Machine Learning Nano Degree
Capstone Proposal
Predicting Zoo's animals class

Khloud Ali

July 4, 2018

Domain Background

Taxonomy is a branch of science that includes the identification, nomenclature, description and classification of the organisms.

classification is the science of how living organisms are grouped together.

It started with Aristotle who developed the first classification system, which divided all known organisms into two groups then dividing each of these main groups into three smaller subgroups:

1-Animals

Subgroups: Land, Water, Air.

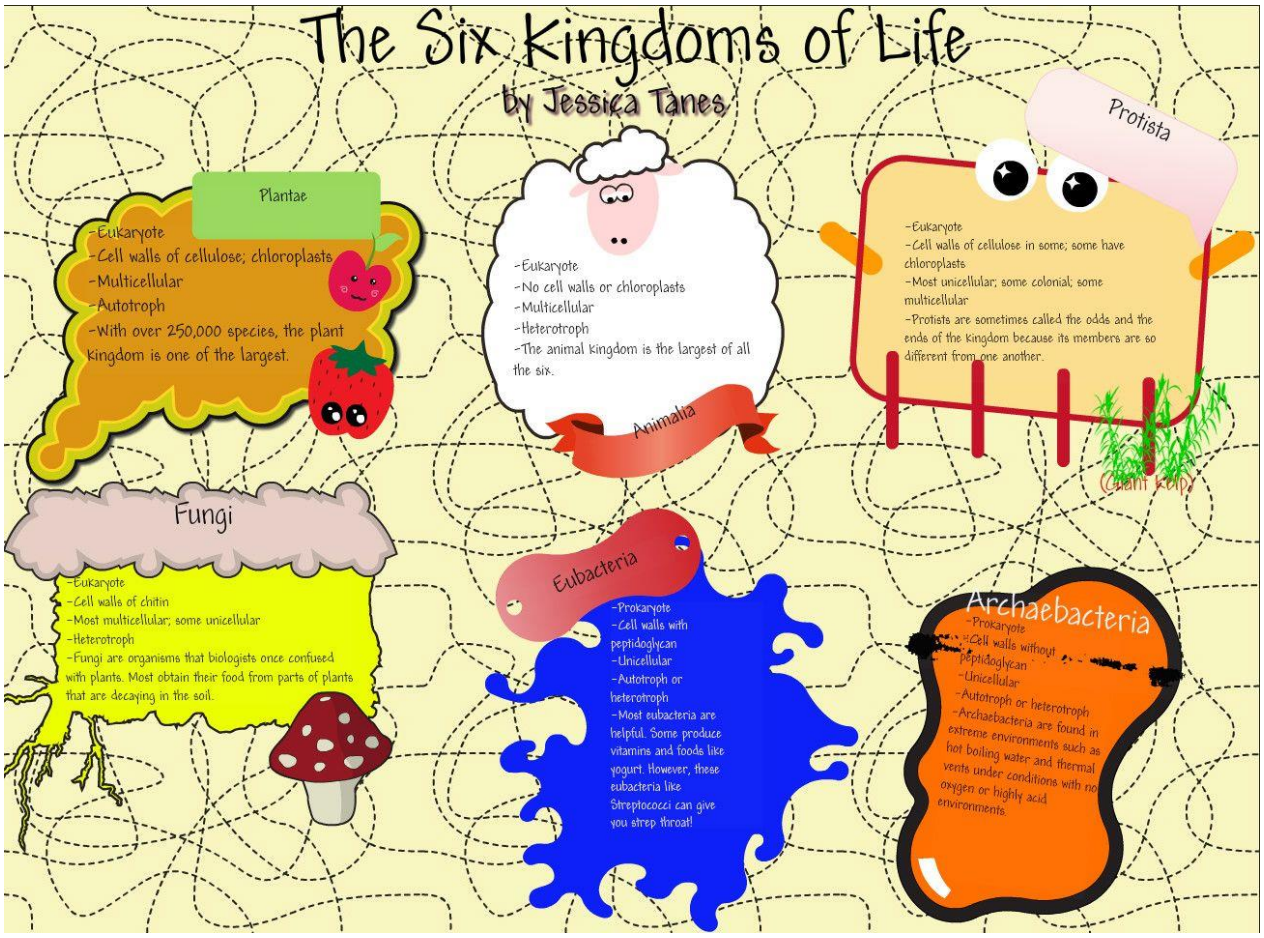
2-Plants

Subgroups: Small, Medium, Large.

