Final Project Data Memo

Dylan Clausen, Noah Holubow, Max Olander, Max Tokman

04/07/2021

Project Overview

This project attempts to use the geometric properties of dried beans to predict, categorically, which type of bean it is. A computer took pictures of the beans, calculating features like shape, type, structure, and form. Then the features were extracted. Using these attributes, we hope to predict the actual bean type.

Data Source

Our data source is the Machine Learning Repository from the Information & Computer Sciences department at the University of California, Irvine, in Irvine, CA. This repository contains many complete, academic datasets for the machine learning community. In this particular dataset, there are a total of 13,611 observations recorded. The dataset is made up of strictly numerical data, again covering various geometric attributes and measurements. These will likely be regressors. The outcome will be bean type.

Link: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset

Proposed timeline

We hope to read in the dataset by the end of week 3 and have the splitting and resampling done by week 4. The recipe could be done by week 5. Afterwards, we'd like to start running the models around week 6 or 7.

Why this data

We think it is an interesting idea to use geometric properties to predict a categorical variable. Since originally these were pictures of beans from which measurements were extracted, this type of predictive modeling is in some ways similar to image recognition, which also analyzes images for various details to predict the type of image. In this instance, the values have already been extracted and we are not inputting actual images.

Potential data issues

There are not any categorical variables, which we would have liked, but the data otherwise seem to be good as far as we can see right now.