Galvanize Capstone 2 Proposal

Motivation

I have always been intrigued by how the cells in our bodies interact with each other, from how we deal with stress to how what we put in our bodies. Several years ago, I came across a book called 'The Optimum Nutrition Bible', but it wasn't until last year that I was finally able to start reading it. The book explains how a variety of stimulus to our bodies affects our nutrition, which affects our health and wellness. It breaks food, drinks, sleep, stress, and several other things down to what is in them and how those parts affect our bodies in negative and positive ways.
It is easy to understand how this book caught my eye and that is why it is the motivation for this capstone.

Data

I was able to find a data set on kaggle named 'Fruits 360'. The data set is comprised of 90,483 100x100 '.jpgs' of fruits and vegetables. 90,380 of those pictures are either a fruit or vegetable and 103 have multiple fruits or vegetables. The data set has 131 classes of fruits and vegetables as different varieties of the same item were stored as belonging to different classes. I, also, built my own data set of each item and some of its nutrition facts, which I placed into a '.csv' and called as a data frame in pandas.
As part of my check of the data set I verified there were no broken nor fraudulent images.

Goal

I will shrink the images down to 50x50 and create gray scale representations using SkImage. Then I will group the varieties of the same item together and visualize the numbers of each item using MatPlotLib. I will then perform Principal Component Analysis and run a Multinomial Naive Bayes Model. I am curious to find out how this model performs with getting the correct classification. I will be analyzing a confusion matrix and the accuracy of the model as the classes have different numbers of images.

MVP+/++

After completing the NB Model, I would like to plot a ROC Curve for multiple classifications and run a Convolution Neural Network, as an MVP+. I will first check how the model performs as a binary network. I will then run this model in such a way that an item can be passed and matched with the correct nutrition fact in my pandas data frame, as an output. As an MVP++ I would like to run the CNN on the images that have different items in the same image to see how accurately it can identify each image.