Universitat Autònoma de Barcelona

Informe Inicial

Analysis of statistical prediction models for life expectancy

Santiago Pérez Lete

Tutor: Walter Andrés Ortiz Vargas

Grau en Enginyeria de Dades

Escola d'Enginyeria

Index

1. Summary

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.

2. 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 Formación*[4], 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.

3. 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.

4. 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".

5. Planning

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	70%
Datasets	Find a dataset from which to obtain variables for the models. Clean the dataset and adapt it to our needs.	2	50%
Exploration	To make exploratory graphs to analyse the variables and the relationship between them.	1	10%
Apply models	Split the dataset into train and test. And apply the different prediction models and perform the anova.	4	0%
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%

Table 1. Planning of project tasks.

6. Bibliography

1. V. Pando Fernández and R. San Martín Fernández. Regresión Logística Multinomial. 2004.
2. Peter Dalgaard. Introductory Statistics with R. 2008.
3. Kaggle [Internet], [Quoted 20 Febraury 2023]. Url: https://www.kaggle.com/
4. *Máxima Formación* [Internet], [Quoted 22 February 2023]. 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. Towards Data Science [Internet], [Quoted 24 February 2023]. Url: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
6. Dataquest [Internet], [Quoted 24 Febreury 2023]. Url: https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/