IBM Professional Certificate: Capstone Project

**Introduction and Business Problem:**

India is a diverse country with vast variety of transportation systems with heavy reliance on it. One such gigantic network of transportation system is the Indian Railways network. States such as Delhi, Mumbai and Kolkata top the charts in the number of Railway stations and the total KM stretch.

Taking Kolkata for instance, it has more than 350 stations with more than 1300 km of tracks throughout the state. This operates 24 hours a day. With such a heavy reliance, people often find it hard to have a meal while waiting on the train station because of the lack of shops. This is not only the case with common passengers but also with the staff members who work day in and day out.

The objective of this project is to find the nearby food stops (near Kolkata) using Foursquare data to provide the necessary supper to the common passengers and the staff members while escalating the revenues for the local shops.

**Data Required:**

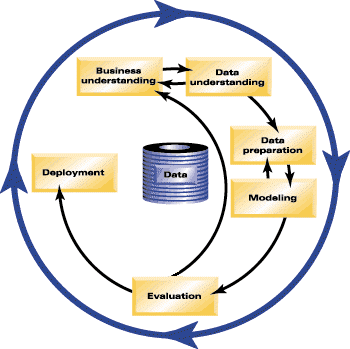
* Kolkata Neighborhood Data:
  + All the neighborhoods of Kolkata state are listed on Wikipedia and this project will perform Web Scrapping using Beautiful Soup to extract the HTML version of it.
  + This data will be cleaned and stored into a pandas Dataframe.
* Geocoding:
  + To get the exact coordinates like the latitude and longitude of each neighborhood on Kolkata state, this project will be using Google / Geocoder Python package.
  + New columns will be appended into the already existing dataframe.
* Venue Data:
  + The nearby venues in the Kolkata state for the locations obtained after web-scrapping will be extracted using Foursquare API call. For this, there is a need to pass in the API code and API secret key.

**Methodology:**

Exploratory Data Analysis was performed on the final dataset to check for any outliers or statistically incorrect measures. Moreover, a partition base clustering (unsupervised) was performed as a Machine Learning measure.

CRISP-DM also known as Cross-Industry Standard Process for Data Mining, is an industry standard way to structure and organize the methodology for a Business Case. The same is used in this project.

The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.



The Description of 6 Major steps is as follows:

1. **Business Understanding:**

Focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan.

1. **Data Understanding:**

Starts with an initial data collection (wiki, foursquare) and proceeds with activities (EDA, Visualization) in order to get familiar with the data, to identify data quality problems, to discover first insights into the data.

1. **Data Preparation:**

The data preparation phase (also called as Data Cleaning phase) covers all activities to construct the final dataset from the initial raw data and putting into a DataFrame.

1. **Modeling:**

Machine Learning Modeling technique is selected and applied. KNN is used for the scope of this project.

1. **Evaluation:**

This phase involves the testing of the data and if the model if working perfectly or not. If not, it loops back to the previous steps and some changes are made to make the model work.

1. **Deployment:**

Once we are satisfied with the model, it can be deployed to the real world.

**Reasons for K-Means Clustering:**

K-Means Clustering is an unsupervised machine learning algorithm as opposed to traditional supervised machine learning algorithms, K-Means attempts to classify data without having first been trained with labelled data. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the most relevant group.

The prime reason to perform K-Means clustering to allow the model to learn by itself to form clusters based on similarity in order to have minimum intra-cluster distance and to maximize inter-cluster distance.

**Results:**

An optimal value of n was found for KNN machine learning model which resulted in a successful clustering algorithm. This has been visualized in the Jupyter Notebook using a gamut of colors to make in identifiable.

**Discussion and Conclusion:**

The food places on the outskirts of Kolkata has lower footfall than the ones which are near the railway station.

Opening more and more food stalls in the near future specially near the major railway stations can not only fulfil the needs of the common passengers or the staff members but can also supplement the average profits of the food outlets. Having a higher footfall will flourish the revenue generated by the local food stores.