Course: Python Machine Learning Labs

Project: Predicting sleep variables in mammals

Instructor: Christophe Bécavin

Students: Zakaria BENGARAA, Salim, TOUATI, Mohamed Aymen KHERARBA and Kevin COLLIN

Table of Contents:

Introduction

The objective of the project is to use machine learning algorithms to predict two features of mammals: their sleeping time and their dreaming time. To do so, we will use a data set composed of 16 attributes (columns), ranked in four categories: general, biological, ecological and sleep attributes. The data set is relatively small, which is common in the biology world.

The first part of the report will present our processes and thoughts to predict the total sleep. It will deal with the exploratory data analysis, the methods description and results interpretation. With the same structure, the second part will discuss about the prediction of the dreaming time. And finally, the third part will explore the links between the other attributes.

Preliminary thoughts

The size of the data set is small (87 rows), so we thought about including more mammals in this data set, in order to increase the rows number. Sadly, our instructor confirms us it is not possible, due to

lack of science-based data. Consequently, our result could be enhanced in the future when new data will be produced by the scientific community.

1. Predicting sleeping time

1.1) EDA

A lot of values are missing in this data set, up to 40 for NonDreaming attribute. As a consequence, we will focus on filling the missing data first (see model 1 and 2), with two different technics. We will also test by removing the rows with missing values (model 3).

Underneath, we explain the deleted attributes from the data set and the reasons:

- TotalSleep and Awake: they are redundant.
- Conservation: this attribute is not linked with any biological characteristics of a mammal. It expresses the conservation status and extinction risk of biological species.
- BodyWt and BrainWt: they have a high degree of correlation (93 %, see Figure 1 below).
- Dreaming and NonDreaming attributes: totalsleep = Nondreaming + Dreaming
- Species: They are all unique values.

	BodyWt	BrainWt	TotalSleep	LifeSpan	Gestation	Predation	Exposure	Danger
BodyWt	1.000000	0.925683	-0.310147	0.302382	0.696004	0.070922	0.370613	0.150341
BrainWt	0.925683	1.000000	-0.319661	0.506326	0.776817	0.027343	0.383869	0.143061
TotalSleep	-0.310147	-0.319661	1.000000	-0.417433	-0.660791	-0.408713	-0.677876	-0.587729
LifeSpan	0.302382	0.506326	-0.417433	1.000000	0.643651	-0.116818	0.372426	0.049262
Gestation	0.696004	0.776817	-0.660791	0.643651	1.000000	0.169895	0.659636	0.356378
Predation	0.070922	0.027343	-0.408713	-0.116818	0.169895	1.000000	0.619839	0.930782
Exposure	0.370613	0.383869	-0.677876	0.372426	0.659636	0.619839	1.000000	0.770361
Danger	0.150341	0.143061	-0.587729	0.049262	0.356378	0.930782	0.770361	1.000000

Figure 1: Correlation Table between attributes.

Two attributes deserve a deeper analysis to determine if they are correlated to the TotalSleep attribute: LifeSpan and BodyWt. Indeed, they demonstrate both a modest correlation, respectively 42% and 31%.

In addition, the Figure 2 below does not show a strong linearity between LifeSpan and BodyWt attributes. It is even more explicit for LifeSpan and TotalSleep, according to the Figure 3. To be sure our intuition is right, we will train each of our three models, with and without these two attributes. We will make sure to fix the train set to compare the evaluation. If our intuition is right, the models without LifeSpan and BodyWt attributes will better perform.

