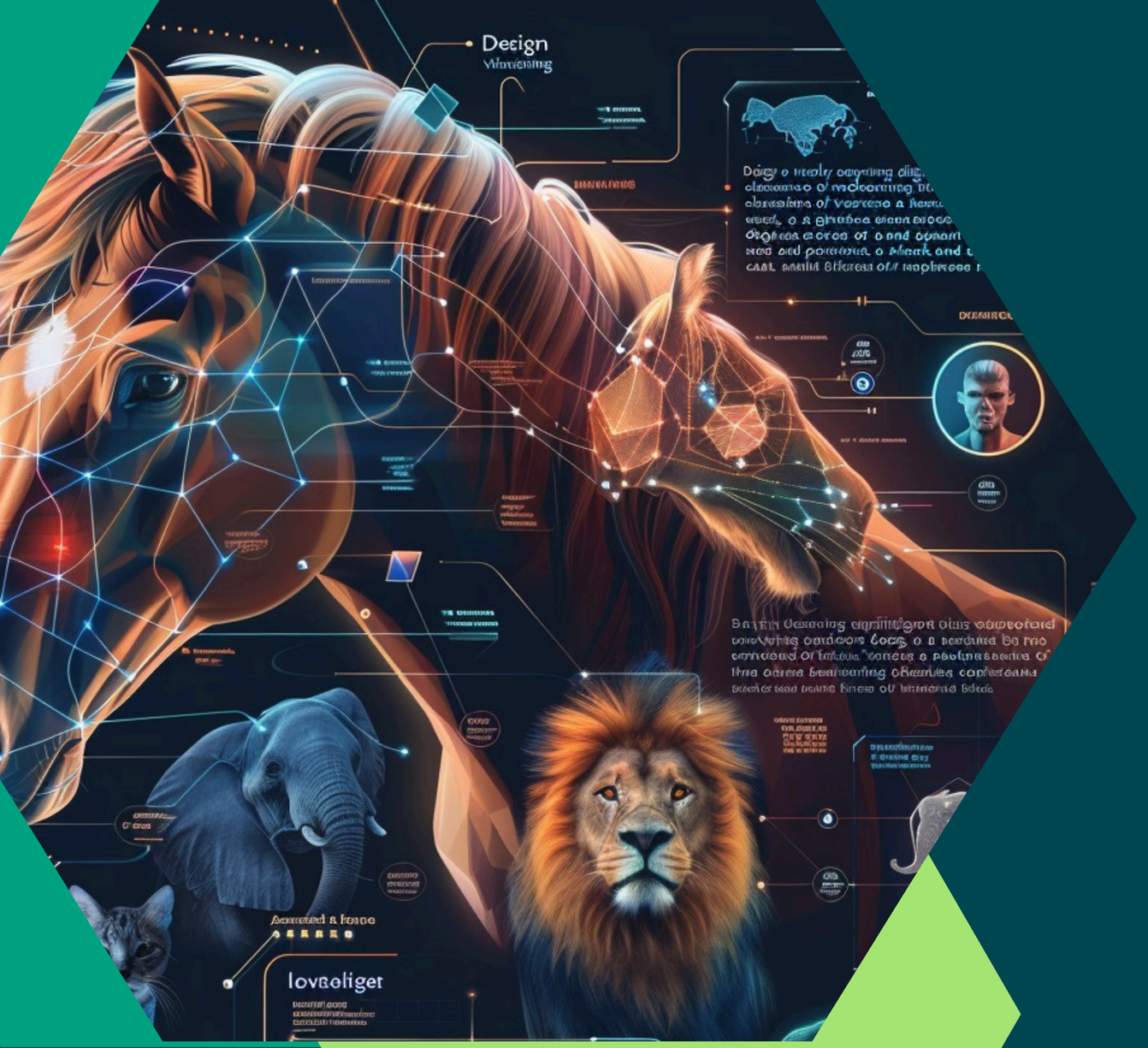
Machine Learning Course

Animal Classification Project

Yair Turgeman
Elad Laster



introduction

- Overview
- Dataset
- Data Preprocessing
- Mechine Learning Algorithms
- Review code & result
- Our Main Questions
- Conclusion

Machine Learning

Overview

The main objective of this project is to perform image classification on a dataset consisting of images of five different mammal.

