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 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 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 Manual Filling: Given the biological significance of lifespan across mammalian species, we manually filled in the missing values, considering related attributes like taxonomy and habitat to guide our decisions and ensure accurate estimations.

- BrainWt Manual Filling: For the BrainWt column, we performed manual imputation. This process ensures that our data handling is precise and accurately reflects the significance of the variable in question.
- Gestation Manual Filling: Although we attempted a linear regression to estimate missing gestation periods, it was unsuccessful. Consequently, we opted for manual filling, drawing on available knowledge to make informed estimations and maintain the coherence and reliability of our dataset.

Distribution Analysis and Correlational Assessment

Following the imputation:

- We reassessed the distributions of 'Dreaming', 'TotalSleep', as well as the newly imputed LifeSpan 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 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 understand how combined effects influence sleep duration. Notably, we introduced a variable 'Ratio,' calculated as Dreaming/TotalSleep, which was instrumental in filling the missing values in the 'Dreaming' column. Although this 'Ratio' variable provided valuable initial insights for data completion, it was not retained for the subsequent EDA process. This approach allowed us to maintain focus on primary variables while ensuring the integrity of our dataset.

• Feature Selection: Our approach to feature selection involved discarding overly granular columns (such as 'species' and 'Genus') and those we were unable to reliably impute (such as 'Conservation'). This strategy was aimed at honing in on features that offered the most

substantial predictive value and maintaining model accuracy without the application of Recursive Feature Elimination (RFE). Our focus was on ensuring that the remaining features were relevant and contributed meaningfully to our analyses.

Model Training and Evaluation

We selected Random Forest for its proven effectiveness in handling complex datasets, treating our problem as a regression task to predict continuous values. The training process was meticulously documented, highlighting:

- Model Configuration: We meticulously detailed the configuration of each algorithm, emphasizing the tuning of hyperparameters for optimal performance. Given our focus on a regression task aimed at predicting continuous outcomes, we systematically experimented with various hyperparameter settings for each model. This rigorous tuning process was essential to ascertain the configurations that best fit the models, ensuring precise and reliable predictions for continuous variables.
- Performance Metrics: In-depth analysis of model performance was conducted, using metrics appropriate for regression analysis such as MSE, MAE, Pi-Score and the R2-score. The significance of each metric in the context of our study was explained, underlining their relevance in assessing the performance of regression models.

Cross-Validation: To ensure the robustness and generalizability of our models, we employed stratified k-fold cross-validation. This method was particularly crucial given our approach to treat the problem as a regression, ensuring our models' performance was consistent across different subsets of the data.

Additionally, we tested various algorithms beyond Random Forest to identify the most effective approach for our regression problem, ensuring a comprehensive evaluation of potential models.

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 and minimizing the biases of the current dataset.
- 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 has demonstrated exceptional effectiveness for the 'TotalSleep' and 'Dreaming' variables, showcasing the lowest MSE and MAE, which signify superior accuracy. Additionally, this model has achieved a relatively high R2 score and a moderate Pi-score, outperforming other models in our comparative analysis. Therefore, we have decided to proceed with the Random Forest model for our further analyses and predictions. It is, however, important to highlight that within our assessments, the Random Forest model specifically applied to the 'Dreaming' variable exhibited superior performance across all metrics compared to the one for 'TotalSleep.' This distinction underscores a notably higher predictive accuracy for 'Dreaming,' establishing it as a more robust model outcome within our study.

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.