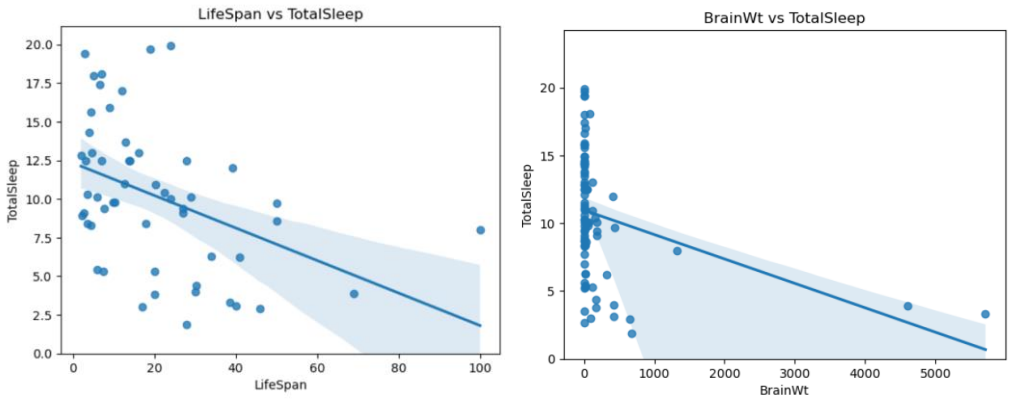

Figure 2 (left). Graph of TotalSleep vs LifeSpan & Graph of TotalSleep vs of BrainWt.

Furthermore, we are exploring few complementary interpretations from the EDA:

- The insectivore mammal is poorly represented in our data (9.20%, see annexed figure 5), meaning that our models will not be able to accurately predict the sleeping time of an insectivore mammal.
- A mammal which is more likely preyed upon will tend to sleep less (see annexed figure 6).
- A mammal which sleeps in a well-protected area will tend to sleep for a longer time (see annexed figure 6).
- A mammal which is in danger will tend to sleep for a shorter time (see annexed figure 7).
- A mammal which has a long time of gestation will tend to sleep less (see annexed figure 7).

1.2) Models' description

Model 1: filling missing values with Gene attribute.

In this first model, we decided to fill the missing of the data set based on the gene attribute, which represent a "layer" of the taxonomic rank. In fact, taxonomic rank is the relative level of a group of organisms (a taxon) in an ancestral or hereditary hierarchy. As an example, the figure 4 below represent the taxonomic rank for one of the mammals of our data set, the African elephant. It is important to note that the provided data set only gives us the "layers" of the taxonomic rank that are circled in red: Order, Genus and Species.

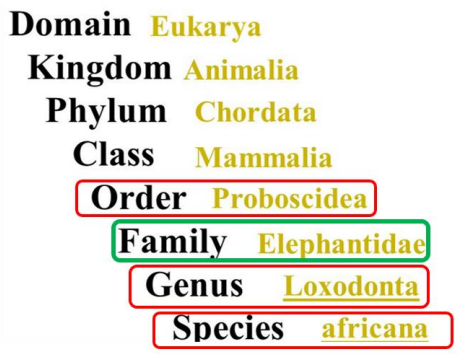

Figure 3. Taxonomic rank of the African elephant

The process to fill the missing values for each specie in each row is the following:

1. Detecting where are the missing values for a given Order: which species for which attributes.
2. Calculating the mean of each attribute, in the given Order.
3. Filling the missing values by the calculated mean.

Example from the dataset:

1. In the Order called Primates, three species have missing values: Squirrel monkey, Potto and Mongoose lemur. These missing values are located in the following attributes: Lifespan, Gestation, Predation, Exposure, Danger.
2. The program calculates the mean for the attributes above.
3. The program fills the missing values.

This method allows to keep 83 rows over 87.

Model 2: filling missing values with Family attribute.

In this model we decided to include a column in the data set, which represent the Family layer from the taxonomy rank, in green in Figure 4 above. We picked up these data from the National Library of Medicine. Therefore, there more groups to rank the species than the Order.

The same process than the model 1 above is applied to fill the values.

This method allows to keep 75 rows over 87.

Model 3 with Family attribute

For this model, we delete from the data set all the rows with a missing value. This is not the preferred model, since we withdraw a lot of rows.