The ultimate goal is to develop machine learning models that can accurately determine the specific mammal species depicted in a given image.

Machine Learning

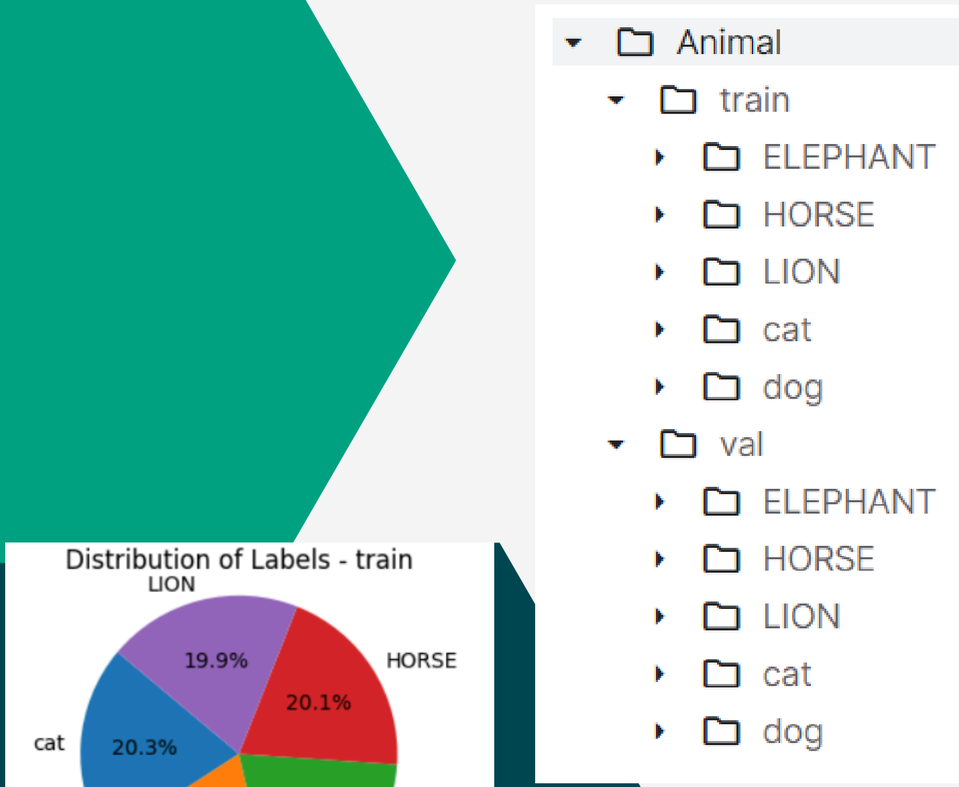
Dataset

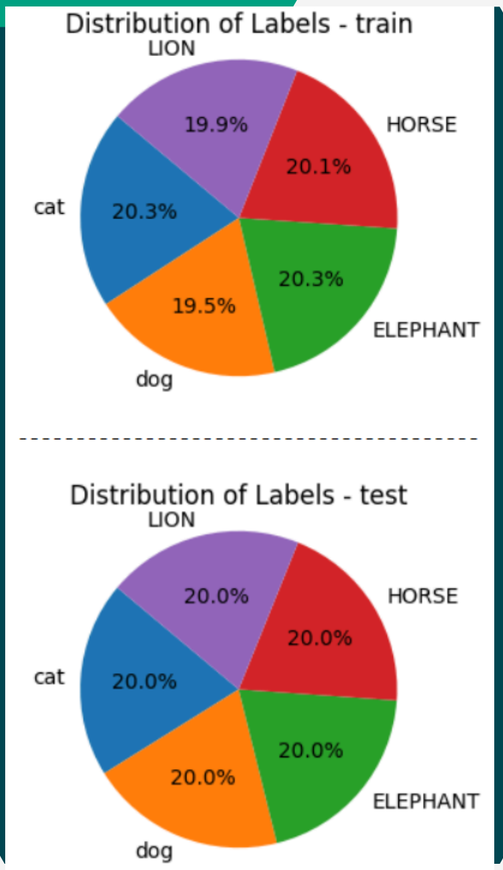
The dataset comprises 15,000 medium quality animal images, divided into two main folders: train and test.

