

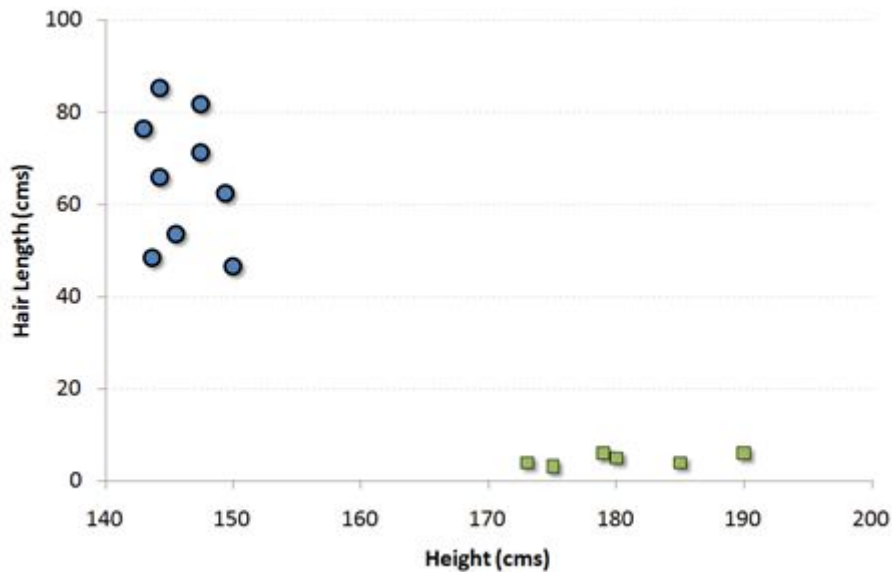
## Animal Attribute Prediction

### Justification

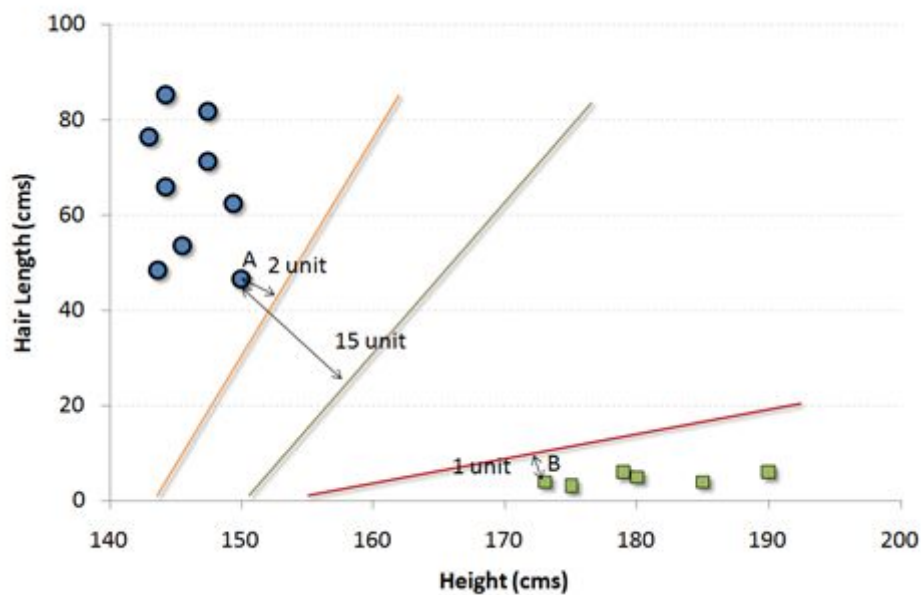
More than 18,000 species were discovered in 2017, and while not all of them may be animals, this is just a little scratch of all the millions of species more that are still undiscovered, but some of the discovered ones are very hard to investigate, document or even classify. Some of these new species where already found, this is because there are cases where in museum collections, nobody looked at the specimens closely enough to identify if it was a new species found, also technology has led to identify even more animals by analyzing their DNA even if the species looks exactly alike, they can have lots of dissimilarities in their genes. Since science has identified about 2 million of species of animals, plants and microbes, and researchers are looking in every corner of the world for new ones, it might be hard for scientists to find some missing attributes for a given animal, with whatever characteristics, attributes and facts that were documented.

