Homework3- Sheena Sharma

2. This data set contains data from a group of Pima Indians, a group that is known to have a large BMI but relatively low incidence of diabetes compared to Americans with the same BMI. This data sets has many variables that generally predict diabetes such as results of a glucose tolerance test, BMI, and skin fold thickness testing.

3. There are many feautures of this data set. The first is the number of times a person has been pregnant. The next is how patients respond to an oral dose of glucose as measured by a glucose tolerance test. Next is blood pressure and the triceps skin fold. Next the 2-hour serum insulin determines the amount of insulin present in serum two-hours post test. Body Mass index or BMI is the next feature. The next feature is diabetes pedigree function, a likelihood that they will develop diabetes based on ancestor data. The next feature is age and the final feature is class variable (1 or 0), where 1 means the person tested positive for diabetes. I hypothesize that BMI, pedigree function, and glucose will be relevant in predicting diabetes.

4a. Using the .isnull feature I checked for null values and there are none. Using .info I can see that all rows have the same number of values so there are no missing values either. There are zero values, but that is OK in Times\_Pregnant and Class\_Variable. In the remaining columns, I replaced the zeros with the mean of that column.

b. There are 8 features and 1 outcome variable. To be normalized the mean should be 0 and the Stdev should be 1. They didn’t appear normalized so I normalized them.

c. The data appears ordered by label