# Progress Update 2 Part II

## Data Preparation

The dataset prepared contains 1605 product from over 200 brands. The features target is the same as reported in the last progress update. However, multiple scenarios of data duplication have been carried out (Ensemble Boosting) to overcome class imbalance scenario and even Ensemble bagging hasn’t proved much vital in the scenario. For overcoming noise and misleading values and to maintain the integrity of the dataset NaN values or null values for features required in the models have been dropped such as missing brand names. Along with this for providing valuable data to Decision trees and K-Neighbor user with only 1 review were dropped along with products that had only a single review, leaving the important data for review analysis collaborative user-item feedback to around 6000.

## Content Based Filtering

The [notebook](#notebook) attached at the end contains all the results and evaluation for the work done today. EDA is still to be added in the process making a vital part of the report. The model implemented for the collaborative feedback include KNN, decision trees, Tf-IDF vectorization plus additional formats of these algorithms implemented using Ensemble learning and K-Fold cross validation. Still the process results in about 50-60% accuracy for these models due to a rather stochastic nature of the data collected. Data specially in the domain of skincare products proves to be a challenge, with uneven features and lack of amount of data available on the internet even after a thorough scraping carried out from the internet. These models require consistent data for a single user rather than the number of users and even if we collect specific data from a single source with less data count, but consistent data it would prove to be unambiguous for testing/predicting on data from an external source. However, Tf-IDF vectorization in this scenario proves to be vital with matching ingeredients+brand+category and provides sufficient results. This model-based approach will be carried forward for integrating in to the application.

## Collaborative Filtering

Multiple continuous learning algorithms tried and tested have failed to produce sufficient results for this scenario. Following an approach for item-item based filtering which includes rating-based feature matrix has proved to be the more viable than those implemented from the previous approach, due to the similar features of lacking sufficient and quality data. This feature matrix calculates the ratings provided by the users along with the product ID and then recommends future items based on correlation of the ratings in the matrix. Other detailed implementation can be viewed from the attached notebook or its equivalent html for quick viewing. Another last approach could try to feed the user-user features into a RNN which could prove a bit better results probably, but still the consistency and data count will prove to be challenge.

## Tomorrow’s Goal

Working on to improve accuracy and integrate these algorithms to the flask application and make a firebase connection.

 