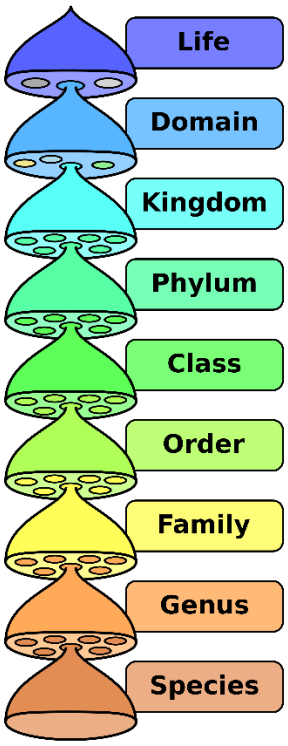
Then came Linnaeus, who classified the living organisms according to its traits and called this groups "Kingdoms", and also Linnaeus has divided each kingdom into five levels: class, order, genus, species, and variety. Each living organism was placed in those levels based on its traits, including similarities of physical body parts such as size, shape, and the way they feed.

But with the use of microscope leading to the discovery of new organisms the today classification system has developed to include 5 kingdoms which are:

1-Plantae.
2-Animalia.
3-Fungi.
4-Protista.
5-Monera (includes Eubacteria and Archaebacteria).



But in this domain we're going to focus on the Animals kingdom "word animal derived from a latin word called animalis meaning …Having Breath…" where the animalia "kingdom" is divided into 40 smaller groups each is known as a "Phylum" which is divided into further groups each is known as a "Class" and animals in a class have so much more in common more than the whole entire phylum, then class is divided into further more groups each is known as an "order" and in each order there's different types of families with similar features that's further more divided into "genus" where each genus contain the animals with very similar features that's closely related.



Levels (from the highest to the lowest)	Example
Kingdom	Kingdom Animalia is the broadest category of all in the animal classification system. It includes every animal.
Phylum (plural: Phyla)	Phylum Chordata includes all animals of the Kingdom Animalia that have spinal cords.
Class	Class Mammalia includes all warm-blooded animals of the Phylum Chordata that have hair and feed their young with milk.
Order	Order Artiodactyla includes all animals of the Class Mammalia that have an even number of toes in their hooves.
Family	Family Giraffidae includes all animals of the Order Artiodactyla that have long legs, a long narrow head with small horns, thin lips, and long tongues.
Genus (plural: Genera)	Genus Okapia and Genus Giraffa
Species	Species Camelopardalis, also known as giraffes in English.

focusing on the famous seven classes we're going to classify each zoo animal in our dataset according to the classes they belong to which are :

1-Mammels

If the animal's body is covered with hair and it drinks milk when it's a baby then it belongs to the mammal's class.

2-Birds

When the animal has feather and lays eggs as it's born out of an egg then it belongs to the bird's class.

3-Fish

If the animal is vertebrate that lives in the water and have fins, scales and gills on their body then it belongs to the fishes' class.

4-Reptiles

If the animal has a scaly skin, born in land with tail and are cold blooded then it belongs to the reptile's class.

5-Amphibians

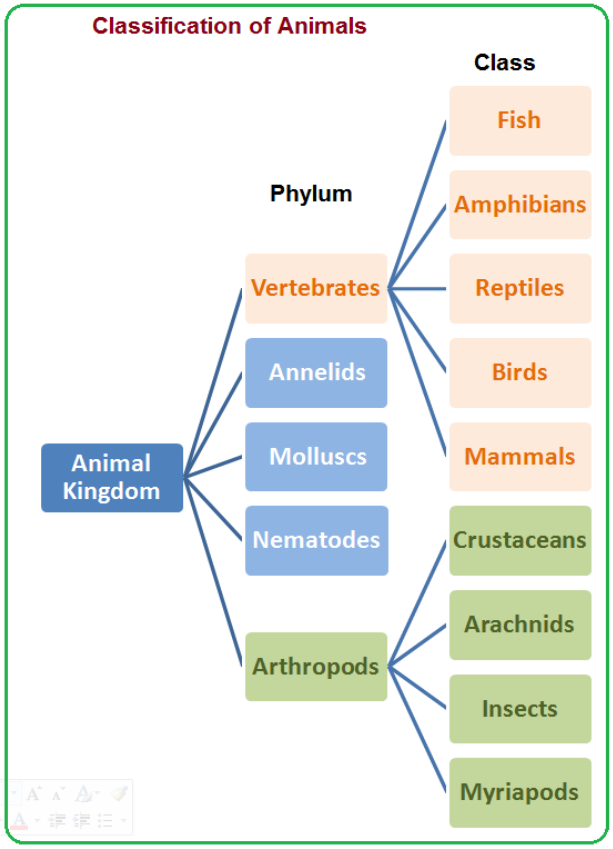
If the animal is born in water and it breaths through gills like fishes, but they develop lungs as they grow and can live on land then it belongs to the amphibian's class.

