# assignment4

November 20, 2022

## 0.1 Assignment 4 - Spark ML

Name: Bruno Gonzalez

## 0.2 Learning Outcomes

In this assignment you will:

- Use ML piplenes
- Improve a Random Forest model
- Perform Hyperparameter tuning

**Question 1: (5 marks)** In our learning from this module, we have identified a fairly significant link by leveraging the ML pipeline, a more sophisticated model, and better hyperparameter tuning. However these results are still a bit disappointing. With that being said, we're working with very few features and we've likely made some assumptions that just aren't quite valid (like zip code shortening). Also, just because a rich zip code exists doesn't mean that the farmer's market would be held in that zip code too. In fact we might want to start looking at neighboring zip codes or doing some sort of distance measure to predict whether or not there exists a farmer's market in a certain mile radius from a wealthy zip code.

With that being said, we've got a lot of other potential features and plenty of other parameters to tune on our random forest so play around with the above pipeline and see if you can improve it further! Note: adding a feaure for the distance measure is just an example and not a mandatory change to improve the model's performance. We also aren't concerned about if the model's perforamnce is actually improved! We simply want to see if changes have been made to the code for possible improvements.

Learn mode about the Farmers Markets dataset, here: https://catalog.data.gov/dataset/farmers-markets-directory-and-geographic-data

You may use the same classifier we built in the notebook( command cells 65 to 82) in this module.

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5915990090

**Question 2 ( 7 marks)** Using the Apache Spark ML pipeline, build a model to predict the price of a diamond based on the available features.

Read from the following notebook for details about dataset.

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5915990090

Note:
- Please submit the **published** notebook link in a word/pdf document. Do not submit HTML, IPython notebook, or archive (DBC) formats. - If you receive an R_Squared value that is negative, that is okay. This may occur due to the low sample size of the data.

Question 1 link: Notebook1

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/14053396234

Question 2 link: Notebook 2

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/14053396234

[ ]: