of all nations in the belief that knowledge of the truth underpins the functioning of free societies. Within

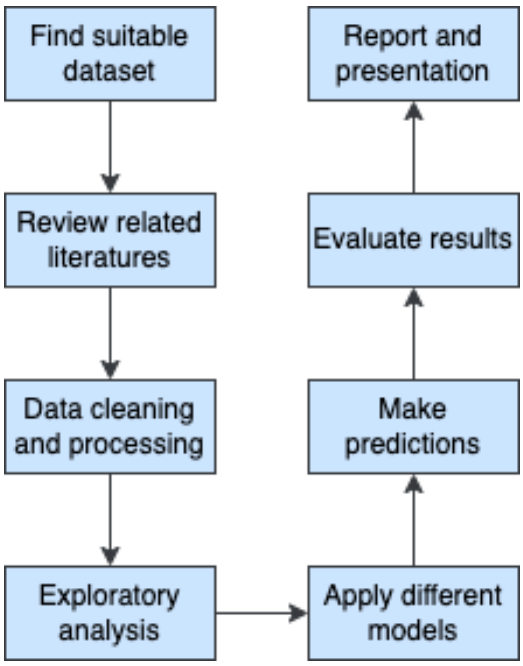

Fig. 1: Flowchart of the study's work process

Task	Description	Duration (weeks)	Level of completion
State of the art	Search for scientific articles on statistical analysis and its predictive models to see the current state of the art.	1	100%
Datasets	Find a dataset from which to obtain variables for the models. Clean the dataset and adapt it to our needs.	2	100%
Exploration	To make exploratory graphs to analyse the variables and the relationship between them.	1	90%
Apply models	Split the dataset into train and test. And apply the different prediction models and perform the anova.	4	5%
Predictions	Make predictions with the test dataset.	2	0%
Results	Analyse the results of the predictions and visualise them.	2	0%
Report and presentation	Writing the report and the final presentation.	2	0%

Fig. 2: Progress table

these records there is data on every country in the world, geographically, economically, environmentally, etc. All this is published on the website of the Central Intelligence Agency (CIA) and has been put in a dataset and published in kaggle, from where I have been able to download the csv. Given the diversity of characteristics of each country in the dataset, it was a complicated but important task to choose the variables. After trying several options such as: studying the Unemployment rate as a function of investment in education, inflation and GDP of each country. The variables chosen were:

- Dependent variable of the model:
  - Life Expectancy: Shows life expectancy from each country.
- Independent variables of the model:
  - Particulate Matter emissions: This variable indicate, in micrograms per cubic meter, the particulate matter emissions.
  - Carbon Dioxide emissions: Shows the carbon dioxide emissions in megatons.
  - Methane emissions: Shows the methane emissions in megatones.

5.2 Data cleaning

The dataset, having been extracted and created manually by the CIA, as I mentioned earlier, is not very buggy and is quite easy to process and use. Despite all this, the dataset has been subjected to a filtering and cleaning process to adapt the data to our needs and objectives. The steps that have been taken for this cleaning have been as follows:

1. The first step has been to filter from the dataset the variables we want to use for our model using Python. Therefore, a sub-dataset has been created with the name, life expectancy, particulate matter emissions, carbon dioxide emissions and methane emissions of each country. That is, the dataset created consists of all the same countries as the original dataset but only has the columns of the variables we are interested in for our model.

2. Once you have filtered the columns you want to use, with Python and its pandas library you can rename the columns, since in the original dataset, the name of each column is the subject of the column. For example, in the particulate matter emissions variable, in the original dataset the column name was: Environment: Air Pollutants - particulate matter emissions. This is done for ease of use when using the dataset for statistical analysis.

3. Subsequently, it is searched for null values in any column of any row of the dataset and in case there is a null value in a row, that row is deleted in order to avoid inconsistencies in the models and that the examples used to train it have all the independent variables with total security.

4. Once we have filtered and eliminated the null values, the next problem we encounter is that in the values of the columns there are both numerical values, which are the ones we need, and text. For example in the Life expectancy column the value is 76.3 years or in the particulate matter emissions column it says 2.12 micrograms per cubic meter. For the analysis we are only interested in the numerical value, so we proceed to delete the text in each value of each column and keep only the numbers, this as in the previous steps has been done using Python and its Pandas library.

5. To finish the work in Python, when we have the dataset suitable for our needs, it is exported in csv format to be processed in Rstudio, where we will use the dataset for statistical analysis.

6. In the project different models will be realised and not all of them require the same type of dependent variable. Therefore, 3 different variants of this dataset have been created from the previously exported dataset.

