Palmer Penguins, discovering a new classic Project Proposal

Group 11
Spring 2024
Machine Learning
Master's Degree in Data Science
UPC - BarcelonaTech

Team members

Our group is formed by Adrià Casanova Lloveras and Víctor García Pizarro, both students from group 11.

Problem and dataset we are going to work on

We have chosen the palmer penguins dataset to study penguin species classification. It represents data from 344 observations of penguins in the Palmer Archipelago collected and made available by Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program during 2020. Given categorical and continuous variables such as island, sex, flipper length (mm) or body mass (g), our goal is to predict the species of a penguin, which is a variable of the dataset as well.

The raw data will be obtained from the *palmerpenguins* module, running the function *load_penguins_raw()*. Given it is a suggested data source for this project, we will not provide information on the data.

Why have we chosen this problem

Palmer penguins is a relatively new dataset that aims to become a classical source for data science learning, substituting the ubiquitous iris dataset. It contains more variables than iris and some missing values. Furthermore, it has unequal sample sizes, and Simpson's Paradox examples. Importantly, the Palmer penguins dataset encompasses real-world information derived from several species with regional breeding populations notably responding to environmental change, so it is a currently relevant topic. Finally, a member of the group, Víctor, is a graduate in Marine Science, so he can provide domain knowledge to the analysis.

Previous work in this problem

Many previous studies of this dataset can be found on the internet, such as in https://medium.com/@Fortune_/visualizing-the-palmerpenguins-dataset-d3d70bb619b4, https://www.neuraldesigner.com/learning/examples/palmer-penguins/ or https://www.kaggle.com/code/florianspire/palmer-penguins-data-preprocessing-and-analysis. We will try to find new conclusions and perform different models based on these articles.

Project Title

Our project will be named “Palmer Penguins, discovering a new classic”.