Data ScienceTech Institute

Course: Machine Learning with Python

Project: Predicting sleep variables in mammals

Student:

Germain Meli Tsamoh germain.meli-tsamoh@edu.dsti.institute **Github:**

https://github.com/Gtmel/Sleep_regression

Abir Lammari abir.lammari@edu.dsti.institute **Github:**
https://github.com/Abirlmr/ML-Project

Yuting Wu yuting.wu@edu.dsti.institute **Github: https://github.com/Yw0991/ML-Project**

Professor: Hanna Abi Akl

Instructor: Christophe Bécavin

Project Report: Sleep Regression Analysis

Objective

This project's main objective was to develop a machine learning model to predict sleep patterns based on the attributes of the animals. Given the focus on sleep regression, the project aims to uncover insights into how different factors influence sleep and dreaming duration. The libraries used in this model are: numpy, pandas, matplotlib, seaborn, scikit-learn.

Data Preprocessing

- Addressing missing values: For certain attributes such as lifespan, Gestation and Conservation, we searched on the internet (Wikipedia) to fill in the missing values.
- Imputation of missing values, particularly in the BrainWt, Predation, Exposure, and Danger columns, through imputation strategies such as linear regression for BrainWt and median imputation for ecological factors based on dietary preferences (Vore).
- Handling highly skewed distributions in features like BodyWt, Gestation and LifeSpan through logarithmic transformations to normalize their distributions.

Feature engineering

- The dataset underwent significant feature engineering, including the creation of dummy variables for the categorical **Vore** column
- A mapping was into numerical values was applied to the **Conservation** status keeping the existing relation of increasing extinction danger between the categories.
- Log transformations were applied to several skewed features to improve model performance and interpretability, also reducing the outliers such as Gestation and LifeSpan.

Unusable categories and redundant columns were removed to focus the analysis on the most impactful predictors. We excluded Species, Genus, Family, Order as they contained many different categories without much information, Vore, Conservation as they were replaced by new features, Predation which was multicollinear with Danger and Exposure and finally Awake and NonDreaming, redundant with the variables to predict.

Model Training

- Splitting the dataset into training (75%) and test sets (25%), which is a golden rule for evaluating machine learning models, this approach allows us to find the model hyper-parameter and estimate the generalization performance.

Two linear regression models were developed:

- **Total Sleep Time Prediction**: The first model predicted total sleep time based on body weight, brain weight, lifespan, gestation period, ecological dangers, dietary habits, and conservation status.
- **Dreaming Time Prediction**: The second model predicted dreaming time, incorporating the predicted total sleep time from the first model as an additional predictor.
- Evaluation metrics included Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), which provided insights into the models accuracy and performance. The results are:

	MAPE	RSME
Total_Sleep	0.362	3.653
Dreaming	0.473	0.789

Model Evaluation

- Evaluation of the model: the models performance was conducted using predicted vs. actual sleep times on the test dataset. Typical evaluation metrics for regression tasks include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Discussion

This project successfully established separate regression models for predicting "TotalSleep" and "Dreaming" duration. While the models demonstrate promise, there is further room for improvement. Techniques like fine-tuning (hyperparameter) and feature selection can be explored to potentially enhance model performance.

Findings and Conclusions

The study has developed predictive models for animal sleep patterns, successfully selected the attributes that influence sleep duration and quality of sleep across species: BodyWt, BrainWt, LifeSpan, Gestation, Danger, Exposure,

The findings underscore the importance of physiological and ecological characteristics in sleep behavior.