Unsupervised Machine Learning (DS 5230) - Project Abstract
Shagun Jindal (001596493), Pranshu Diwan (001347331)

Our aim is to classify a set of food items into their appropriate categories using various unsupervised Machine Learning Algorithms. We have a dataset available that contains a list of more than 13,000 food items and their id's. These food items are actually the items on the menu of restaurants for various restaurants in Mumbai, India. Some examples include 'Chocolate Belgian Waffles', 'Butter Chicken', '30ml - Smirnoff', etc. The dataset also contains the corresponding item_id for every item. It may so happen that the item_id is unique but the food item may be the same. For example, the food item 'Egg Fried Rice' might have item_id 34 and 46. This is because some food items are common across all restaurants, and item_ids are assigned based on the food items for a particular restaurant.

The aim of this project is to classify all these food items into appropriate categories. These categories can be decided by us. Categorizing these items will help the restaurants gain insights into consumers and their favorite choices. Let's say I as a consumer love to have desserts a lot. This will enable the business to identify dessert-related recommendations and offers tailored and personalized to me. Please note that the orders dataset isn't publicly available to protect the customer's identities.

Initially, we wish to begin by cleaning the dataset by dropping unnecessary columns, converting data to lowercase and removing digits or specific words that have no significance. Next, we wish to vectorize all menu items. Vectorizing text data is important as it enables us to feed that data into our models. We would also have our categories decided. We wish to use vectorization techniques like Word2Vec, GloVe and fastText. The task is now to assign each row in the dataset to a category defined in the previous sentence. To explore the data, we could begin by finding the most common dishes and creating a word cloud of menus. EDA will play an important part in finalizing our categories. We were thinking to begin with vegetarian, non-vegetarian, desserts, alcoholic beverages, and non-alcoholic beverages as categories.

To categorize the vectorized food items, we will first use PCA for dimensionality reduction and manually check for any clusters or patterns. We will then begin with clustering algorithms like K-means, K-medoids and Spectral clustering. We answer the question of our model accuracy by applying various evaluation metrics taught in class, like Silhouette analysis and Elbow method.

Our end result will contain the database with every food item categorized into one category. This will help the business analyze the customer buying behavior and make decisions on personalized communications and offers for these customers.