**MI2 Sea Monsters Checkpoint**
**Group: The Sea Monsters**
Group Leader: Sharaf Tariq Group Members: Alexa Owen, Audrey Himes
**DS 4002**
**March 17, 2023**

**Hypothesis:** The model will correctly be able to identify the type of plant based on an image of its leaf with an accuracy of 0.75 or higher.

**Research Question:** Can different types of plants be classified solely based on images of their leaves?

**Model Approach:** We will conduct image classification on various plant images using the Keras model in R. This model is trained using a set of images and then will classify new images based on the data provided by the training data set.

**Executive Summary:** Image classification will be performed on the chosen plant leaves data set. The Karas model used will accurately predict the type of plant based off of an image 0.75 or 75% of the time.

**Data Set Establishment Details:** The data set contains 4,503 images of plant leaves for 12 different species of plant. The part of the data set used to train the model contains 4,274 images with 2,163 of those images being of healthy leaves of various plants and the remaining 2,111 images are of diseased leaves. The folder of images used to test the model contains 110 images, half of which are healthy leaves and the other half are diseased leaves. The dataset also contains a folder with 8 images to have the model predict and a folder containing 110 images to validate the results of the model.

| Element/Variable Display Name | Description | Number of Images |
|---|---|---|
| Images to Predict | Images to predict the species based on the leaf | 8 |
| Test | 24 folders containing images of diseased and healthy leaves for each plant species to test the model | 110 |
| Train | 24 folders containing images of diseased and healthy leaves for each plant species to train the model | 4274 |
| Valid | 24 folders containing images of diseased and healthy leaves for each plant species used to validate results | 110 |

This dataset is available on Kaggle [1].

What types of leaves are in the images in the datasets?

There are twelve different types of leaves indicated in the four datasets. These are Mango, Arjun, *Alstonia scholaris*, Guava, Bael, Jamun, Jatropha, *Pongamia pinnata*, Basil, Pomegranate, Lemon, and Chinar. *Alstonia scholaris*, also known as the blackboard tree or the scholar tree, is mainly found in Southeast Asia and Tropical Oceania. Their natural habitats are rainforest and monsoon forest [2]. *Pongamia pinnata* is commonly known as the malapari or the karanja tree and is mainly found in humid and subtropical regions of Southeast Asia and Oceania as well [3]. All the plants in the datasets are found in more tropical regions of Asia, Oceania, and the Pacific Islands.

Hopefully, we can use the algorithm we will develop to classify plant leaves in different regions of the world as well.

We had originally planned to work on a dataset of pictures of fruits and vegetables, but we made a change because we thought that people are more likely to want to classify plants instead. We know what different fruits and vegetables are because they are labeled at the store. However, it is much harder to know what different plants are out in nature, so this would be more applicable.

Is the data balanced?

For the sake of building our model, the data is divided into training, testing, validation, and prediction sets. It is important that the data is well-balanced for the sake of the accuracy of our model. The distributions are as follows:

- Training Data:
  - 11 different species
  - 2,111 photos of diseased plants
  - 2,163 photos of healthy plants
- Testing Data:
  - 11 different species
  - 55 photos of diseased plants
  - 55 photos of healthy plants
- Validation Data:
  - 11 different species
  - 55 photos of diseased plants
  - 55 photos of healthy plants
- Prediction Data:
  - 8 unclassified images

We can see that of the four different sets of data, both the testing and validation sets are perfectly balanced. The training set however, has a slight skew towards healthy plants. This difference is small, so it is not probable that this will cause a large discrepancy within our model.

How large are each of the images?

It appears as though each image is between 1.3 and 1.9 megabytes. The physical size, however, could be altered when importing the data for analysis. It could be possible that there is some type of correlation between the memory size of the image and what sort of classification it has, although this is not the case. There are no significant correlations between the memory size of the image and whether the plant is diseased or healthy. However, some species have lower average memories than others – this could be due to the coloring of leaves or the size of leaves that contributes to the memory space needed to store the image. It does not seem like the size of the image is at all an extraneous variable that could affect the efficacy of our model for image classification.

What do the pictures look like?

There is only 1 leaf in each picture on a dark gray background. This makes it easier to see what the different leaves look like on their own. However, we do not know what the trees look like. Many people may see a tree from afar and want to know what the tree is without having to go closer to it. On the other hand, many different types of trees may look similar but their leaves are more distinct and unique, which would make our model more accurate in predicting what tree or plant the leaf comes from. Overall, there are both strengths and limitations of this dataset that can determine how accurate the model is.

Below is an image of a healthy Chinar leaf from the training dataset [1].

**Analysis Plan:**


EDA and Data Cleaning:
The data is unzipped, imported into R, and explored. Specifically, we can examine some data points that might contribute towards a deeper understanding of the data or extra factors to consider when developing and analyzing our model. We will look at the balance of the data as well as the size distribution of the images. The balance of the data is especially important when training the model so that the model does not become skewed towards any particular classification. Additionally, the size distribution of the images is just one potential extraneous variable that could also affect the results of our model. We may need to return and conduct more levels of EDA (or model tuning) if we think that there are some variables affecting the model.

Keras Model Training:
After the data is cleaned and explored, we can begin to train the Keras image classification model. First, we will set up the layers of the model by pulling data from the images. Furthermore, we will extract multiple layers and chain them together to better analyze and train the model [4]. After establishing layers, we will then compile the model so that we can properly train it. This will be done by adding parameters such as 'Loss of Function' which measures the accuracy of the model, an 'Optimizer' that will help to keep the model updated, and 'Metrics' which we will use to monitor the results of the training and testing [4]. Finally, to train the model the folder with the training images will be fed into the model to learn what image is associated with its label [4].

Keras Model Testing:
The model will be tested by feeding the testing images into the model and then seeing what the accuracy of the model is [4].

Keras Model Validation:
The model will be validated by feeding some images already tested to see if the model produces the same results as when the model was tested [4].

Analyze Accuracy of the Model:
Finally, after training and testing the model, we will analyze how accurate the model is. The images in the "Images to Test" dataset will be fed into the model. There are 8 unclassified images in this dataset. The model will return 3 values between 0 and 1 indicating the rate of precision, recall, and the F1 score for each image. The model will also provide an overall accuracy rate, also between 0 and 1, and the average of all the precision, recall and F1 score values as well.

**References:**

[1] “Plant Leaves for Image Classification,” *Kaggle*, September, 2022. [Online]. Available: https://www.kaggle.com/datasets/csafrit2/plant-leaves-for-image-classification. [Accessed Mar. 14, 2023].

[2] National Parks Board, “Alstonia scholaris (L.) R. Br.,” *Government of Singapore*, Mar. 10, 2023. [Online]. Available: https://www.nparks.gov.sg/florafaunaweb/flora/2/7/2705. [Accessed Mar. 16, 2023].

[3] Center for International Forestry Research, “Energy From Forests: Pongamia pinnata,” *CIFOR-ICRAF*, 2023. [Online]. Available: https://www.cifor.org/feature/energy-from-forests/millettia-pongamia-pinnata/. [Accessed Mar. 16, 2023].

[4] TenserFlow, “Image classification,” *tensorflow.org*, para. 1, Dec, 15, 2022. [Online]. Available: https://www.tensorflow.org/tutorials/images/classification. [Accessed Mar. 14, 2023].