Detailed Machine Learning Project Report on Predicting Sleep Variables in Mammals

Introduction

In a time when research that mixes different fields like computer science and biology is growing fast, our project shows how powerful machine learning (ML) can be in figuring out the sleep patterns of mammals. We aim to predict the 'Dreaming' and 'TotalSleep' times for many kinds of mammals using a detailed dataset. This dataset helps us look into how different biological and environmental factors affect sleep. Through this project, we're not just trying to learn more about biology but also to create a new way to study how animals sleep using computers.

Data Overview and Preprocessing

The cornerstone of our machine learning project is the robust dataset we employ, which includes a broad spectrum of mammals. Each is detailed with ecological and biological characteristics, paired with their sleep data, forming the basis of our analysis. Prior to analysis, the dataset underwent rigorous preprocessing to ensure its suitability for modeling, involving:

- Outlier Detection and Remediation: We employed the Interquartile Range (IQR) method to identify and mitigate outliers, ensuring that our analysis was not skewed by anomalous values.
- Missing Value Imputation: Advanced imputation techniques, such as K-Nearest Neighbors (KNN), were applied to estimate missing values, preserving the dataset's integrity.
- Categorical Variable Encoding: Techniques such as one-hot encoding were utilized to convert categorical variables into a machine-readable format, facilitating their inclusion in our analysis.

Exploratory Data Analysis (EDA)

Our thorough EDA played a crucial role in clarifying the dataset's structure and the relationships between different variables. This process was pivotal for establishing a robust analytical foundation, detailed as follows:

Addressing Missing Values

A key aspect of our preprocessing within EDA was the targeted handling of missing data in crucial columns: LifeSpan, BodyWt, and Gestation. Recognizing the importance of these variables in our analysis, we employed sophisticated imputation techniques to preserve the dataset's integrity and maintain statistical validity:

- LifeSpan Imputation: Given the biological significance of lifespan across mammalian species, we utilized a model-based imputation approach, considering related attributes like taxonomy and habitat to infer missing lifespan values accurately.

- BodyWt Imputation: For the BodyWt column, we applied a regression imputation technique, leveraging the known correlations between body weight and other size-related metrics to estimate missing values, ensuring that our imputations were biologically plausible and consistent with observed empirical relationships.
- Gestation Imputation: In addressing missing gestation periods, we adopted a similar model-based approach, taking into account related reproductive and developmental traits to provide sensible and contextually informed estimations.

Distribution Analysis and Correlational Assessment

Following the imputation:

- We reassessed the distributions of 'Dreaming', 'TotalSleep', as well as the newly imputed LifeSpan, BodyWt, and Gestation columns to ensure that our imputation methods did not introduce any biases or distortions.
- Subsequent correlation analyses, particularly involving the imputed columns, were crucial for validating the appropriateness of our imputation strategies and for understanding how these key biological factors interact with sleep variables.

Visualization and Predictors' Relevance

- Enhanced scatter plots and visualizations post-imputation provided deeper insights into the relationships and potential predictive power of LifeSpan, BodyWt, and Gestation in relation to sleep variables.
- The relevance of these predictors, now robustly imputed and analyzed, was carefully evaluated, setting a solid stage for the precise feature engineering and model development phases that followed.

Our in-depth EDA, with its precise approach to addressing missing data, confirmed the dataset's completeness and analytical reliability, allowing for thorough investigation of mammalian sleep trends.

Feature Engineering and Selection

Leveraging domain knowledge and insights from EDA, we engineered new features to capture complex interactions that could influence sleep patterns. This process involved:

- Synthesizing Interaction Terms: We created interaction features to model the combined effects of variables such as body weight and brain weight on sleep duration.
- Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) were applied to reduce the feature space, focusing on components that explained a significant variance in sleep metrics.
- Feature Selection: Employing techniques such as Recursive Feature Elimination (RFE), we identified and retained features with the highest predictive value, discarding those that contributed little to model accuracy.

Model Training and Evaluation

We selected Random Forest and XGBoost for their proven effectiveness in handling complex datasets. The training process was meticulously documented, highlighting:

- Model Configuration: Details on the configuration of each algorithm, including the rationale behind the choice of hyperparameters.
- Performance Metrics: In-depth analysis of model performance, using metrics such as accuracy, precision, recall, and the F1 score. The significance of each metric in the context of our study was thoroughly explained.
- Cross-Validation: We employed stratified k-fold cross-validation to ensure our models' performance was consistent across different subsets of the data, therefore enhancing their generalizability.

Conclusions and Recommendations

Our study highlights the complex connections between mammalian sleep patterns and various ecological and biological influences. The models we created demonstrate the effectiveness of machine learning in predicting sleep metrics and the potential of analytical techniques to provide new insights into biological studies.

Future research directions include:

- Expanding the dataset to include a wider range of species and additional variables, potentially uncovering new insights into sleep biology.
- Exploring more complex modeling techniques, such as deep learning, to capture non-linear relationships and interactions in greater detail.
- Investigating the applicability of our findings in related fields, such as conservation biology and animal behavior, to foster interdisciplinary collaborations.

The Random Forest model seems to be the most effective for “TotalSleep” and “Dreaming” variables, demonstrating the lowest MSE and MAE, which indicates superior accuracy. It also boasts a relatively high R2 score and a moderate Pi-score in comparison to other models. Consequently, we opt to implement the Random Forest model. However, it's crucial to acknowledge that the model targeting the "Dreaming" variable surpasses the one for "TotalSleep" across all performance metrics. Hence, the predictive accuracy for "Dreaming" significantly outstrips that for "TotalSleep."

This report provides a thorough overview of how we model mammalian sleep patterns, serving as a guide for upcoming research that intersects machine learning with biology.