**Instacart - Customer cart prediction and recommendation**

**Team Members**: Siddartha Rao, Siddharth Kothari and Vishal Singh

**Link for project repository:** Project files and data sets are uploaded [here](https://drive.google.com/drive/folders/1dRQ6EkNtY2fhbxMVoetP-eE983DIL1UT?usp=sharing).

**Data set:** The dataset consists of 6 files:

1. ’aisles.csv’: Contains aisle id and respective aisle name
2. ’departments.csv’: Contains department id and respective aisle name
3. ’orderproductsprior.csv’: Contains transactional history of which product were purchased in each order, were they reordered, and their add to cart order id and respective aisle name
4. ’orderproductstrain.csv’: Contains details of which product were purchased in the last order of few users, were they reordered, and their add to cart order id and respective aisle name
5. ’orders.csv’: Contains more details about each order like day of week, order number etc, also tell which set (prior, train, test) does the order belong to
6. ’products.csv’: Contains product details of product name, their aisle and department

**Files and flow of project:**

Please run the file in Kaggle as the data is large and cannot be run on individual machine, and also the reading of the file is done according to Kaggle standard

1. Exploratory Data Analysis: EDA has been done in file ‘EDA.ipynb’
2. Prediction: Two models has been build for prediction. The files were runned on Kaggle kernels.
   1. Light GBM: The model has been build in file ‘Light GBM - feature, parameter tuning.ipynb’. The file has an output sub\_lgbm which is the prediction on test data and need to be submitted in kaggle to get accuracy.
   2. XGBoost: The model has been build in file ‘XGBoost.ipynb’. The file has an output sub\_xgb which is the prediction on test data and need to be submitted in kaggle to get accuracy.
3. Clustering: Clustering has three files , all of them are run on Kaggle or Big Red II–
   1. ‘Customer Segmentation - Product Pairs.ipynb’ – Clustering done based on product pairs.
   2. ‘Customer Segmentation - Aisle.ipynb’ – Clustering done based on aisle ids. Use this file’s output for recommendation
   3. ‘pca-product\_pairs.py’ – PCA done on product pairs
4. Recommendation: The files are –
   1. ‘Apriori.ipynb’ – Apriori algorithm was done in this file and output was used in recommendation.
   2. ‘Recommendation.ipynb’ – output files from ‘Customer Segmentation - Aisle.ipynb’ and ‘Apriori.ipynb’ with data set givens to create recommendation list