**ChBE 6745 – Group 6 Project definition**

**Github: https://github.com/yship1002/ChBE-DataScience-Group6**

**Car evaluation dataset - Training a classification model that achieves a F1 score of 80% in predicting if a car is acceptable or not based on the six input parameters given in the data**

**Background:**

The automotive industry is one of the biggest contributors to the global economy with millions of cars being sold every year. As the markets are becoming more competitive, it is important for manufacturers, dealers, and policy makers to understand consumer preferences and car performance.

The Car evaluation database is derived from a simple hierarchical decision model which evaluates a car based on six parameters: car acceptability, overall price, buying price, maintenance price, boot size and safety of the car. This model is more focused on understanding what parameters would make a car more acceptable for a consumer to purchase.

**Project goal:**

The cars have been categorized in the acceptability section under four different classes: unacceptable, acceptable, good, and very good.

The goal of this project is as follows: Training the classification model that achieves an F1 score of **80%** in predicting if a car is acceptable or not based on the six input parameters given in the dataset.

**Strategy:**

We will start with exploratory data analysis to understand the categorical features (buying, maint, doors, persons, lug\_boot, safety) and their relationship to the target variable (acceptability). Since all features are categorical, we will preprocess them using one-hot encoding to make them suitable for learning algorithms. The dataset will be split into training and test sets, with stratified sampling to preserve class proportions since there are very few good cars in the dataset. We plan to test multiple supervised classifiers—such as decision trees, random forests, logistic regression, and gradient boosting—because they handle categorical and structured data effectively. Hyperparameter tuning with cross-validation will be applied to improve generalization, and model performance will be evaluated using accuracy, precision, recall, F1-score, and confusion matrices. We will compare our best-performing model to the baseline model, logistic regression.

**Risks:**

The first risk is class imbalance, as some acceptability categories may have fewer observations, which can affect the performance of the model. To mitigate this, we will use techniques like stratified sampling, oversampling, or class-weighted loss functions. Another risk is overfitting due to the limited size of the data we have. We will address this by using pruning, depth limits, or regularization to make sure our model can be generalized. Interpretability is another risk. We could have made a model that achieves good performance on the test set, but without interpretability, it is difficult to convince stakeholders to use this model. Overall, our contingency plan is to balance model complexity with generalization, ensuring that the final model is both reliable and interpretable for predicting car acceptability.

**Dataset Description:**

For this project, we use the Car Evaluation dataset from the UCI Machine Learning Repository. The dataset contains 1,728 instances with 6 categorical features describing car characteristics: buying price, maintenance cost, number of doors, seating capacity, luggage boot size, and safety. The target variable, class, categorizes car acceptability into four levels: unacceptable, acceptable, good, and very good. The dataset is well-structured, with no missing values and consistent labeling. Preprocessing included applying one-hot encoding to convert the categorical features into a numerical format suitable for machine learning models, ensuring compatibility with common classification algorithms.