# Department of Computing

**CS 471: Machine Learning**

**BESE: 11A**

**Lab 10: feature Engineering**

**Date: 14 April 2023**

**Time: 02:00 PM-05:00 PM**

**Instructor: Dr. Seemab Latif**

**Lab 10: Feature Engineering**

**Introduction:**

Feature engineering is the process of selecting, extracting, and transforming data features in a way that improves the performance of machine learning models. It is one of the most important steps in any data science project and can significantly impact the accuracy and efficiency of the models. In this lab, we will explore the concept of feature engineering, its importance, and various techniques to apply it to a dataset.

**Objectives:**

* Understand the concept of feature engineering
* Learn different feature engineering techniques and tools
* Apply feature engineering techniques to a real-world dataset

**Tools:**

Python, Pandas, Numpy, Scikit-learn, Matplotlib, Seaborn

**Description/Methodology:**

Feature engineering is a critical step in any data science project. It requires a combination of technical skills, domain knowledge, and creativity. The goal of feature engineering is to extract relevant information from raw data that can be used to improve the performance of machine learning models. Feature engineering involves several steps, including data cleaning, data transformation, feature selection, and feature scaling.

To apply feature engineering, one needs to have a good understanding of the dataset and the problem at hand. This requires critical thinking and domain knowledge. One should be able to identify the relevant features that can help to solve the problem and remove the irrelevant ones.

The process of applying feature engineering involves the following steps:

1. Data Cleaning: This involves identifying and handling missing values, duplicates, and outliers in the dataset.
2. Feature Selection: This involves identifying the relevant features that can help to solve the problem and removing the irrelevant ones.
3. Feature Scaling: This involves transforming the data features to a similar scale to improve the performance of machine learning models.
4. Feature Transformation: This involves creating new features by combining, extracting, or transforming existing features.

**Tasks:**

In this lab, we will be using a real-world dataset (Automobile Data Set <https://archive.ics.uci.edu/ml/datasets/automobile>) to apply feature engineering techniques. The tasks involved are:

1. Import the necessary libraries and load the dataset
2. Perform data cleaning and handle missing values, duplicates, and outliers
3. Perform feature selection and remove the irrelevant features
4. Perform feature scaling and transform the data features to a similar scale
5. Perform feature transformation and create new features by combining, extracting, or transforming existing features
6. Evaluate the performance of machine learning models before and after feature engineering

**Deliverables:**

The deliverables of this lab include:

1. A Jupyter Notebook containing the code used to apply feature engineering to the dataset
2. A report detailing the steps taken to apply feature engineering and the results obtained and summarize the findings of the analysis

**Steps taken to apply feature engineering:**

1. **Handling Missing values in data:**

First the data was checked for missing values and then missing values were replaced with mean value of that column.

1. **Encoding the data:**

Categorical variables with two values were binary encoded and more than two values were categorical encoded using category encoder.

1. **Plotting the correlation map:**

**Among the features with multicollinearity between them the features that had higher correlation with target variable were kept and the other one removed**

* Correlation map of all features were plot. The features with high multicollinearity such as city\_mpg and highway\_mpg and length and wheel\_base was removed to keep only one feature among the two.
* Curb\_weight also had high correlation with length and width and with engine\_size so I dropped this column as well.
* The features with low correlation with target variable such as symbolling, num\_doors and compression\_ratio was removed from the dataset.

1. **Feature Transformation:**

Feature transformation could not be applied to dataset because all the features are independent and combining them would produce faulty results.

1. **Applying Model and Evaluating Performance:**

I selected Linear Regression model because output variable had continuous values. The model gave better score with data that had feature engineering applied to it than data that did not had feature engineering applied to it.