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

**Course Numbe­r: CEIS312**

**Edward Alvarado**

**September 2022**

**Course Project**

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.

# Objectives

* Predict the consumer reports rating on cereal based on the data given.
* Achieve an R2 Value above 0.70

# 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

Diagram

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Chart, histogram

Description automatically generated

# Iteration process

* Why you chose to exclude certain features
  + In the most current iteration process I decided to exclude weight and number of cups from our analysis. After updating the data set the Scored labels improved.
* New evaluation when those features were excluded
* What features are most influential on the rating
  + The most influential to the rating was calories, sugars, and fiber from my testing.

# Conclusion

The objective of this analyzing this data set was to determine the values that contributed to the overall ratings on cereal. Looking at the project map we can see we started off cleaning up some of the missing data of our dataset. There were multiple sections that had 0 values. We replaced those values with the mean average value. We then normalized our dataset to try and account for outlier values. We split our dataset and utilized linear regression in our training model. In the most current iteration process I decided to exclude weight and number of cups from our analysis. After updating the data set the Scored labels improved. The most influential to the rating was calories, sugars, and fiber from my testing.

# Career Skills

In this course I’ve learned a lot of skills I had little to no experience with previously. The biggest skill was learning how to take a data set and transform the data into a workable set of data we can use to provide answers to our questions. Azure Machine Learning was the primary took we used during the course. During the course we also utilized Microsoft word as our primary lab work prep software. Overall my biggest takeaway was the value of machine learning and how we can manipulate data to solve real world problems.