This method allows to keep 50 rows over 87.

1.3) Results interpretation

Considering the use of the machine learning algorithm, we thought the limited size of our dataset would force us to use simple machine learning algorithm. Our intuition was correct, because we obtain higher precision with a linear regression than the random forest (result in the notebook, in this report we show only the best obtained model), with the same split and train dataset.

The table below show the precision with the linear regression algorithm for each model.

Model		**1**	**2**	**3**
Specificity		Filling values based on Order attribute	Filling values based on Family attribute	Deleting rows with missing values
Precision score	With LifeSpan and BodyWt	0.59	0.42	0.61
	Without LifeSpan and BodyWt	0.59	0.64	0.79

We can see that the precision is egal or higher without LifeSpan and Bodyweight attributes. This corroborates our assumption that these two attributes should not be included.

In addition, we are surprised to conclude that the most efficient model is the third one, where we deleted all the rows with missing values. The precision score is 0.79, this model is save under the name "bestmodel.sav" in document.

To determine which attributes contribute the most and the least to predict TotalSleep, we analyse the weight of each variable from the TotalSleep multi linear equation. The table below summarises the analysis:

Weight of each variable (absolute value)

Variable	Model 1	Model 2	Model 3	Mean of the weight
Danger	3.50	3.50	1.00	2.67
Predation	1.80	2.00	0.28	1.36
Insecti	1.90	0.76	0.23	0.96
Carni	1.10	0.48	0.73	0.77
Omni	0.43	0.07	0.61	0.37
Herbi	0.36	0.35	0.34	0.35
Exposure	0.21	0.17	0.14	0.17
Gestation	0.02	0.02	0.02	0.02
LifeSpan	0.00	0.01	0.01	0.01
BodyWt	0.00	0.00	0.00	0.00

Based on the mean of the variable's weights from the models, we can demonstrate that:

- Danger and Predation represent the attributes that contribute the most to predict TotalSleep.
- LifeSpan and BodyWt are indeed the attributes that contribute the less to predict TotalSleep.

2. Predicting Dreaming time:

Using the same logic and analyses as in model 3 for predicting total sleep, we applied a different function and normalized the data frame before training. 1st data frame with BodyWt, BrainWt and LifeSpan and the 2nd data frame is without previous columns.

We cleaned the data and got 41 rows in the first data frame and 43 in the second. We also learned that data normalization is crucial for machine learning models like linear regression, but only when the features have different scales (for example: BodyWt [0.005000: 6654]).

As the previous analyses, we found that there's no strong linearity between BodyWt, BrainWt and LifeSpan with dreaming also, that why we tested two data frames to see the improvement of the model.

2.1. Results:

Model Specificity		**Random state**	**Deleting rows with missing values**
Precision score	With LifeSpan, BodyWt and BodyWt	103	0.62
	Without LifeSpan, BodyWt and BodyWt	117	0.66

We can also observe that the precision is the same or little better without LifeSpan and Bodyweight attributes.

The table below shows the analysis of the Dreaming multi linear equation, which helps us identify the most and the least influential attributes for predicting Dreaming.

Variable	With LifeSpan, BodyWt and BodyWt	Without LifeSpan, BodyWt and BodyWt
Danger	0.98	0.49
Predation	0.63	0.28
TotalSleep	0.11	0.32
Insecti	0.06	0.001
Carni	0.07	0.09
Omni	0.04	0.002
Herbi	0.04	0.02
Exposure	0.19	0.01
Gestation	0.69	0.01
LifeSpan	0.23	/
BodyWt	1.03	/
BrainWt	0.29	/

The normalization process has altered the previous observations. We can see that **BodyWt** has an impact on the Dreaming prediction as **Danger**, while **LifeSpan** has a minor effect.

From our analysis, we can conclude that the attributes which significantly influence the prediction of Dreaming are Danger, Predation, Gestation, and BodyWt.

