A logo for a company

Description automatically generated

Python Machine Learning Project

Project 3

Abed E. D. ADOUNKPE, Nazim Akram HADJ SEYD, Zarzi Rissel Mohamed El Moundhir, Alexandre Benamenyo, Chloé Grangeon

DSTI

## Introduction

The primary objective of this project is to construct a comprehensive pipeline for forecasting sleep variables in mammals using an extensive dataset. The diverse sleeping patterns observed among mammalian species prompt inquiries into the underlying factors influencing these patterns.

Despite being a fundamental activity for all mammals, sleep characteristics vary significantly across species. This project is centered around investigating and predicting sleep attributes utilizing a dataset containing information on 87 mammalian species. The dataset encompasses a broad spectrum of general, biological, and ecological attributes, along with specific variables related to sleep patterns.

## Data Exploration and Analysis

The dataset comprises 87 entries with 17 features, categorized into 6 categorical (e.g., Family, Species, Genus, Order, Vore, Conservation) and 11 numerical values.

Initial exploration revealed missing values in several columns, indicating the need for careful preprocessing. A correlation analysis, visualized through a heatmap, highlighted significant negative correlations between Total Sleep and factors such as Exposure, Gestation, and Danger. This suggests that mammals facing greater exposure to predators, longer gestation periods, or higher levels of danger tend to sleep less. Graphs elucidated these relationships, showing a decrease in Total Sleep with increasing levels of exposure and danger.

However, the presence of missing values complicated the analysis, underscoring the importance of accurate data imputation.

## Data Preprocessing

Data preprocessing involved addressing missing values without significantly reducing the dataset size. Missing values were filled using online documentation for some categorical features and an iterative imputer for numerical features. The preprocessing phase also included encoding categorical variables to numerical formats, enabling their use in machine learning models. The multivariate imputation method was chosen for its efficiency in predicting missing values based on other features, optimizing feature combinations to minimize the MSE.

## Data Validation and Correction

A critical validation step was performed to ensure the accuracy of TotalSleep values. Visualizations were employed to confirm the validity of data corrections. Specifically, a plot comparing TotalSleep and corrected RealTotSleep values provided visual confirmation of the accuracy of corrections.

## Encoding Categorical Variables

Categorical variables were encoded into integers using LabelEncoder to prepare them for further analysis. This step was crucial for transforming string values into numerical format, facilitating subsequent modeling tasks.

## Predictive Modeling for Missing Data

Advanced techniques such as univariate testing and variable selection were utilized to predict missing values for Predation, Exposure, Danger, Dreaming, and NonDreaming variables. The approach involved identifying variables related to each target variable and employing iterative imputation to make predictions.

## IV. Machine Learning

The machine learning phase began with a decision to exclude certain features to avoid multicollinearity, based on their correlations with Total Sleep and Dreaming. Various models were explored:

* Linear Regression: Initially chosen for its simplicity and the assumption of linear relationships between features and target variables. However, it showed limited success, likely due to the dataset's small size.
* SMOTE Application: To overcome data limitations, SMOTE was applied, creating synthetic samples to enhance the dataset. This was particularly effective, as it maintained the categorical distribution while expanding the dataset to 1000 rows.
* Random Forest and XGBoost Models: Both models improved predictions over linear regression, benefiting from the larger, SMOTE-enhanced dataset. Feature importance metrics guided the selection of variables for these models.
* Neural Network Model: This model offered the most accurate predictions, with MSE values indicating a strong model fit without signs of overfitting. The choice of neurons and epochs was optimized based on performance and execution time.

## Conclusion

The project's comprehensive approach, from data preprocessing to machine learning model evaluation, highlighted the efficacy of SMOTE in enhancing dataset quality for machine learning applications. The comparative analysis of different models underscored the neural network's superior ability to predict Total Sleep and Dreaming durations accurately. This machine learning project not only demonstrates the potential of advanced analytical techniques in understanding mammalian sleep patterns but also sets a precedent for future research in physiological phenomena prediction using machine learning.