|  |  |
| --- | --- |
| Subject: Big Data  Class: IS405.O21  Instructor: Nguyen Ho Duy Tri | Group: 07  Le Minh Nguyet – 21521211  Nguyen Cong Nguyen – 21521200  Ho Duc Truong – 21522730  Nguyen Phuong Tung – 21520524 |

**Project: Customer Segmentation With K-means Clustering**

The project undertaken by the team focuses on applying the K-Means algorithm combined with Apache Spark (PySpark) for customer segmentation in the marketing and customer relationship management domain. The primary objective is to enhance business performance and optimize marketing strategies by gaining a deeper understanding of customer segments and customizing strategies based on the characteristics and behaviors of each segment.

The dataset was collected from Kaggle with the context of an automobile company planning to expand into new markets with their existing products. It comprises nine columns, including information on gender, marital status, age, graduation status, occupation, work experience, spending score, and family size. After preprocessing steps such as categorical variable encoding, data normalization, and handling null values, the data is prepared for the segmentation process.

The K-Means algorithm is deployed on the preprocessed and exploratory data analysis (EDA) conducted data, and the results are evaluated using the Silhouette score to assess the clustering quality. The project also emphasizes the utilization of Apache Spark (PySpark) for efficient processing of large-scale data, optimizing performance, and reducing dependency on available support libraries.

Finally, in the conclusion section, the project summarizes the results and proposes future directions such as performance optimization and integration of solutions for noisy data.