Saliou CISSE
Quentin CAMILLERI
Tom SALVADORE
Ethan ADA

Project report

Project 3.

First of all, for this project, we first decided to import the numpy, pandas, matplotlib, seaborn and math library to manipulate the dateframe, for visualization and computation if needed.

We can split this project into 3 phases.

The first phase consists of the completion of the dataset using information from the internet. Indeed, we need a clean dataset before doing any kind of machine learning work on it. We chose that option (manual completion) knowing that we didn't have a lot of rows.

We completed the rows to the best of our capabilities manually, but it wasn't enough, there were still missing values but only for the dreaming, predation, exposure and danger. We then chose to complete these values using the mean values of the already existing other species which have the same genus, family or order (in this order of preference). We added the family column with values found on the internet particularly to use this technique.

We also compared the different features to the Totalsleep which is a target variable and based on these comparison we decided to get rid of the NonDreaming column because it gave the same informations that the Dreaming column, the Genus and Family columns because they didn't give much informations by having too much groups which can complexify the model and can cause overfitting. We also dropped the BrainWt column and the Awake column because they were redundant with the BodyWt and TotalSleep columns respectively.

We then turn the Order, Vore and Conservation columns into Booleans with 0 and 1 values, by turning each group into a new column.

The second phase was modelling phase. We decided to try the linear regression despite its simplicity, suspecting linear relationships between features and our target variable. In addition, we tried the Random Forest that can capture non-linear relationships on top of the linear regression and a Neural Network to see if a more complex model would have better results, knowing that a neural network is capable of feature learning. We choose those 3 models knowing that our constraints were that we have a small dataset and a good number of features. We used these models for the prediction of the TotalSleep of the species and for the Dreaming.

It brings us to the third and last phase which consists of result interpretation.

For the Sleeping model, the best that worked for us was the Random Forest, with an average of error of 2.2 hours. The Linear Regression was close behind, and the Neural Network was not functioning at all.

The error represents about 20% of the mean value of the column. You could use this model, but you must keep in mind that the values are not really accurate and have a 20% error.

Saliou CISSE
Quentin CAMILLERI
Tom SALVADORE
Ethan ADA

For the Dreaming model, the best that worked for us was the Random Forest and the Linear Regression, both with an average of error of 1.1 hours. The Neural Network was not functionning at all.

The error represents about 50% of the mean value of the column which is really a lot. We do not recommend using this model.

In retrospective, it makes sense, other than the fact that we had not much data and 1/3 of the Dreaming column was filled with artificial data, the dreaming state should have more link with the development of the brain, the activity of the brain, more brains related data than the type of food they eat or the state of conservation of their species.

These results are not surprising, knowing that the Neural Network works best with large dataset and a linear regression can easily overfit small datasets in addition to the fact that it assumes a linear relationship between predictors and the response