We consider this an opportunity to develop something that can help to classify, in the name of science. For this, we used an AI model called SVM (Support Vector Machine). A Support Vector Machine is a supervised machine learning algorithm that analyzes data and is often used for **classification** or regression challenges. It transforms your data and based on that, it finds the optimal frontier/boundary between outputs, this is done by separating the input data based on the labels or outputs defined.

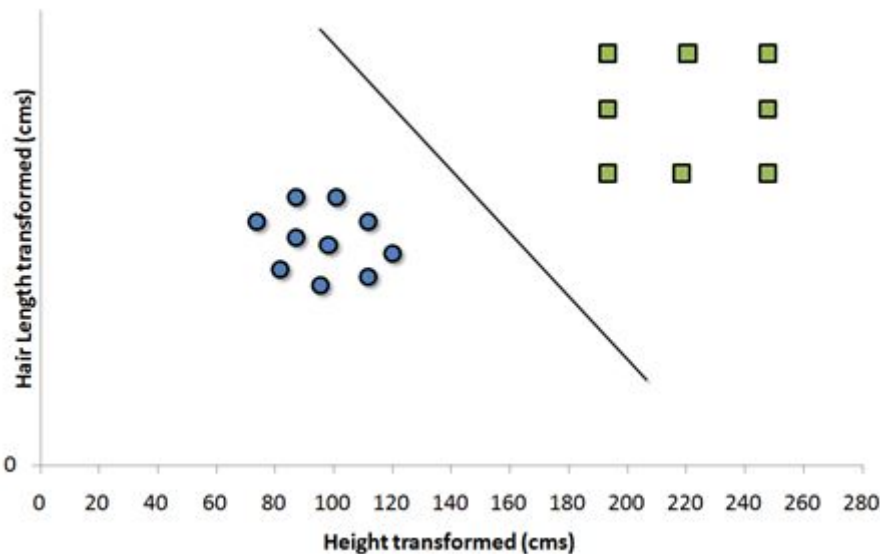
But how do they work? Let's imagine that we have a population composed of 50% Males and 50% Females and we want a computer to identify whether someone is a Male or a Female. In these example, there are only two segments, males have higher average height and females have longer hair.



A Support Vector is a coordinate of individual observation like the coordinate (45, 50), that corresponds to a female in this case, also its easier to identify the classes because they are notably separated. To find the SVM you need to decide which is the best frontier for a particular statement, for this you need to find the minimum distance between a frontier and a support vector, like (45, 50) or any other from other class.



In these cases, we have three frontiers, and we calculate the minimum distance from the nearest Support Vector coordinate and then we select the farthest distance. If there's not a frontier that helps us segregate between classes, then we need to map the vectors to a higher dimension so that we can segregate each class.



There are also Non-linear SVMs that calculate the frontiers without a straight line, giving the possibility to capture more complex relations between Support Vector coordinates and without having to perform more difficult transformations.

Based on these, we used Support Vector Machines for a linear classifying, to predict each animal attributes, given a dataset full of attributes like Color, Strength, Speed, Type, etc. We used several methods that use Support Vector Machines, these were:

- Support Vector Classification (SVC)
- Linear SVC – Is more flexible in the choice of penalties and loss functions (it's better for large data samples).
- NuSVC – Uses a parameter named “Nu” to control the number of support vectors.

We also were able to implement a Neural Network algorithm model called Multi-layer Perceptron Classifier that maps sets of input data to a set of appropriate outputs. This consists of multiple layers, with each layer fully connected to the following one.

In our solution we obtained a dataset of 50 animals, and 85 attributes (marked with 1 if the animal has it or 0 if it doesn't). Based on these we calculate and predict the full attributes of a new animal by selecting which attribute you want to look for, by filtering them from the input data set received, and then find out which attributes have the most importance to fully describe an animal. Even though a small dataset size is a general condition for overfitting to happen, this is only seen when the accuracy on the test set is diminished. On the examples execution, the resulting probabilities were high (from 80% to 100%) in the majority of the cases, so we could say that, opposite to what we said in the presentation, overfitting did not happen in this challenge.

## Documentation

In order to run this project, navigate to the directory where it is located. Make sure you have python3 installed.