    - One of the variants, identical to the exported one, requires the dependent variable (Life expectancy) in its real values.
    - Another of the variables requires Life expectancy to be a binary variable, so the original dataset has been modified and a 1 has been added to Life expectancy in countries where it is equal to or greater than 75 years and a 0 where it is less than 75 years.
    - For the last variant, the dependent variable needs to be divided into groups, so 4 groups have been created for Life expectancy, which are distributed as follows:
        - 0 if Life expectancy is smaller than 65.
        - 1 if Life expectancy is between 65 and 72,5.
        - 2 if Life expectancy is between 72,5 and 80.
        - 3 if Life expectancy is larger than 80.

7. Finally, for each variant, the dataset was randomly divided into train and test. With a distribution of 80% for the train and 20% for the test.

5.3 Exploratory analysis

After performing the data cleaning and having the data as we are interested in for our objective, we are going to analyse the variables we will use and how they are related to each other in order to see if they would work well in a statistical prediction model. Unlike data cleaning, the exploratory analysis will be done using the R language in RStudio.

To do this, first of all a correlation matrix has been made with the variables mentioned above. This is done by calculating the correlation matrix, of the 3 types of dataset that have been created for the project (explained in the previous section) one of these matrices is calculated for each one. Here we can see that the variable ”methane emissions” is highly correlated with the variable ”carbon dioxide emissions” and that the dependent variable of the model, Life Expectancy, is highly correlated with the variable ”particulate matter emissions”. The following image shows the correlation matrix of the dataset for the multi-models (without grouping Life expectancy). Although the correlation matrix of the three types of datasets are similar. Fig3 shows the correlation matrix of multiple regression dataset

Also as an exploratory analysis and to help in the division of the binary and grouped datasets, a histogram of the Life Expectancy variable has been made. With this we can see the frequencies of Life Expectancy and help us to divide the dataset in half or in equal groups. We can see the histogram in the following image Fig4.

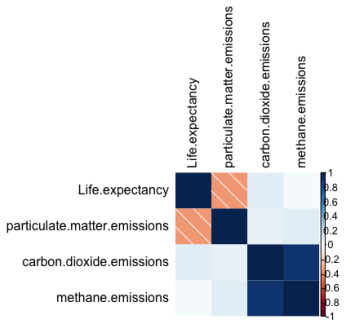

Fig. 3: Correlation matrix dataset ”multiple”

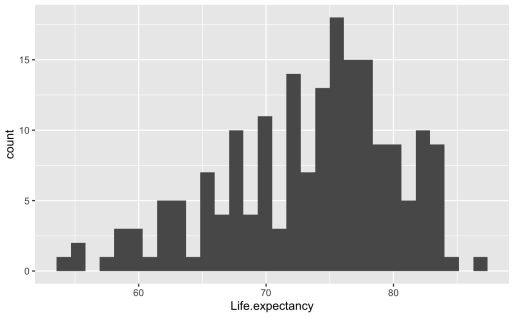

Fig. 4: Life Expectancy histogram

6 Models implementation

Still to be implemented.

6.1 Binomial

Still to be implemented.

6.2 Poisson

Still to be implemented.

6.3 Multiple

Still to be implemented.

6.4 Inverse Gaussian

Still to be implemented.

6.5 Gamma

Still to be implemented.

7 Results

8 Conclusions

Still to be implemented.

Agraïments

Still to be implemented.

REFERENCES

[1] V.PandoFernándezandR.SanMartínFernández.RegresiónLogística Multinomial. 2004.

[2] PeterDalgaard.IntroductoryStatisticswithR.2008.

[3] Kaggle[Internet],[Quoted20Febraury2023].Url: https://www.kaggle.com/

[4] MáximaFormación[Internet],[Quoted22February2023].Url: https://www.maximaformacion.es/blog-dat/el-papel-de-la-estadistica-en-el-machine-learning/#: :text=El%20Machine%20Learning%20es%20una%20rama%20de%20las%20estad%C3%ADsticas%2C%20por,el%20campo%20del%20Machine%20Learning.

[5] TowardsDataScience[Internet],[Quoted24February2023].Url: https://towardsdatascience.com/logistic-regression-detailed-overview- 46c4da4303bc

[6] Dataquest[Internet],[Quoted24Febreury2023].Url: https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/

[7] GitHub[Internet],[Quoted24Febreury2023].Url: https://github.com/santipe01/TFG_AnalisisEstadistico.git

APÈNDIX

A.1 Secció d'Apèndix

Still to be implemented.

A.2 Secció d'Apèndix

Still to be implemented.