Within each folder, there are subfolders for each mammal category.

All the images have been collected from "google images" and have been checked by human.

dataset source : https://www.kaggle.com/datasets/shiv28/animal-5-mammal



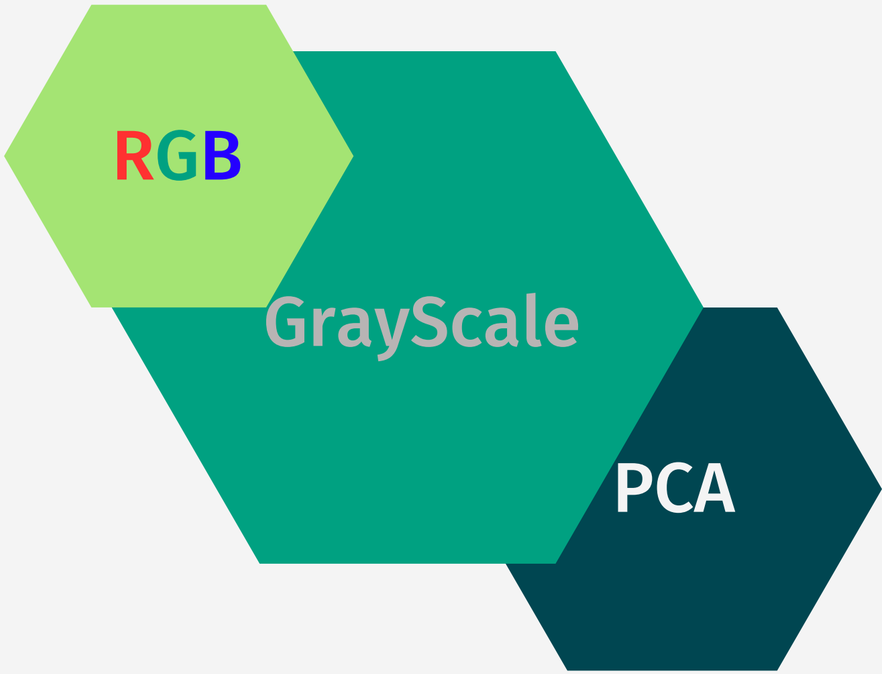


Machine Learning

Data Preprocessing

- Train-Test Split & Shuffle & Resize
- Show Few Pictures from the Data in RGB
- Show Data Distribution
- Classification on Images with the Models (on RGB)
- Moving the Images to Grayscale
- Show Few Pictures from the Data in Grayscale
- Classification Again on Images with the Models (On Grayscale)
- Applying PCA (Dimensionality Reduction)
- Classification Again on Images with the Models (On RGB using PCA)

Mechine Learning Algorithms



1. K-Nearest Neighbors
2. Logistic Regression
3. Decision Tree
4. AdaBoost

K-Nearest Neighbors

Assigns a class or value to a new data point based on the majority class or average of its nearest neighbors in feature space, using a predefined number of neighbors (k).

Logistic Regression

A statistical method used for binary classification tasks, which predicts the probability that an instance belongs to a certain class.

Decision Tree

organizes data into a tree-like structure where each branch represents a decision based on a feature, leading to a final prediction at the leaf nodes.

AdaBoost

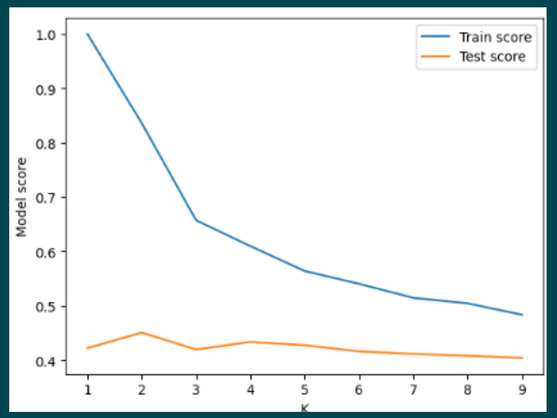
Adaboost combines multiple weak classifiers into a strong one by iteratively giving more weight to misclassified data, improving overall classification accuracy.

