**NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

-----🙞🙜🕮🙞🙜-----



**BIG DATA**

**CLASS: IS405.O22.HTCL**

**Instructor: Nguyen Ho Duy Tri**

**TOPIC: CAR RATES CLUSTERING AND PREDICTION**

1. **Summary of topic content:**

The analysis commenced with a meticulous data preprocessing phase, which involved cleaning and organizing the dataset from Kaggle. This included tasks such as handling missing values, encoding categorical variables, and scaling numerical features for consistency. The processed data was then thoroughly explored through visualization techniques, using heatmaps and line charts to uncover patterns and relationships. Clear correlations were found between the General\_rate and other attributes in the dataset. This led to the decision to use three algorithms: K-means, GBM, and Linear Regression for data mining and prediction of the General\_rate compared to the actual General\_rate of the dataset. The models were trained and evaluated using the processed data, capturing the identified relationships. Visualizations of actual versus predicted General\_rate further confirmed the effectiveness of the models. The decision to use these algorithms was reinforced by the observable trends in the visualizations, providing a robust basis for predictive analysis. The comprehensive findings, including visualizations and the models, are presented in the attached final report for thorough review and reference.