Name: Canyu Hankins

Project Name: Pollen classification and database

1. **Introduction**

The idea for this project came from one of my friends, who wanted to create a database for some hardware inventory for a student organization at UTA. Then I started to wonder if there is a way to use data science techniques to tackle down this problem: instead of enter information and create a database manually, can I let the computer do it for me?

I figured we can use image processing algorithms to train datasets so that the computer could automatically recognize the item from an image and create a query in the database. On the other hand, the same technique and process can also be used for similar cases, such as classifying different classes of pollen and create databases to store their information for a museum, and this will enhance their management. This white paper is intended to provide an overview of data mining on the Pollen image dataset, including data preparation, exploratory data analysis (EAD), modeling and evaluation.

1. **Dataset:**

The dataset is in jpg format, they are all images of 18 classes of pollen, and each folder contains about 20 images per pollen type. I got the dataset from Dr. Rostami. Each image is different in size, the image in the anadenanthera folder is about 15 KB per image, but in the myrcia folder, the image is average about 160 KB per image.

1. **Data Preparation:**

The first step in data mining of the Pollen dataset is to prepare the data. But unlike the traditional dataset, it does not come in as the regular csv file, or contains any text. I can utilize google colab to do the image processing. I will need to convert the image into some pixel values, so I can run analysis on them, to do this, I need to convert images into a numpy array.

1. **Exploratory Data Analysis (EDA):**

After the data has been prepared, the next step is to perform exploratory data analysis (EDA). EAD involves visualizing and summarizing the data to gain insights. For the Pollen dataset, this may involve checking the average image for each class, to do so, we can take the average value of each pixel across all observation. Then we can compare the contrast, variability between average images. Some dimension reduction technique could also help in our case, such as using principal component analysis (PCA) to visualize the components. Furthermore, we can use feature extraction to break down different details for the images. EDA can help identify patterns in each image and help select a more appropriate model in the model selection.

1. **Modeling**

The next step is to select a modeling technique. I want to select 3 algorithms such as VGG16, this is one of the most popular pre-trained models for image classification. Xception could be another model. On top of that, Keras and TensorFlow also have some pre-trained models I can use, and I will explore more on different models to see which one might be more suitable for the Pollen data set.

1. **Evaluation**

Model evaluation and calculating the error is always an important step. I would use F1 score to calculate accuracy, precision, recall to determine how the model performs.

The F1 score is calculated as follows: .