3. To go further: the links between other attributes

Apart from the sleeping attributes, this part will examine the correlations and regressions within biological and ecological attributes.

From the biology point of view, a correlation is considered “strong” when it is above or equal to 0.3.

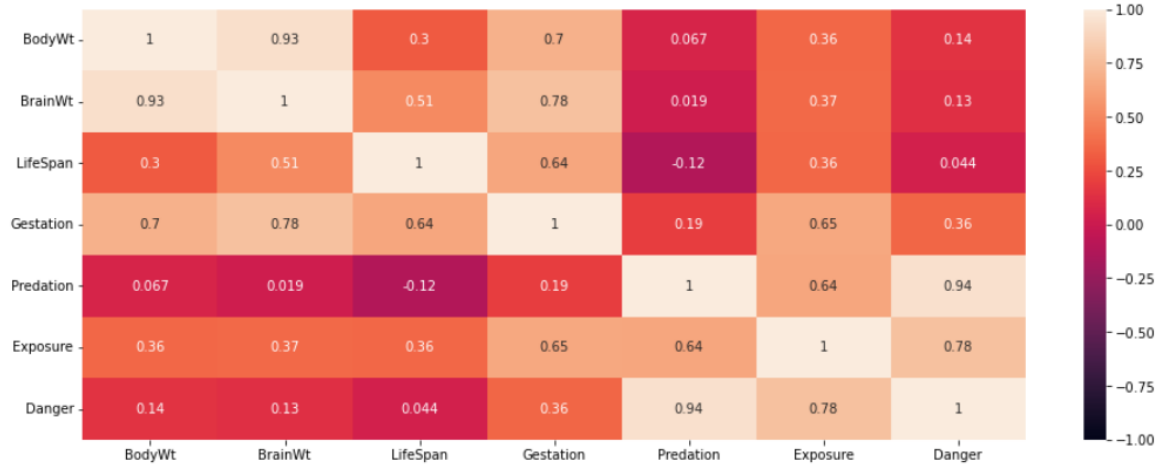

Figure 4. Heatmap of the attributes, apart from sleeping and dreaming attributes

For the sake of our analysis, we decided to drop all the null values.

Regarding biological and ecological attributes, several links and correlations seem interesting to emphasize.

Within the biological attributes:

To begin with, we can observe a significant relation between the two weight characteristics (Body and Brain) and the time of Gestation (0.7 and 0.78). We can assume here that bigger is the animal, longer could be the gestation.

Second, the heatmap shows us another relevant correlation between the time of gestation and the life span (0.64).

Then, a last correlation seems strong in biological attributes: the life span and the two weight characteristics (Body and Brain) with a correlation of 0.3 and 0.51. Bigger is the animal and its brain, longer it lives.

To conclude about the biological, we can observe that each of them provide a positive correlation.

Within the ecological attributes:

Concerning the ecological attributes, as the Predation and the Exposure are part of the Danger, it seemed not relevant to interpret this correlation with Danger.

However, the correlation between Predation and Exposure seems interesting and quite logical. With a positive correlation of 0.64, it indicates us the strong relation between the place where an animal sleep and the likelihood of being prey.

Between biological and ecological attributes:

When we try to establish a correlation between these two attributes, we observe a positive and strong correlation between biological attributes and Exposure. Indeed, all of them are higher than 0.35:

BodyWt / Exposure => 0.36: Higher is your body, bigger will be the exposition. Probably massive mammals encounter some difficulties to find or reach a safe place to sleep.

LifeSpan / Exposure => 0.36: This correlation goes against our beliefs because it exposes that the more dangerous the place is, the longer is the life span. This is maybe something to cross with another data like the diet groups.

Gestation / Exposure => 0.65: The correlation between these two parameters is strong. It exposes us that animals with long time of gestation are often the ones which sleep in the most unsafe places. It could be explained by the strong correlation between the body weight and the time of gestation. Bigger animals have a longer time of gestation, and bigger animals could have some difficulties to access a safe place to sleep.

