**Plant Seedling Classification Project**

**Executive Summary**

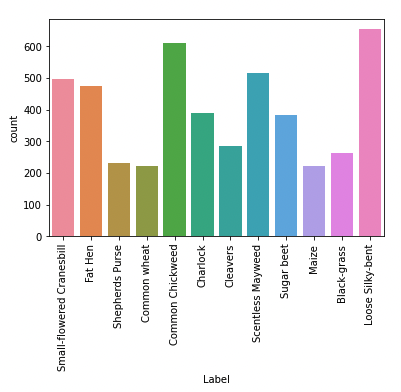
* We developed a series of Convolutional Neural Network models, that can help classify and predict from different images, different species of plants.
* Different techniques were used such as Data Augmentation, Spatial Dropout, Batch Normalization and Transfer Learning.
* We were able to come up a model that predicts with 90% of accuracy a plant’s species from a set of images.

**Business Problem Overview**

* Creating a classifier capable of determining a plant species from an image can be applied in; detecting and identifying when some crops have pests and diseases. Which consequently can help:
* Reduce the manual and visual inspection of plants (crops) that can be consuming in human labor and can help reduce the bias and human errors associated with it.
* A more rational and cost-effective use of fertilizers that can change the costs structures of agricultural plants and producers.
* Reduce waste, and help disease prevention in future crops.

**EDA Results**

* A critical aspect of the data used for the modelling is that there is some imbalance among classes, that is, the data set does not contain the same number of cases for each class. As noted in the chart below. We tried to correct this issue using SMOTE, but we weren’t able to do it satisfactorily.
* Although, our final model achieved an acceptable predictive and generalization capacity, we think that correcting for class imbalance would it make it even better.

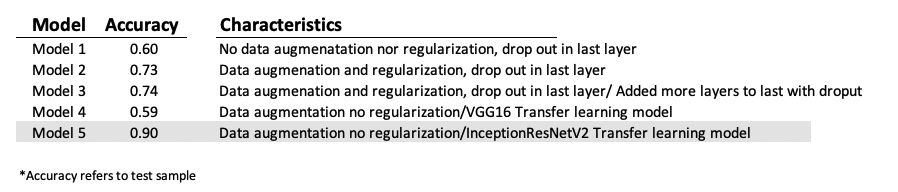


**Data Preprocessing**

* Images were resized to make the CNN work faster and efficiently
* The data set was divided in three distinct samples for crossvalidation and prevent overfitting; (training set, validation set and test set)
* Label Binarizer was used to transform the labels into numerical outputs; Label Binarizer is different from one hot encoding because it reduces the sparsity of the data.
* The data was normalized dividing the samples by 255 which is the max number of pixels for an image, so that each image can get values between 0 and 1, and thus, help us avoid getting stuck at local optima and avoiding exploding gradient problems.
* Data augmentation was employed for some models such as geometrical transformations, that can help avoid overfitting

**Model Performance Summary**

* Five different models were developed, data augmentation, batch normalization, spatial dropout and transfer learning were employed to tune and prune this models; these were the results:



**Conclusion**

* Transfer learning algorithm” InceptionResNetV2” improved significantly the accuracy of all the models, still, due to the class imbalance of the data set, class 0 is not very well classified and still present some room for improvement.
* The take away from these models is that transfer learning helped us achieve significantly higher performance, very fast with a limited dataset, using a pre-trained model.