**Projects Background**

Based on a Kaggle study, most of the data used for training machine learning models are stored in a relational format inside a database. Therefore, in recent years companies and teams such as Relational AI and Google’s Bigquery ML have started implementing databases that accept relational inputs. Performing the join on the database tables is time-consuming and increases the size of the data due to redundancies that are factorized in a relational database. As a result, it will be faster to design and implement algorithms for training machine learning models without performing the join itself. One such algorithm for training KMeans relationally is proposed in [1] which consists of two sections. This project’s goal is the implementation of that algorithm.

**Project Summary 1**

The student needs to study and understand the algorithm in [1]. Then, the student should implement K-means++ algorithm from that paper in C++ and the code should be runnable on at least 3 relational datasets namely the Favorita grocery dataset, Yelp dataset, and Retailer’s dataset. There should be enough documentation and build instruction for the code. The code must be implemented efficiently, this means some aspects of the original algorithm will be slightly modified in the implementation, and the students will be instructed about the modifications needed.

The steps are the following:

* Study the relational algorithms including the algorithm in [1]
* Download the datasets
* Implementation of KMeans++ section of [1]
* Writing the documentations for usage of the code

**Project Summary 2**

The student needs to study and understand the algorithm in [1]. Then, the student should implement the adaptive K-means algorithm from [1] in C++. The code should be runnable on at least 3 relational datasets namely the Favorita grocery dataset, Yelp dataset, and Retailer’s dataset. The students should assume the sampled centers in the weighting algorithm is given as an input to them. There should be enough documentation and build instruction for the code. The code must be implemented efficiently, this means some aspects of the original algorithm will be slightly modified in the implementation, and the students will be instructed about the modifications needed.

The steps are the following:

* Study the relational algorithms including the algorithm in [1]
* Download the datasets
* Implementation of the adaptive KMeans section of [1]
* Writing the documentations for usage of the code

[1] Moseley, B., Pruhs, K., Samadian, A., & Wang, Y. (2020). Relational Algorithms for k-means Clustering. arXiv preprint arXiv:2008.00358.