Conclusion

Our project dealt with biological data that had two main challenges: low volume (87 lines) and high nullity. To overcome these, we had to modify our model and our training method. Based on our research and our tutor's guidance, we chose to discard the missing values and focus on the ones that were present.

Perspective:

Despite the challenges we faced due to the lack of data and difficulties in filling it, we achieved satisfactory results (**R=0.75**), which are acceptable given the dataset. If the scientific community were to create a larger and more robust dataset in the future, we could improve our predictions of TotalSleep or Dreaming.

Additionally, we could enhance the process of filling in missing values using libraries from **scikit-learn**, such as **sklearn.impute.SimpleImputer or sklearn.impute.KNNImputer**, to obtain a more robust dataset.

Annexes

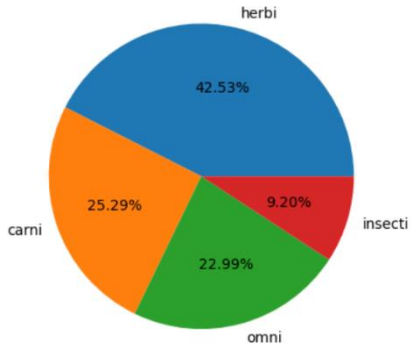

Figure 5: Vore attribute representation.

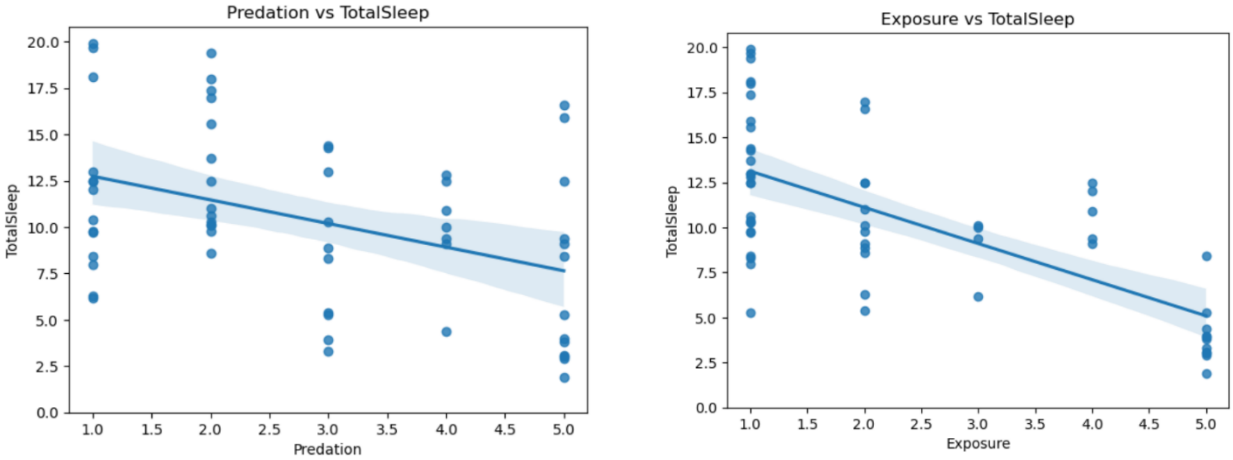

Figure 6. Graph of TotalSleep vs Predation (left) & Graph of TotalSleep vs of Exposure (right).

