**Data Drift Monitoring for Images:**

During the Training phase, all machine learning models are based on patterns in predictors or features used in training data. However, in production, Machine learning models often deal with corrupted, late, or incomplete data which might compromise the predictions and hence the business outcome. This is the business case behind model monitoring. We will look at one type of model monitoring called data -drift for images in this case study.

**What is Data drift?**

**Data drift**, **feature drift**, **population**, or **covariate shift**. Quite a few names to describe essentially the same thing.

**Which is: the input data has changed. The distribution of the variables is meaningfully different. As a result, the trained model is not relevant for this new data.**

**Does data drift affect machine learning models for images?**

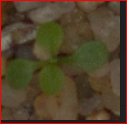
The use case for data drifts in images occurs when the training data comprises of a set of images say plant seedlings of 2 leaves but, the model in production receives images of mature seedlings or it receives images of seedlings of a different type of plant which was not a class in the model.

The picture below shows a sample of images for 2 types of plant seedlings. Our goal is to determine if there is a way to identify data drift of the features between the two samples of images below.

Sugar Beet:

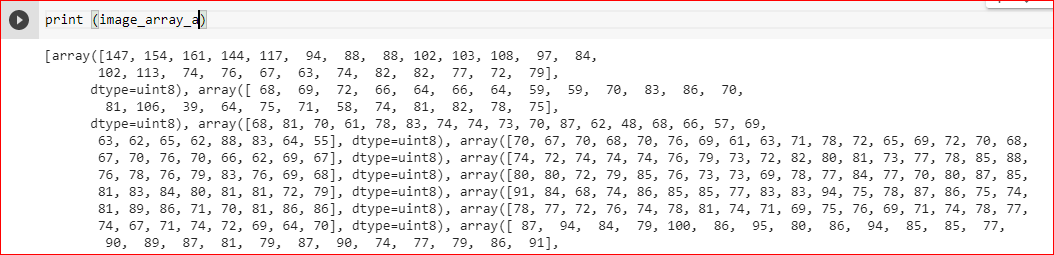
   

Shepherds Purse:

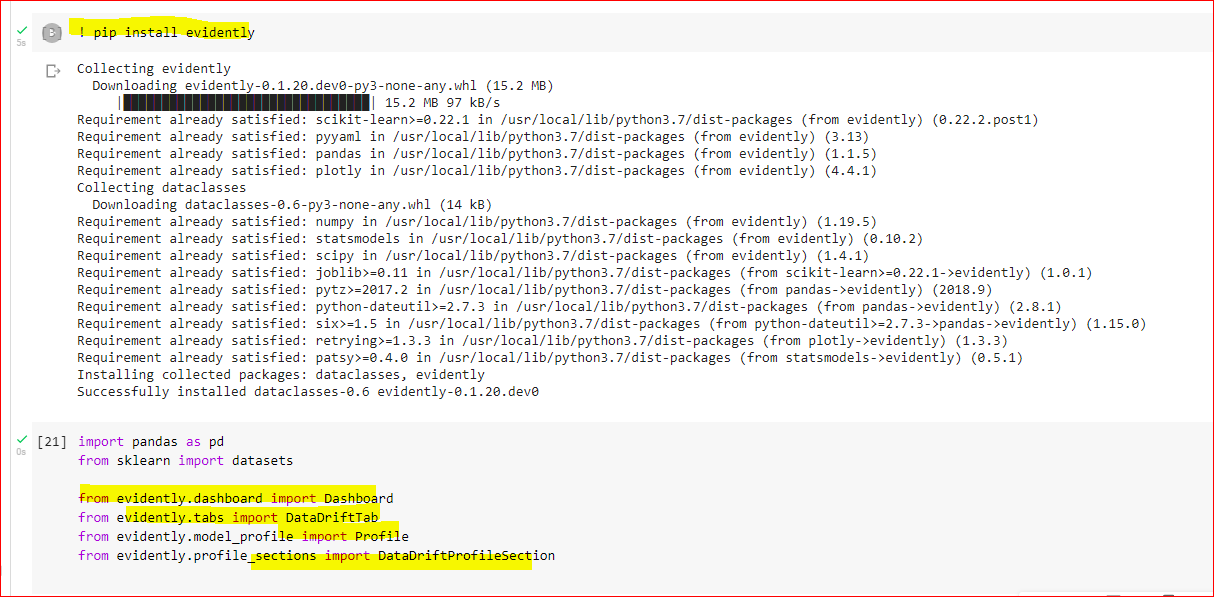
  

Here are the steps used in EvidentlyAI to identify data drift in the images.

1. Read images of original training dataset as an array of pixels.
2. Convert the images into grayscale to reduce the channels in the data.
3. Re-size the images into a 5 by 5 array of pixels.
4. The image\_array of Sugar\_Beet images created now contains 25 features per image.



1. Same steps are repeated for the images of Shepherd’s Purse.
2. Installing Evidently libraries and import the relevant modules.



1. Create dataframes from the image arrays above.
2. Create Evidently dashboards for comparing two image sets
   1. Comparing images of sugar beet with different images of sugar beet should show no drifts.



* 1. Comparing images of sugar beet with images of shepherds Purse should show data drift.



Advantages of using Evidently for data drifts for images:

1. Evidently is an open-source library which was easily installed on Colab and the modules needed were imported quite easily.
2. The image data needed to be re-sized and changed to grayscale prior to processing. This part could get tedious for certain image types.
3. Once the libraries are imported and the images converted into dataframes the data comparison dashboards can be created to check for data drifts.