# CS 474 Capstone Proseminar
# Data Analysis Classwork

Natalie Lee
nml3@hood.edu

Fall 2023

## 1 The Data Set

`Prices.cvs` is a data set with information about the fruit and vegetable prices in the USA from the year 2020 [1].

The CSV file contains 8 columns and 156 rows. Each row holds data about a specific fruit or vegetable. The column values of the data set represent attributes of said fruit or vegetable ranging from the form (i.e. frozen, canned, fresh) to the average retail price of the fruit.

Fruit and vegetable prices are made comparable to each other using the $CupEquivalentPrice$ variable or the price per edible cup equivalent of each item. This is the chosen unit of measurement of federal recommendations for fruit and vegetable consumption and provides a normalization of price across all fruits and vegetables.

The data set is sourced directly from the USDA Economic Research Service report which examined retail scanner data provided by the company Circana, a consumer behavior analysis company.

## 2 Price Prediction Exploration

Fruit and vegetable price prediction is an interesting exploration point of this data set. The exploration below seeks to find a model that takes information about a fruit/vegetable and accurately predicts its price.

Using the WEKA Linear Regression model, the nominal variables of $Item$ and $Form$ and the numerical variable of $Yeild$ were used to predict the value of $CupEquivalentPrice$.

This combination of variables produced a well-performing linear regression model. The correlation coefficient was 0.9668, meaning that there was a strong positive linear relationship between the predicted fruit prices and the actual data values. The mean absolute error (MAE) revealed that on average the model's predictions were off by $0.086.

To explore the topic of price prediction using another method, the WEKA SMOreg function was also run using the same variables. This method proved slightly more accurate with a correlation coefficient of 0.9674. The MAE revealed that on average the SVM model's predictions were off by $0.0475. These error margins are preferable to the linear regression model since in the context of price an improvement of about 4 cents can be non-trivial.

# A WEKA Terms

- **SMOreg:** A Sequential Minimal Optimization algorithm implemented in WEKA for training a support vector regression model.
- **Sequential Minimal Optimization (SMO):** A popular algorithm used for training support vector machines. In essence, the algorithm breaks down the larger quadratic programming optimization problem required in a training SVM into smaller parts.
- **Support Vector Machine (SVM):** A supervised machine learning technique used to solve classification or regression problems. A SVM can handle nonlinear solutions, unlike a linear regression model.
- **Regression:** In the context of machine learning, regression refers to the task of predicting a continuous target variable. In this case, the continuous numerical value of fruit/vegetable price.

# B Data Dictionary

| **Variable** | **Description** | **Type** | **Example** |
|---|---|---|---|
| Item | Name of the fruit or the vegetable. | nominal | Apples |
| Form | The form of the item, i.e., canned, fresh, juice, dried, or frozen. | nominal | Fresh |
| RetailPrice | Average retail price of the item in the year(USD). | numerical | 1.1804 |
| RetailPriceUnit | Average retail price's measurement unit. | nominal | per pound |
| Yield | Average yield of the item in the year. | numerical | 0.4586 |
| CupEquivalentSize | Comparison done with one edible cup of food. | numerical | 0.4519 |
| CupEquivalentUnit | Comparison's measurement unit. (pounds or fluid ounces) | nominal | pounds |
| CupEquivalentPrice | Price per edible cup equivalent (The Unit of Measurement for Federal Recommendations for Fruit and Vegetable Consumption) | numerical | 1.1633 |

## References

[1] USDA-ERS, “Fruits and vegetables prices in usa,” 2023. [Online]. Available: https://www.kaggle.com/dsv/6775400