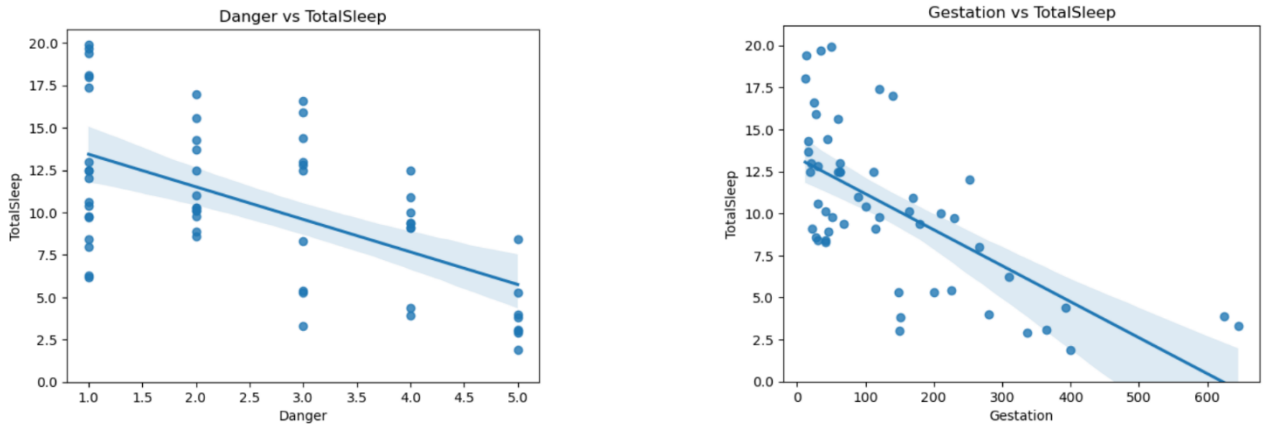

Figure 7. Graph of TotalSleep vs of Danger (left) &Graph of TotalSleep vs of Gestation (right).

Dreaming prediction:

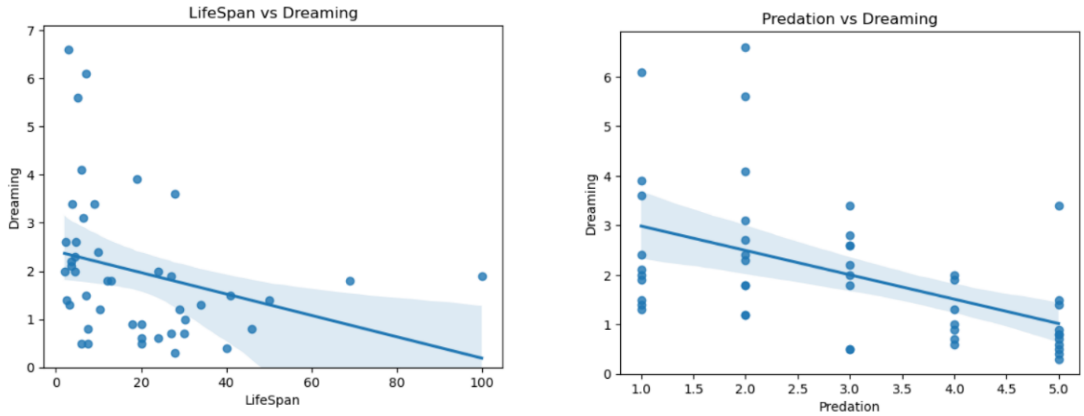

Figure 8 Graph of Dreaming vs LifeSpan (left) & Graph of Dreaming vs of predation (right).

