DSTI

Python Machine Learning Project

PROJECT 3: PREDICTING SLEEP VARIABLES IN MAMMALS
GROUP 5

ABED E. D. ADOUNKPE, NAZIM AKRAM HADJ SEYD, ZARZI RISSEL MOHAMED EL MOUNDHIR, ALEXANDRE BENAMENYO, CHLOÉ GRANGEON

GitHub link to the project: https://github.com/AbedAdounkpe/PredSleep/

Introduction

The primary objective of this project is to construct a comprehensive pipeline for forecasting sleep variables in mammals using a 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 according to the species. This project is centered around investigating and predicting sleep attributes utilizing a dataset containing information on 87 mammalian species. It encompasses a broad spectrum of general, ecological, and biological 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 (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 of the dataset, underscoring the importance of accurate data imputation.

Data Preprocessing

Data preprocessing consisted of processing missing values without reducing the size of the data set. Missing values were filled in using the online documentation for some categorical characteristics and an iterative imputer for numerical characteristics. The preprocessing phase also included encoding categorical variables in digital formats, allowing their use in machine learning models. The multivariate imputation method was chosen for its effectiveness in predicting missing values based on other characteristics, optimizing the combinations of characteristics to minimize 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 in order to prepare them for further analysis. This step was important 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, Danger, Dreaming, Exposure 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 started with a decision to exclude certain features to avoid multicollinearity, based on their correlations with Total Sleep and Dreaming. For non-linear models, features' importances were studied.

Various models were explored:

- Linear Regression: Initially chosen for its simplicity and the assumption of linear relationships between features and target variables. But this model showed limited success, likely due to the dataset's small size.
- SMOTE Application: To overcome data limitations, SMOTE was applied, creating synthetic samples to improve the dataset. This was particularly effective, as it maintained the categorical distribution while extending the dataset to 1000 rows. Linear Regression was applied again, this time, on the SMOTE dataframe, but the MSE and RMSE values suggested again poor accuracy between the predicted and the actual values.
- 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 and RMSE values indicating a strong model fit without signs of overfitting. The choice of neurons and epochs was optimized, based on performance and execution time.

Total Sleep durations

Model	MSE (hours^2)	RMSE (hours)	RMSE (minutes)
Linear Regression	12.11	3.48	209
Linear Regression on SMOTE dataframe	4.62	2.15	129
Random Forest on SMOTE dataframe	0.72	0.85	51
XGboost on SMOTE dataframe	0.64	0.8	48
Neural Network on SMOTE dataframe	0.25	0.5	30

Dreaming durations

Model	MSE (hours^2)	RMSE (hours)	RMSE (minutes)
Linear Regression	3.11	1.76	106
Linear Regression on SMOTE dataframe	0.62	0.79	47
Random Forest on SMOTE dataframe	0.09	0.30	18
XGboost on SMOTE dataframe	0.08	0.29	17
Neural Network on SMOTE dataframe	0.04	0.21	13

Conclusion

The overall approach of the project, from data preprocessing to the evaluation of the machine learning model, highlighted the effectiveness of SMOTE in improving the quality of data sets for machine learning applications. Comparative analysis of different models has highlighted the superior ability of the neural network to accurately predict total and dream sleep times. This machine learning project not only demonstrates the potential of advanced analytical techniques to understand mammalian sleep patterns, but also sets a precedent for future research into physiological phenomena using machine learning.