(Could not find the one I wrote and showed Bill, so am writing a new one taking out the math!)

**The difference between a regression analysis and a classification analysis:**

Principally, the two analytical techniques are after the same thing, to predict what something is. The two techniques just go about the end results in a different manner.

Regression analysis uses a loss function to predict a given outcome for a specific input to a system using a function called a regression line or “best fit” line, which is called a loss function, among other names. The loss function seeks to minimize the cumulative average differences between the input data and the line with respect to the output variation and is adjusted using parameters or weight values to minimize the error values between the line and the output values (the “y” values are subtracted from each other and the difference is the error of prediction for that particular data domain value and predicted value). A very important aspect of the analysis using regression is that the data will probably have several data points that have the same domain value with various range values, thereby creating a distribution from minimum range value to maximum range value, and this distribution for each domain value is assumed to be a Gaussian distribution for most of the models used that I have experienced, which means that it is likely to be true for the majority of models, so there are little Gaussian curves associated with each domain value that has an expected value for the range and a variance, both of which can be calculated. Never-the-less, regression analysis is used to predict an expected result for a system given a specific value within the domain. The difficulties with regression lies in values outside the domain as well as the variation associated with each value of the domain not being consistant from domain value to domain value, which adds even more error and uncertainty to the predicted results. The best models to use are a gradient model that adjusts the parameter values to minimize the error between predicted and data values.

As stated before, classification is similar to regression analysis in that it is a technique that is trying to predict an expected result. The difference is that the result is a classification of an object or item, which can be visual, physical or verbal in nature. Like classifying a type of food from a group of pictures of various fruits, vegetables and meats. The algorithm used is different and is called a K-Means calculation, where the data is clustered into categories due to similarities of the data around a specific centroid. A very good method to use to create categories is to have known categorized data sets to “train” the parameters of a loss function (similar to regression analysis), and then once the parameters are set to acceptable correlation values for each classification, then a test set is run through the model, again, knowing the classifications of the data to test the model’s capability. SciKit Learn has a good train/test algorithm for conducting the model building. The trick is to set the iterative steps so that the model is optimized for the best predictor given the data set used, so that the steps toward convergence are not too small (over fitting) or too big (under fitting) making a less than ideal non-optimized model, which means that several runs must be conducted to verify convergence to the same set of parameters for the model by changing the perplexity value a little at a time in both directions to see if the model moves with the perplxity value. If not, then the model fitting is not complete and optimized for the best result.