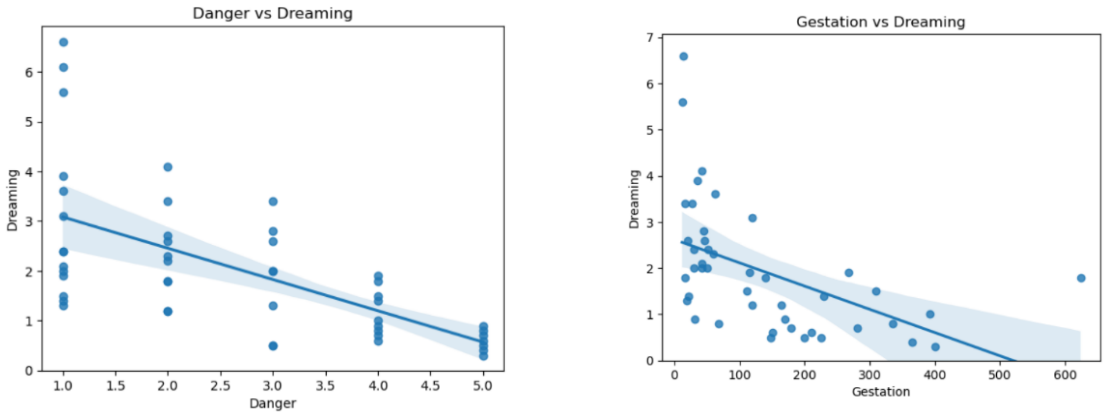

Figure 9 Graph of Dreaming vs Danger (left) & Graph of Dreaming vs of Gestation (right).

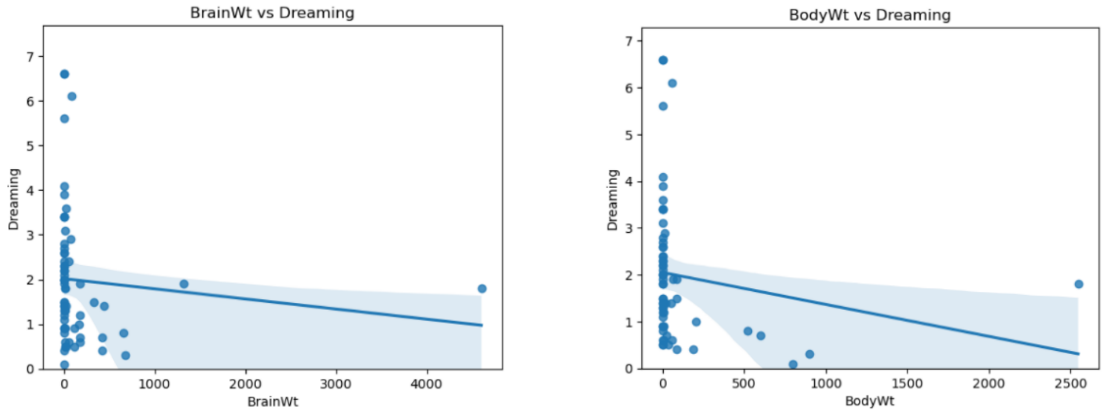

Figure 10 Graph of Dreaming vs BrainWt (left) & Graph of Dreaming vs of BodyWt (right).

Dataset With **LifeSpan, BodyWt and BodyWt**:

Random state: [40, 60, 62, 71, 81, 88, 91, 93, 103, 117, 133]

Mean of Coefficient of determination: 0.55

Random state	Coefficient of determination
40.00	0.5979078874051293
60.00	0.5051103323867021
62.00	0.5876815141332861
71.00	0.5718804298046869
81.00	0.5559618274700693
88.00	0.514693067839314
91.00	0.5012726809198227
93.00	0.5802822682801523
103.00	0.605991588633833
117.00	0.5579215638583419
133.00	0.5054069724487612

Dataset Without **LifeSpan, BodyWt and BodyWt**:

Random state: [4, 6, 13, 62, 68, 72, 91, 103, 112, 116, 117, 131, 134, 136, 143, 148]

Mean of Coefficient of determination: 0.57

Random state	Coefficient of determination
4.00	0.6025292760604604
6.00	0.5046460888558176
13.00	0.5545367713174107
62.00	0.5417500452216832
68.00	0.5697199883112252
72.00	0.5351001416984507
91.00	0.659929651790213
103.00	0.5538865845025717
112.00	0.5863334285616326
116.00	0.5780052374547529
117.00	0.6554730675570641
131.00	0.6082925894006119
134.00	0.5085882525113883
136.00	0.5818053374863965
143.00	0.5070451886377223
148.00	0.5119710871054745