**CSE 523 ML Project**

**Faculty mentor:**

**Prof. Mehul Raval**

Project Number: 9

Project Title: Data-driven imputation scheme for human-subject-based dataset

Group name: Syntellect

Group member details:

| ENROLLMENT NUMBER | NAME |
| --- | --- |
| Kashish Jethmalani | AU2140029 |
| Srushti Thakar | AU2140117 |
| Priyal Patel | AU2140204 |
| Riya Patel | AU2140214 |

Progress report : (Week 7)

Date: 24-03-2024 to 30-03-2024

This week we tried to make the graphs more clear. The K-means graphs were very mixed and together and due to it they were not clear to understand. To visualize the graphs more clearly we changed the scale of the graphs to see the clusters.

For getting better clusters we tried K-means++ on the datasets. K-Means++ is preferred over K-Means due to its improved initialization strategy, better convergence properties, theoretical guarantees, and scalability.

And thus the obtained clusters are better than the clusters obtained from using the K-means algorithm.

Now from the obtained results we will make 4 different dataset on the basis of the clusters obtained. After making 4 datasets we will train a model on all the datasets.

We are using XGBoost models on each cluster and trying softmax similarity scores on the datasets. XGBoost is extreme gradient boosting. XGBoost is a robust gradient-boosting algorithm commonly used for regression tasks. XGBoost is designed for speed and efficiency. It implements parallelization and tree pruning algorithms, making it scalable and able to handle large datasets efficiently. As we are dealing with a large dataset and major missing values it is best fit for it.

The softmax function is a mathematical function that converts a vector of arbitrary real values into a vector of values between 0 and 1, such that the sum of the values is 1. Softmax function makes things easy for XGBoost.

We will check the similarity score with each cluster for a record. After that we will feed it to a softmax function, it will give us the probabilistic weight for the record’s similarity with each cluster. Then we will consider these probabilities as weights. Then we will feed the record to each trained model and generate predictions.