ML Project Proposal: **Palmer Penguins Dataset**

Authors: Silvia Ferrer & Iñigo Pikabea & Ignacio Lloret

Problem Statement

We aim to apply preprocessing techniques and both supervised and unsupervised learning techniques on our dataset in order to classify Palmer penguins by their characteristics. Moreover, our goal is to learn how to implement all these steps with python and the different libraries used for ML methods.

Dataset Selection

Our choice is motivated by the dataset's simplicity, which is ideal for advanced machine learning techniques, aligning with our course objective to implement ML methods. This approach also ensures detailed traceability of our analysis process. Besides, we find the topic interesting and different.

Previous Work

We reviewed existing work focusing on feature selection, and data exploration and visualization. Notable references include:

- There are many small projects in Kaggle Codes which focus on varied approaches and are implemented with different programming languages.
- An Analytics Vidhya blog post, "Data Exploration and Visualisation Using Palmer Penguins Dataset", discussing various data exploration and visualization techniques.

Data Overview

The dataset for our analysis comprises observations of three penguin species, collated from multiple studies conducted in the Antarctic region. The data has the following structure:

Number of records	220
Number of variables	17
Number of categorical variables	14 + 1 datetime
Number of numerical variables	2
Number of binary variables	1

Variables include both categorical and numerical types, such as 'Species' (3 levels representing different penguin species), island location, physical measurements (e.g., culmen length and depth, flipper length, body mass), and environmental isotopes. Our target variable will be 'Species'.