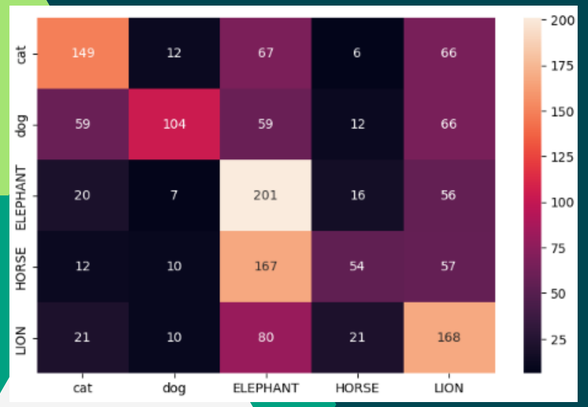
Review code & result

source code : https://github.com/yair489/ML-animal-classification.git

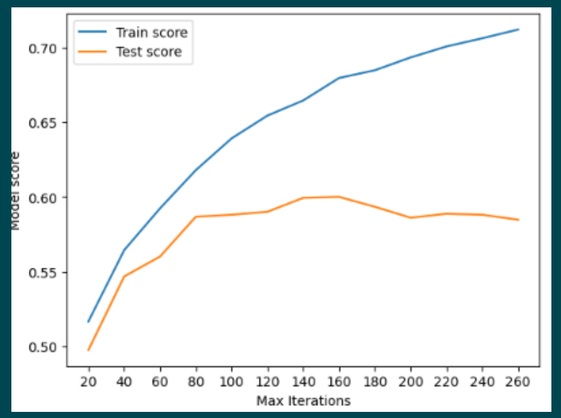
RGB

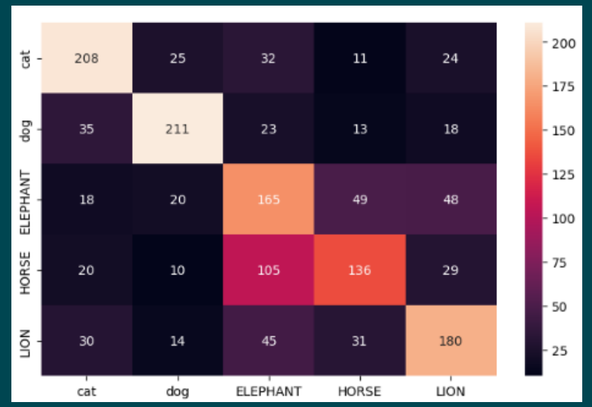
KNN



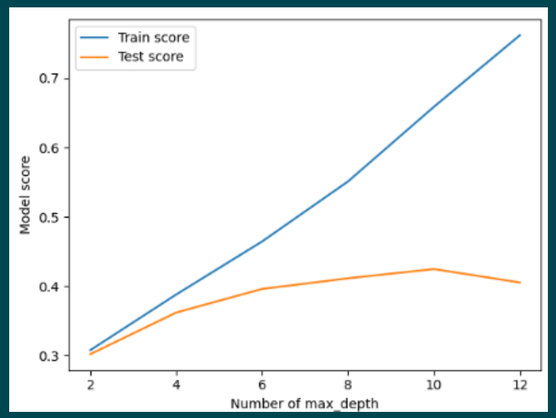


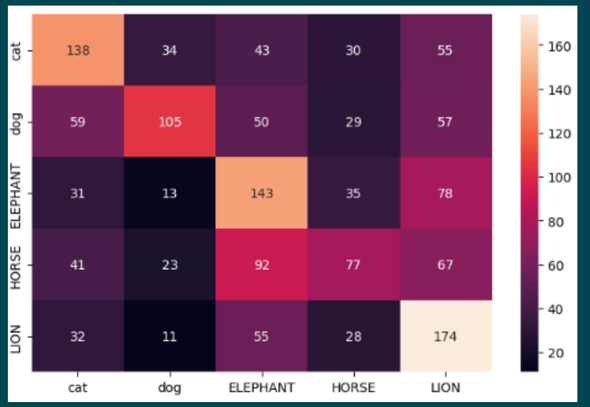
LR



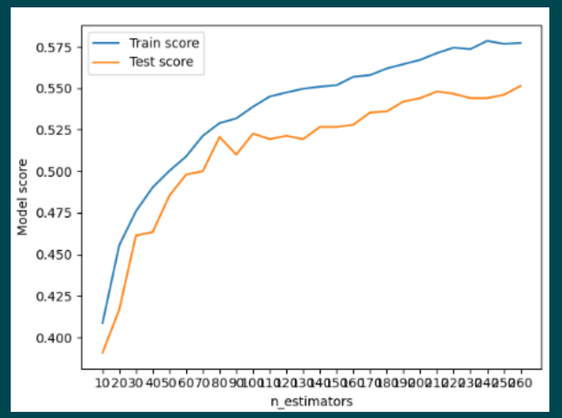


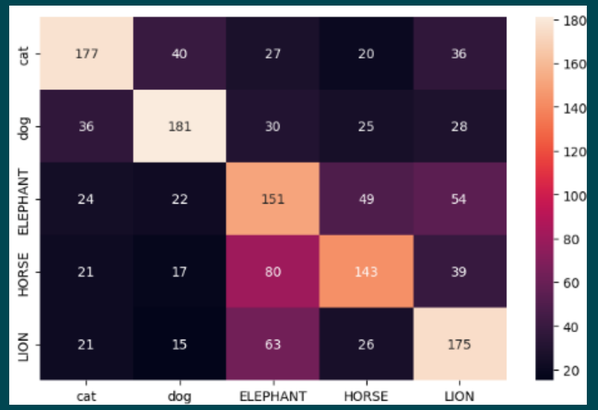
Desicion Tree





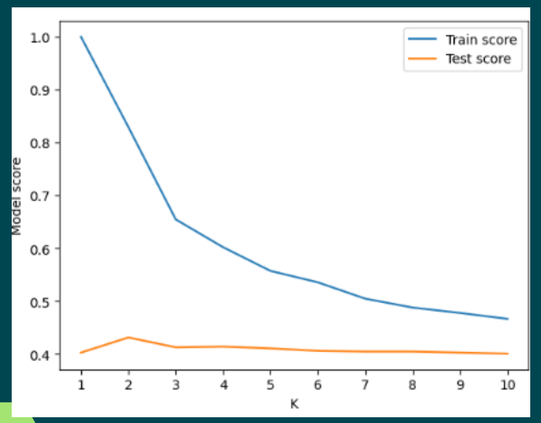
Adaboost

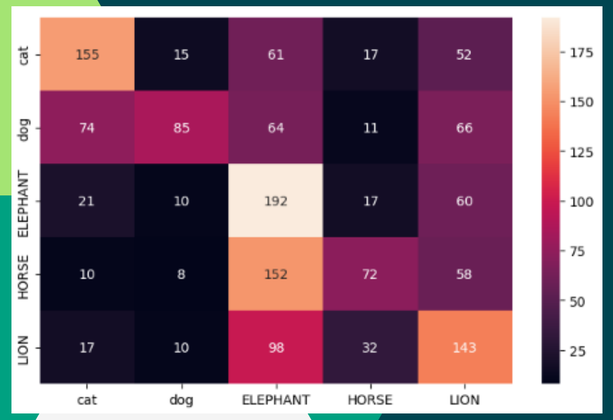




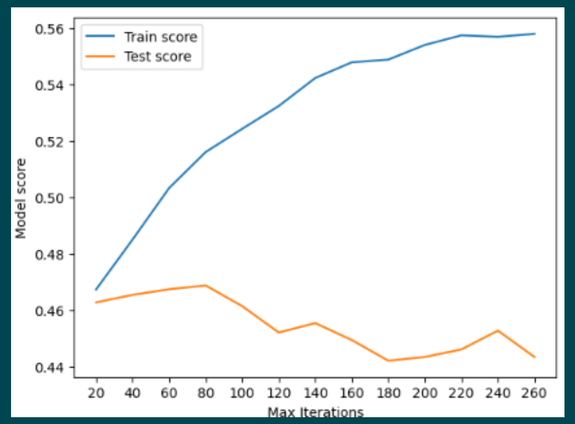
GrayScale

KNN

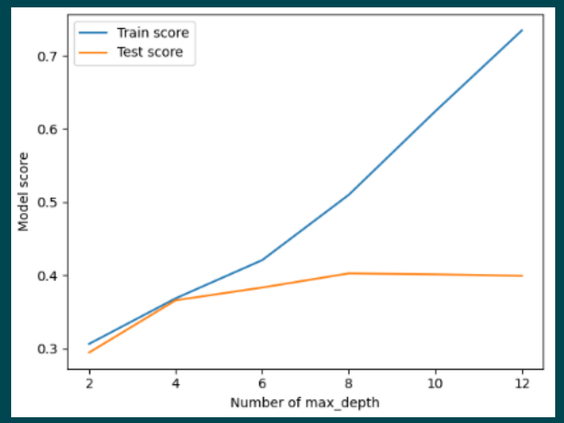


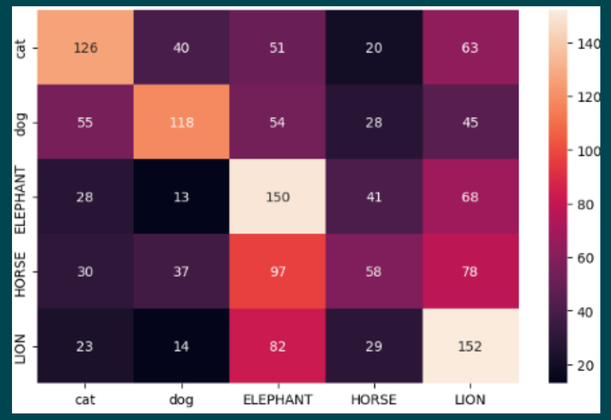


LR

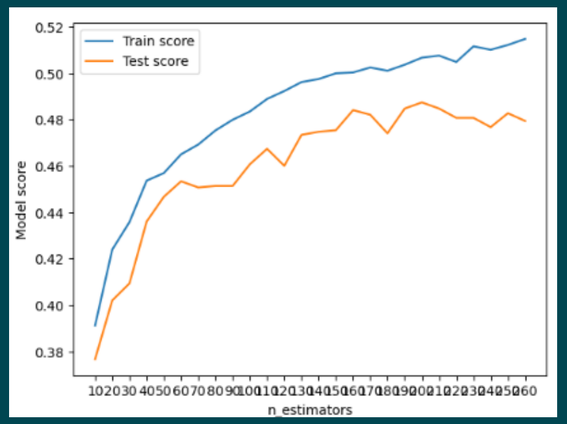


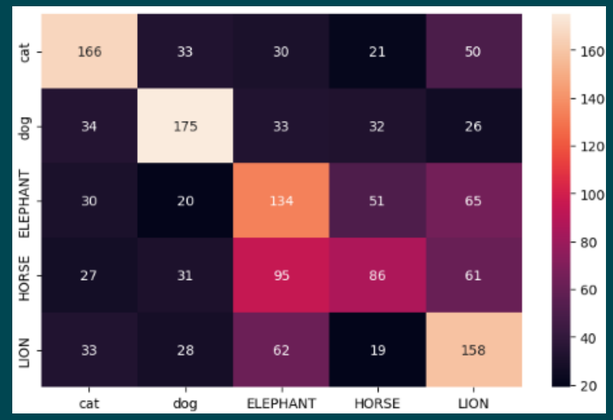
Desicion Tree





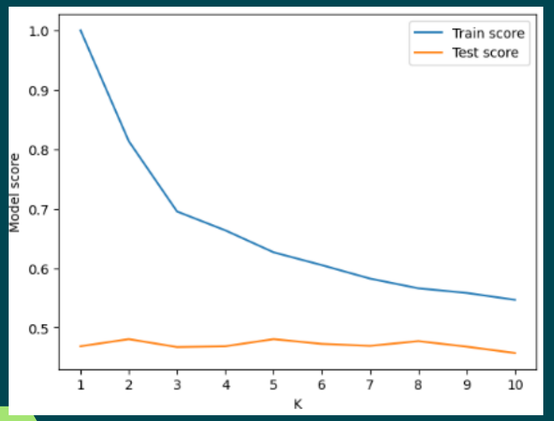
Adaboost

