**Course: Python Machine Learning Labs**

**Project: Predicting sleep variables in mammals**

**Team Lead:** Mohamed Jouhari, [mohamed.jouhari@edu.dsti.institute](mailto:mohamed.jouhari@edu.dsti.institute)

**Team Members**: Hugo Beffeyte, Kawtar Abidine, Sourena Mohit Tabatabaie, Brandt Olson

**Summary**

For this project, we undertook an analytical journey to see if it was possible to model mammalian sleep patterns. Utilizing data science methodologies, our aim was to construct a predictive model that could explain the factors influencing mammalian total sleep duration.

We have also modelled mammals’ total dream time noting that REM dreaming data has been largely derived from captive animals where the environment is likely a factor and that observed, i.e. non-EEG collected, REM sleep data in wild animals is difficult to capture and to the exclusion of NREM dreaming1. So, why we have been able to produce a model, its predictive utility is debatable at present.

While this project has been an academic exercise in applied machine learning; it has been insightful exercise for the team in understanding an aspect of animal behaviour none of had considered previously, and in enhancing our knowledge of mammalian biology. Food for thought as we dream of elephants and wonder if they dream of us.

**1. Data Analysis and Feature Selection**

**Data Processing and Cleaning**

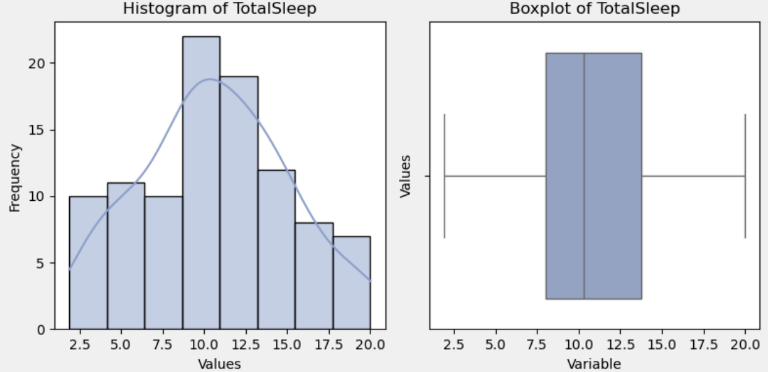
* **Dataset Overview**: The dataset provided for this project encompasses a diverse set of attributes, such as order, life span and weight, which have been considered as potentially having a relationship to mammalian sleep patterns.
* **Data Processing:** Raw data, in the form of CSV, was imported into Python using pandas to make it available for analysis.
* **Data Cleaning**: Columns with excessive missing values were identified as targets for either removal or imputation.

A purple and yellow chart

Description automatically generated

**Data Analysis and Visualization**

* A graph showing a number of species per order

  Description automatically generated**Statistical Exploration**: Our exploratory phase employed descriptive statistics to gain a preliminary understanding of the data, such as the number of species present by order.
* **Data Visualization**: We employed a variety of plots— scatter plots for initial relationship gauging, pair plots for multi-dimensional analysis, histograms and boxplots for distribution assessment, such as the distribution of total sleep values and variables.
* A screenshot of a graph

  Description automatically generated**Insights Gained**: These visual tools, such as Pearson and Spearman Correlations, were instrumental in identifying key trends and patterns, crucial for subsequent modelling stages.

**Feature Selection and Engineering**

* **A comparison of a graph

  Description automatically generatedInnovative Feature Engineering**: We introduced new features, such as the logarithmic transformation of body weight, to better capture the complex relationships in the data.
* **Prudent Feature Pruning**: To enhance model performance, we eliminated redundant features such as awake, the inverse of total sleep, and less informative features, such as conservation, (what is excluded and why?), focusing on variables with substantial influence on sleep patterns.
* **Rationale for Choices**: Selections were underpinned by a combination of statistical significance and biological relevance, ensuring that our model remained grounded in empirical realities.
* **[Placeholder for Feature Importance Chart]** (has anyone done a feature ranking?)

**2. Model Training and Evaluation**

**Model Training Approach**

* **Linear Regression Model**: The cornerstone of our analysis was a linear regression model, prized for its simplicity and interpretability—attributes crucial for stakeholder comprehension.
* **Exploratory Polynomial Regression**: To ensure thoroughness, we also explored polynomial regression, allowing us to investigate potential non-linear relationships without overly complicating the model.

**Robust Model Evaluation**

* **Rigorous Evaluation Metrics**: Our model's efficacy was scrutinized using Mean Squared Error (MSE) and R-squared metrics—standard, yet powerful indicators in regression analysis.
* **Interpretation of Results**: The model exhibited a moderate R-squared value, signifying a respectable fit. The close alignment of linear and polynomial regression lines in our residual analysis further endorsed the linear model's appropriateness.
* **[Placeholder for Model Evaluation Graphs]**

**3. Detailed Project Report**

This project has provided great insight for the team with regards to the potential of data science to extract academic insights from complex biological data. Our initial ‘guess’ was that the thermodynamics and associated metabolic rate for large animals vs small animals would be at play, that sleeping as a means of energy conversation could be a factor. While we did not have the data to confirm or not this ‘thermodynamic guess’, the linear regression model shed surprisingly light on the link between mammal size and the total sleep time in mammals, offering a window into their ecological adaptations and survival strategies.

**4. Ensuring Project Reproducibility**

* **Comprehensive Requirements File**: A meticulously compiled requirements.txt file ensures anyone can replicate our analysis, fostering transparency and scientific rigor.
* **Detailed README File**: Accompanying the project is an instructive README file, detailing the setup and execution process, making the project accessible to a broad audience.
* **[Placeholder for Requirements and README Snippets]**

**5. Project Hosting and Deployment**

* **GitHub Repository**: The project's home on GitHub not only ensures wide accessibility but also fosters collaborative improvement and review.
* **Structured for Accessibility**: The repository's architecture is designed for ease of navigation, ensuring stakeholders can effortlessly access and understand our work.
* **[Placeholder for GitHub Repository Interface or Link]**

**Conclusion**

Our journey into the domain of mammalian sleep patterns highlighted for us the intersection of data science and biology. In applying data science we were able to understand the type of insights which can be gleaned and how they can contribute to progressing scientific discovery.

[**End of Report**]

**Citations**

1. Rattenborg NC, de la Iglesia HO, Kempenaers B, Lesku JA, Meerlo P, Scriba MF. 2017 Sleep research goes wild: new methods and approaches to investigate the ecology, evolution and functions of sleep. Phil. Trans. R. Soc. B 372: 20160251. http://dx.doi.org/10.1098/rstb.2016.0251