METHODOLOGY



**PROJECT UNDERSTANDING**

Customer segmentation using the KNN classification algorithm. To find groups of potential new customers who have high similarities with existing customers. This will also give insights into the right marketing campaign. We have been given a dataset with several customers and their details and they have been separated into 4 segments which is based on the combination of features like work experience, age, profession, etc.

**DATA REQUIREMENT**

The Dataset was gotten from Kaggle [here](https://www.kaggle.com/datasets/abisheksudarshan/customer-segmentation?select=train.csv). The dataset is titled 'Customer Segmentation'. It shows multiclass classification of an automobile company with plans to enter new markets with their existing products (P1, P2, P3, P4 and P5). The data set have identified 2627 new potential customer. The segmentation needs to match the customers interests as close as possible so their marketing efforts would be better served.

**ANALYTIC UNDERSTANDING**

KNN is a supervised learning approach which uses proximity to make classification or predictions about the grouping of an individual data point. While it is commonly used for classification problems, it can also be used to solve regression problems. Classification problems are what we intend to solve in this project. How it solves the problem when classifying is just to assign the data point to the group that has the most observed data point out of its neighbors, and this depends on the number of neighbors selected. KNN is prone to the effects of curse of dimensionality, which means, the model doesn’t do well with high dimensional data. Data with high dimensions are the ones with a lot of features. To avoid this, we try to reduce the dimension of our data by feature engineering or dropping irrelevant features.

**DATA PREPARATION**

The features that were not needed such as ID number of the customers and other anonymized categories were dropped. Features such as gender and profession were label encoded because our model would not be able to work with strings. Age and work experience columns were also scaled, so the model doesn’t assign too much importance to them because of their relatively large values. Some missing values were dropped and some were filled. Pandas was imported to read the csv files and numpy was also imported to fix missing values. Label encoder was imported from ‘sklearn.preprocessing’ to convert sting values to numeric.

**MODELING**

After we were done with all our pre-processing work, the KNN model were fitted with 4 neighbors as we have 4 different groups. For every other parameters, the default values were used. The features (independent variable) to be used were stored in the variable ‘X’, and the target variable, which was segmentation was stored in the variable ‘y’.The dataset was divided into training and testing data. The model was fitted on the training data, and tested for accuracy on the testing data. KNearestNeighbors was imported from the neighbors module of sklearn.

**EVALUATION**

After fitting our model, the model was evaluated by scoring both the training data and the testing data. Accuracy score and classification score were used to test our model. They were both imported from ‘sklearn.metrics’.