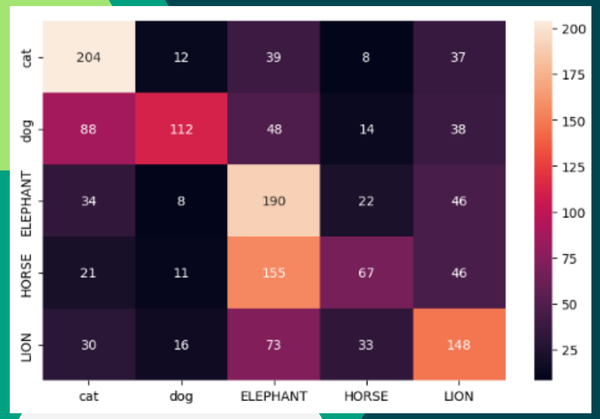




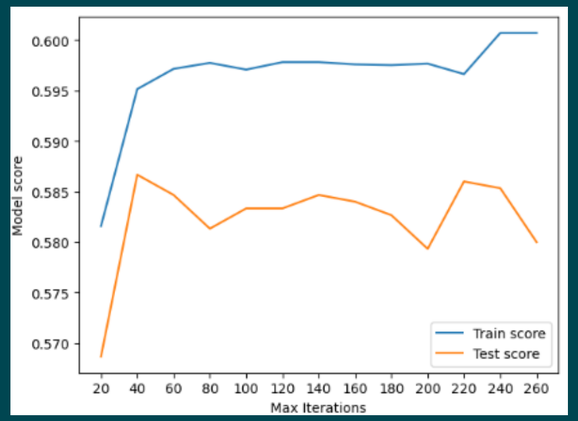
After PCA

KNN

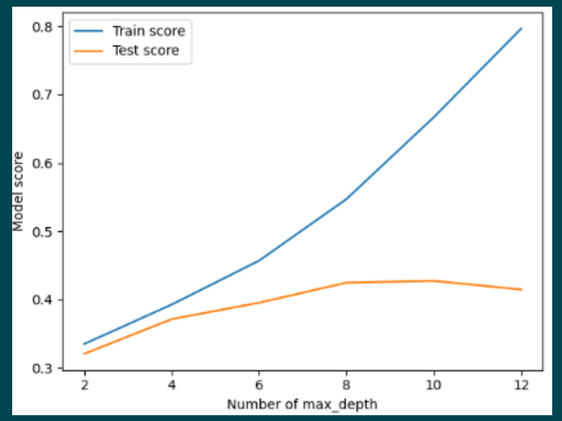


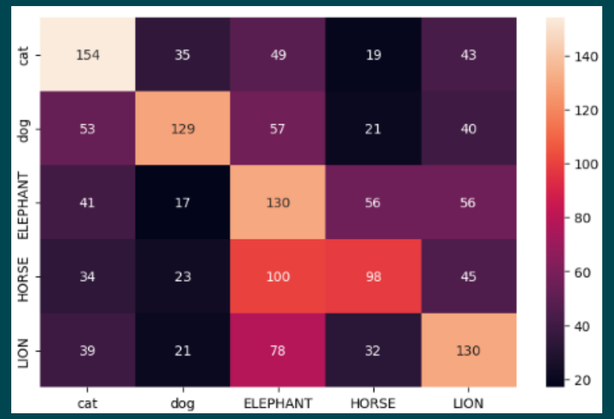


LR

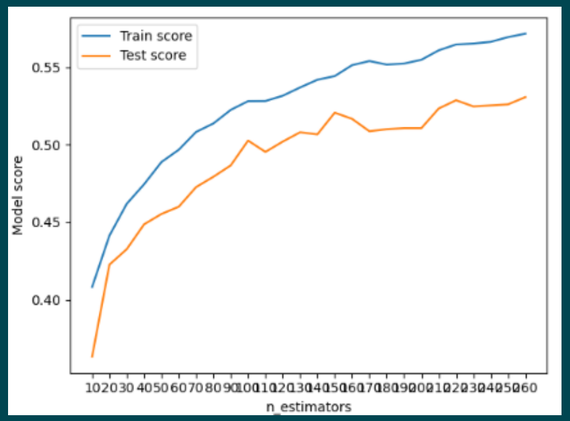


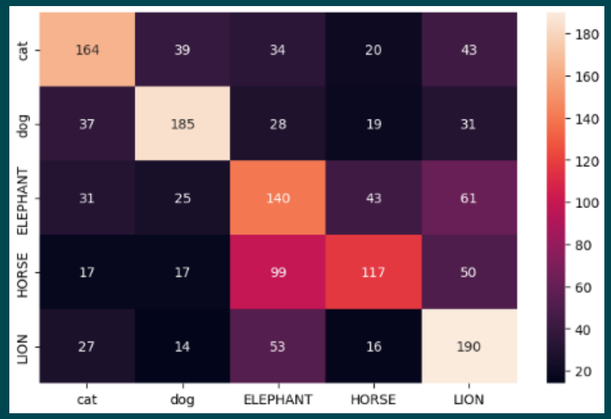
Desicion Tree





Adaboost





Our Main Questions

1) Does one of the algorithms we used worked better than the others?

---

2) Are there types of mammals that are particularly similar and thus more difficult to classify?

---

3) Did using PCA on the images or converting them to grayscale lead to better results compared to classification on the original images?

1) Does one of the algorithms we used worked better than the others?