6-Arthropods "Bugs"

If the animal has more than four jointed legs and it doesn't have a backbone then it belongs to the arthropod's class.

7- Invertebrates

If an animal doesn't have a backbone or vertebra then it belongs to the invertebrates' class.



Problem Statement

It's a multi-class classification problem, where the problem is finding to which animal kingdom class each zoo animal belongs.

The goal is to predict which zoo animal is from a certain class from the provided classes, seven target class types are provided in this dataset:
Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate

Machine learning algorithm can learn from given data and make classification with a good accuracy giving a trusted output.

Datasets and Inputs

I used a dataset that features 7 different classes of animals and a 100 different zoo animal type with 16 features each is a Boolean provided by Kaggle…
https://www.kaggle.com/uciml/zoo-animal-classification#zoo.csv
The distinguishing features included in the dataset:

No.	Feature	Description
1.	Hair	Is its body is covered with hair?
2.	Feather	Is its body is covered with feathers?
3.	Eggs	Does it lay eggs or born out of an egg?
4.	Milk	Does it drink milk when it's a baby?
5.	Airborne	Does it fly?
6.	Aquatic	Does it live in water?
7.	Predator	Does it feed on row meat?
8.	Toothed	Does it have teeth?
9.	Backbone	Is it vertebra or invertebrate?
10.	Breathes	Does it have a nose or breath on land?
11.	Venomous	Is it poisonous?
12.	Fins	Does it have fins?
13.	Legs	Does it have legs and how many legs it has?
14.	Tail	Does it have a tail?
15.	Domestic	Is it housebroken?
16.	CatSize	Is it in a cat size or not?
17.	Class type	To which animal's class does it belong

Solution Statement

I am going to make a prediction using different models that either linear model, non-linear models or even ensemble models

then i am going to feed the model with the animal's name and its features making it predict to which class this given animal belongs to and the model with best score would be the final solution.

Benchmark Model

I will use kaggle F1 scores found on the kernels (since there was no completion on that data so there's no leaderboard), The highest F1 score I have seen till now on kaggle for this problem is 95%.

Evaluation Metrics

The dataset is imbalanced so Accuracy is no longer an option, I will use instead F1 score to use both precision and recall where:

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Also, I will use another metric for fine-tuning the final model like Roc-Auc or log-loss.

Project Design

Programming language: python 3.7
Libraries used: Pandas, sklearn, seaborn, matplotlib, numpy.

I am going to start by

- uploading the data and exploring it by presenting the data
- Exploring the data much further by showing its shape and how the classes are distributed and
- Through data visualization, we could gain more insight from the data and figure out how We can deal with it
- Train the models on training data and using cross-validation data to test various parameters tuning to choose the best
- Test the models using test data then choosing the best model with the highest accuracy then applying the best model on the data to get the best score.
- Finally Evaluating data through metrics and visuals.

References

https://a-z-animals.com/reference/animal-classification/
http://www.kidzone.ws/animals/animal_classes.htm
https://museumsvictoria.com.au/bugs/aboutbugs/types.aspx
https://biology.tutorvista.com/organism/kingdom-animalia.html
http://scienceprojectideasforkids.com/2010/history-of-classification/
http://www.softschools.com/science/biology/classification_of_living_things/
http://www.softschools.com/science/biology/classification_of_living_things/
http://www.softschools.com/science/biology/the_five_kingdoms/
https://www.mrsd.org/cms/lib/NH01912397/Centricity/Domain/245/animal%20classification%2
https://github.com/udacity/machine-learning/blob/master/projects/capstone/capstone_proposal_template.md

https://github.com/davidrobles/mlnd-capstone-proposal/blob/master/proposal.pdf
https://github.com/Tahsin-Mayeesha/udacity-mlnd-deeplearning-capstone/blob/master/capstone%20proposal.md
https://github.com/mvirgo/MLND-Capstone-Proposal
https://github.com/sgreenberg/seedling-classification/blob/master/proposal.pdf
https://github.com/sgreenberg/seedling-classification/blob/master/proposal.pdf
https://en.wikipedia.org/wiki/Taxonomy