Applied MSc in Data Analytics

Applied MSc in Data Science & Artificial Intelligence

Applied MSc in Data Engineering & Artificial Intelligence

Course: Python Machine Learning Labs

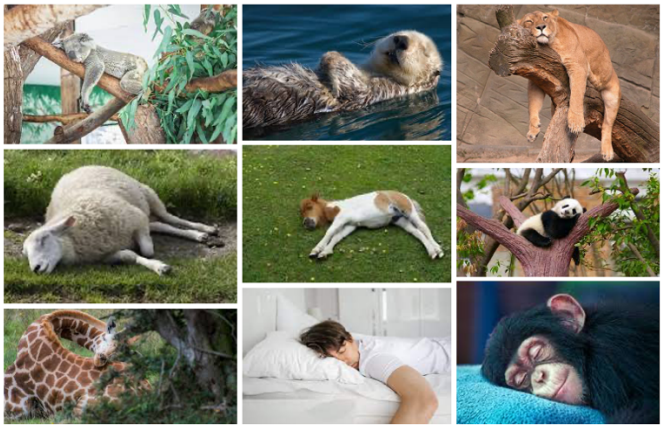
Project: Predicting sleep variables in mammals

Instructor: Christophe Bécavin

Project Overview:

The aim of the course project is to ensure students are comfortable enough developing an end-to-end pipeline to answer a given problem or use case.
The project is a group project. Since this is a mixed course (DA/DS/DE), students are encouraged to form mixed groups to benefit from the competences of their teammates when working on different components of the pipeline. An ideal group is a group of 3 members: 1 Data Analyst, 1 Data Scientist and 1 Data Engineer. However, since this is not an ideal world, and group formation relies heavily on student distribution, groups of 2-4 are allowed (group diversity is still **highly encouraged**).

- Project: Predicting sleep variables in mammals



Project Summary:

Introduction

Sleeping is a common activity for all the mammals. But there are huge discrepancies in the characteristics of their sleep. For example, some mammals sleep 10% of their day, others sleep 80% of their day. Many questions can be asked:

- Does a large mammal have a better sleep than a small mammal?
- Do predators have a better sleep than preys?
- Who has the longer dreams?

All these questions can be answered with the dataset presented in this project.

The dataset - Sleep_merged.tsv

The dataset provides sleep attributes for 87 mammals. It provides general attributes of the species:

Species	name of the species
Order	lower taxonomic rank
Genus	higher taxonomic rank
Vore	Is it carnivore, omnivore, or herbivore?
Conservation	the conservation status of the mammal in the International Union for Conservation of Nature categories

Biological attributes of the species:

BodyWt	body weight (kg)
BrainWt	brain weight (g)
LifeSpan	maximum life span (years)
Gestation	gestation time (days)

Ecological attributes of the species:

Predation	predation index (1-5) 1 = minimum (least likely to be preyed upon); 5 = maximum (most likely to be preyed upon)
Exposure	sleep exposure index (1-5) 1 = least exposed (e.g. animal sleeps in a well-protected den); 5 = most exposed
Danger	overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals); 5 = most danger (from other animals)

Sleep attributes of the species:

TotalSleep	total sleep, sum of slow wave and paradoxical sleep (hrs/day)
Awake	amount of time spent awake (hrs/day, Awake=24-TotalSleep)
NonDreaming	slow wave ("nondreaming") sleep (hrs/day)
Dreaming	paradoxical ("dreaming") sleep (hrs/day) detected by phase of REM (Rapid Eye Movement)

Project Objectives:

Using this dataset, you should be able to build a model to predict the sleeping attributes **TotalSleep and Dreaming** from the general, ecological and biological attributes. Some

attributes are redundant, like TotalSleep and Awake, so you should determine which variables should be include before building the model. Then, you should determine which attribute is most important to understand **TotalSleep** and **Dreaming** time. Also, you should assess what is the correlation between diet groups, endanger status or Genus to the sleeping attributes. Apart from the sleeping attributes, you can also study the correlations and regressions within biological and ecological attributes.

The project can be submitted as a Jupyter Notebook (at least) and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation and finally, deployment.
You may use additional resources from those that are suggested in the "Project Resources" section or others as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. **However, this comes with a big responsibility: any submission that is over-plagiarised or does not reflect personal work will not be accepted**.

Project Resources:

Here are additional resources that may be helpful for the project. The dataset has been built from a merge of two studies on sleep duration in mammals:

- Allison T, Cicchetti DV. Sleep in mammals: ecological and constitutional correlates. **Science. 1976** Nov 12;194(4266):732-4. doi: 10.1126/science.982039. PMID: 982039.
- Savage VM, West GB. A quantitative, theoretical framework for understanding mammalian sleep. **Proc Natl Acad Sci U S A. 2007** Jan 16;104(3):1051-6. doi: 10.1073/pnas.0610080104. Epub 2007 Jan 10. PMID: 17215372; PMCID: PMC1783362.

Project Evaluation:

The project will be evaluated using the following rubric. It contains the required items for a complete submission. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes) and feature selection (feature engineering, feature pruning, choice justification) **[1 point]**
- Model training (motivation for selected model, comparison of different models) and evaluation (evaluation metric, results interpretation) **[1 point]**
- Project report (short report explaining the approach and results) **[1 point]**
- Project reproducibility (requirements file with necessary packages, README file for running the project) **[1 point]**
- Project hosting and deployment (Github, Docker, AWS, Heroku or any other method) **[1 point]**