**Course Project   
DeVry University  
College of Engineering and Information Sciences**

**Course Numbe­r: CEIS312**

Background

## Artificial intelligence and machine learning are frontiers in the technology field. These areas are often used to address common problems that require difficult tools or skills. AI and ML professionals work with SQL, R, Python, and other tools specific to data science. Different algorithms are used to solve problems and choosing the correct algorithm can be challenging. This project will use Azure Machine Learning, which is a cloud-based service from Microsoft. Azure ML allows you to create and run experiments based on datasets and integrate custom code in SQL, R, or Python.

# Module 8: Your Own Analysis

# Objectives

* Develop your own analysis.
* Predict the consumer reports rating on cereal based on the data given.

# Scenario

Kaggle.com contains a variety of datasets for experimentation. One such dataset is the 80-cereals dataset. This dataset contains the consumer reports rating of several different cereals along with various data about the cereals. The dataset contains the following fields.

### **Content**

Fields in the Dataset

* Name: Name of cereal
* mfr: Manufacturer of cereal
  + A = American Home Food Products
  + G = General Mills
  + K = Kellogg’s
  + N = Nabisco
  + P = Post
  + Q = Quaker Oats
  + R = Ralston Purina
* Type:
  + Cold
  + Hot
* Calories: Calories per serving
* Protein: Grams of protein
* Fat: Grams of fat
* Sodium: Grams of sodium (Note: The original data contains milligrams.)
* Fiber: Grams of dietary fiber
* Carbs: Grams of complex carbohydrates
* Sugars: Grams of sugars
* Potassium: Grams of potassium (Note: The original data contains milligrams.)
* Vitamins: Vitamins and minerals: 0, 25, or 100, indicating the typical percentage of FDA recommended
* Shelf: Display shelf (1, 2, or 3, counting from the floor)
* Weight: Weight in ounces of one serving
* Cups: Number of cups in one serving
* Rating: A rating of the cereals

Create a new experiment in Azure Machine Learning and perform linear regression on the cereals.csv dataset. We are seeking to determine if we can predict the rating based on the above features. The ratings are based on nutrition and taste. Open the cereals.csv file and look through the data to familiarize yourself with the data. You will need to cleanse the data—are there any missing values? Any outliers? Do you need to make any fields categorical? You will also need to normalize the data. You will want to develop python visualizations (feel free to use the code we have used in previous examples to create a pairwise scatter plot). You may also want to focus on a few selected fields based on your visualizations (sugar, calories, fat, fiber, etc). Don’t forget to split the data into a training and test set. You will need to create a linear regression model, train the model, score the model, and then evaluate it.

How accurate is your model? What is the R2 value?

Look at the python pairwise scatter plot. What features do you want to remove from the model? Continue iterating (removing features and adding them back in) and reviewing the evaluation results. Find the best R2 value you can. Try to get above 0.70. What features are good predictors of the cereal rating?

Prepare a report on your regression analysis. In your report include several screenshots and explanations of your model. The rubric is below.

* Introduction to the problem (10 pts)
* Uploading dataset (5 pts)
* Data preparation (normalization) (10 pts)
* Data Visualization (python script) (10 pts)
* Selecting features (5 pts)
* Splitting data (5 pts)
* Linear regression model (10 pts)
* Training the model (10 pts)
* Scoring the model – show scored labels (10 pts)
* Evaluating the model (10 pts)
* Iteration process (15 pts)
  + Why you chose to exclude certain features
  + New evaluation when those features were excluded
  + What features are most influential on the rating
* Conclusion (15 pts)
* Challenges (5 points)
* Career skills obtained (5 points)
* Web service (optional)
* Total = 125pts

Note: Pairwise scatter plot code in Python.

def azureml\_main(frame1):

## import libraries

import matplotlib

matplotlib.use('agg') # Set backend

from pandas.tools.plotting import scatter\_matrix

import pandas.tools.rplot as rplot

import matplotlib.pyplot as plt

import numpy as np

## Create a pair-wise scatter plot

Azure = True

fig1 = plt.figure(1, figsize=(10, 10))

ax = fig1.gca()

sm=scatter\_matrix(frame1, alpha=0.3,

diagonal='kde', ax = ax)

[s.xaxis.label.set\_rotation(45) for s in sm.reshape(-1)]

[s.yaxis.label.set\_rotation(45) for s in sm.reshape(-1)]

plt.show()

fig1.savefig('scatter1.png')