**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 the intricacies of mammalian sleep patterns. Utilizing advanced 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, however, we note from a review of the literature that REM sleep data has been largely derived from visual observation and in doing so is highly subjective and also does not capture the dreaming that also occurs during NREM sleep. So why we have been able to produce a model, its predictive utility is debatable. (Note: Brandt to add citations for this section.)

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, 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**: We leveraged a rich dataset encompassing diverse attributes related to mammalian sleep patterns.
* **Rigorous Data Cleaning**: Initial steps included meticulous data cleaning, vital for the integrity of our analysis. This process involved removing columns with excessive missing values or irrelevant information and imputing missing data in key columns.
* **[Placeholder for Data Cleaning Flowchart]**

**Exploratory Data Analysis and Visualization**

* **Statistical Exploration**: Our exploratory phase employed descriptive statistics to gain a preliminary understanding of the data, such as mean sleep times and body weights.
* **Data Visualization**: We employed a variety of plots—histograms for distribution assessment, scatter plots for initial relationship gauging, and pairplots for multi-dimensional analysis.
* **Insights Gained**: These visual tools were instrumental in identifying key trends and patterns, crucial for subsequent modeling stages.
* **[Placeholder for Exploratory Data Visuals]**

**Feature Selection and Engineering**

* **Innovative 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 and less informative features, 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]**

**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 foray into the domain of mammalian sleep patterns highlights the intersection of data science and biology. The insights gleaned not only contribute to the scientific community but also pave the way for future research endeavors in this fascinating field.

[**End of Report**]