In our analysis, Logistic Regression emerged as the top performer for color (RGB) images, achieving an accuracy of 60.00% before PCA and 58.67% after PCA. However, when it came to grayscale images, AdaBoost surpassed other algorithms with an accuracy of 48.46%.

RGB Images:

- KNN: 45.07%
- LR: 60.00%
- Decision Tree: 42.47%
- AdaBoost: 55.13%

GrayScale Images:

- KNN: 43.13%
- LR: 46.87%
- Decision Tree: 40.27%
- AdaBoost: 48.46%

with PCA:

- KNN: 48.07%
- LR: 58.67%
- Decision Tree: 42.73%
- AdaBoost: 53.06%

2) Are there types of mammals that are particularly similar and thus more difficult to classify?

Indeed, it's evident from our analysis that among the mammals in our dataset, the horse and elephant exhibit the highest degree of visual similarity.

This similarity can present a significant challenge for classification algorithms, as discerning between these two species based solely on visual features may prove more challenging compared to distinguishing between other pairs of mammals in the dataset.

3) Did using PCA on the images or converting them to grayscale lead to better results compared to classification on the original images?

When analyzing the results, it's evident that the performance was poorest on grayscale images. This suggests that color is a crucial factor in image classification. It's reasonable to assume that in images with multiple colors, color serves as a vital aid in classification, thus resulting in improved performance.

After implementing PCA for dimensionality reduction, we observed that the results were similar to those obtained before PCA. This is likely because, despite reducing the details and focusing primarily on the fundamental basis of the data, the quality of classification was not significantly compromised.

Conclusion

In summary, our animal classification project employed machine learning techniques to classify mammal species from images.

We used 4 different algorithms to classify the images - in RGB, GrayScale and after using PCA.
We've seen a significant improvement from educated guesses.

We had a great time researching the improvement of classification by the algorithms, it gave us a lot of tools with which we could enter the field of machine learning in the real world.

The possibility of expanding the research

- reading the data in a different way
- downloading the background and classifying only according to the data that is important to us
- Using built-in functions like Gridsearch
- split to data , train and validation

Do you have any questions?

Thank you for listening!!!