For installing python3 in ubuntu, it is needed to run the following commands in terminal.

```
$ sudo apt-get update
```

```
$ sudo apt-get install python3.6
```

```
$ sudo apt-get install python3-pip
```

After installing python3 and pip3 you then need to install Scikit-learn, Numpy and PrettyTable.

```
$ pip3 install numpy
```

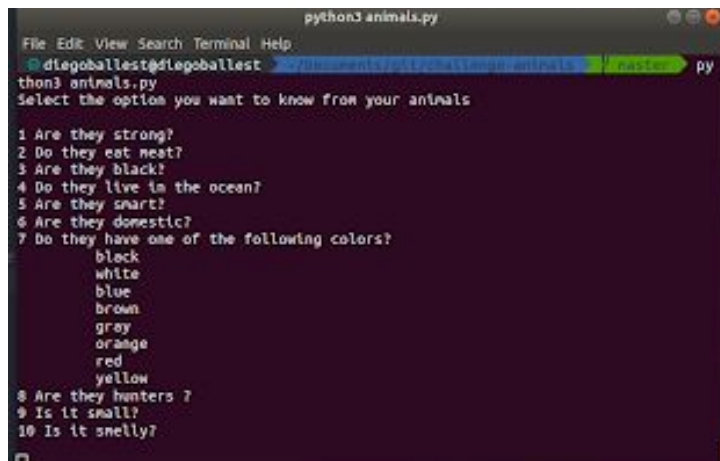
```
$ pip3 install -U scikit-learn
```

```
$ pip3 install prettytable
```

Navigate to the directory where it is located and execute

```
$ python3 animals.py
```

This is going to execute our animals attribute prediction implementation



```
python3 animals.py
File Edit View Search Terminal Help
diegoballest@diegoballest:~/Documents/Programacion/Animals$ python3 animals.py
thon3 animals.py
Select the option you want to know from your animals

1 Are they strong?
2 Do they eat meat?
3 Are they black?
4 Do they live in the ocean?
5 Are they smart?
6 Are they domestic?
7 Do they have one of the following colors?
    black
    white
    blue
    brown
    gray
    orange
    red
    yellow
8 Are they hunters ?
9 Is it small?
10 Is it snelly?
```

Select an option from the menu, these options are the attributes where you are going to predict from. You are going to get the prediction of each one of the Support Vector Machines methods and the MLP Classifier Neural Network, and you will be able to see the different results and how they differ from the expected output as seen on the image.

```
Group 1
Expected result:
[1, 1, 1, 0, 1, 1, 1, 0, 0, 1]
-----
| Animal | Predicate |
-----
| chimpanzee | YES |
| giant+panda | YES |
| leopard | YES |
| persian+cat | NO |
| pig | YES |
| hippopotamus | YES |
| humpback+whale | YES |
| raccoon | NO |
| rat | NO |
| seal | YES |
-----
Actual results:
Linear SVC: [1 1 0 1 1 1 0 0 1] 100.0 %
SVC: [1 1 1 1 1 1 1 0 1] 80.0 %
NuSVC: [1 1 0 1 1 1 0 0 1] 100.0 %
Neural network (lbrgs): [1 1 0 1 1 1 0 0 1] 100.0 %
-----
| Animal | LinearSVC | SVC | NuSVC | Neural |
-----
| chimpanzee | YES | YES | YES | YES |
| giant+panda | YES | YES | YES | YES |
| leopard | YES | YES | YES | YES |
| persian+cat | NO | YES | NO | NO |
| pig | YES | YES | YES | YES |
| hippopotamus | YES | YES | YES | YES |
| humpback+whale | YES | YES | YES | YES |
| raccoon | NO | YES | NO | NO |
| rat | NO | NO | NO | NO |
| seal | YES | YES | YES | YES |
| PERCENTAGE | 100.0 | 80.0 | 100.0 | 100.0 |
-----
diegoballest@diegoballest
```

## References:

Object Recognition with Hidden Attributes (2018). Retrieved from <https://www.ijcai.org/Proceedings/16/Papers/494.pdf>

Simplified, S., Simplified, S., & Srivastava, T. (2018). Classification Algorithm Support Vector Machine. Retrieved from <https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/>

Top 10 New Species Discovered 2017. (2018). Retrieved from <http://www.iflscience.com/plants-and-animals/top-10-new-species-discovered-2017/>

Why Thousands of New Animal Species Are Still Discovered Each Year. (2018). Retrieved from <https://www.atlasobscura.com/articles/new-animal-species>

Why use SVM?. (2018). Retrieved from <https://community.alteryx.com/t5/Data-Science-Blog/Why-use-SVM/ba-p/138440>