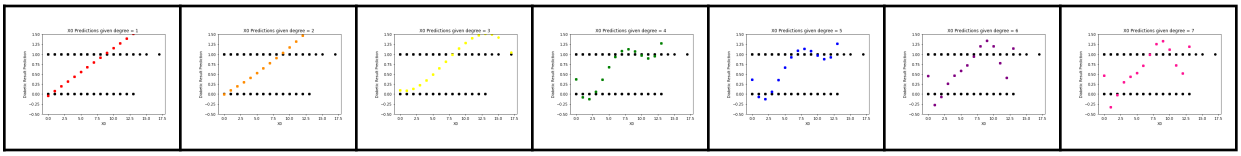
Madison Chamberlain

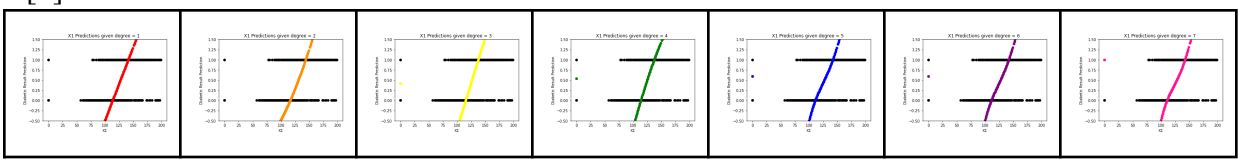
Question 1 Analysis

Below are the polynomial fits for degrees 1-7 plotted against the scatter plot for each X feature in the Pima data set.

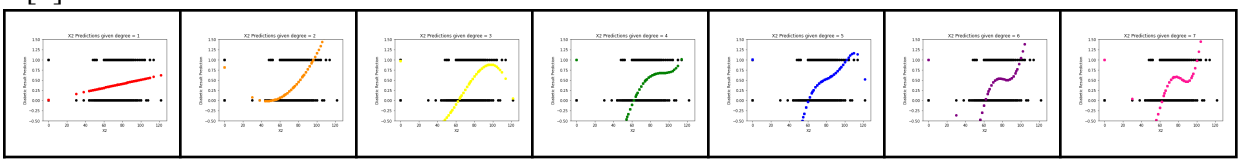
X[0]



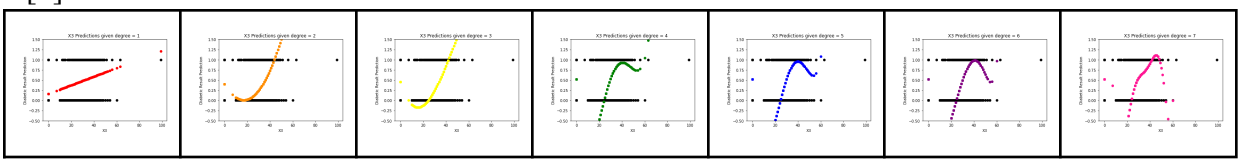
X[1]



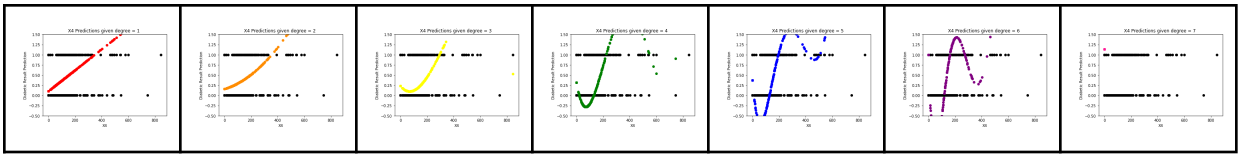
X[2]



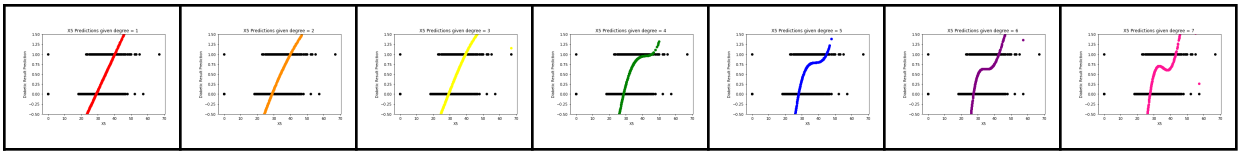
X[3]



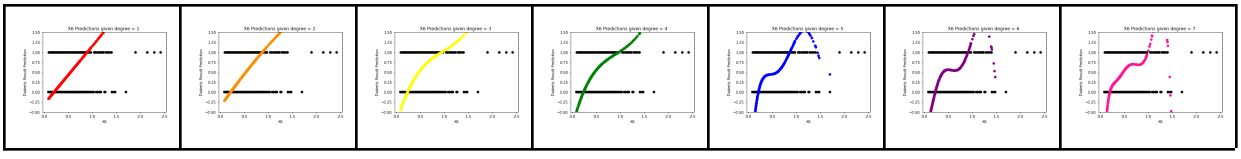
X[4]



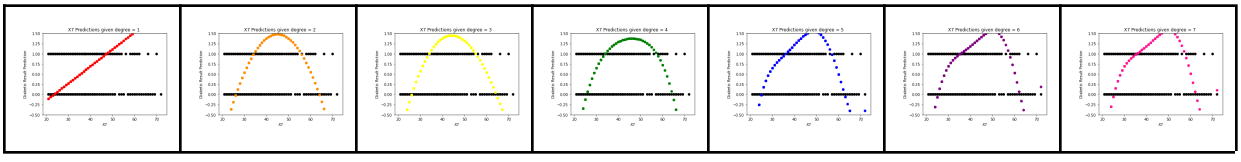
X[5]



X[6]



X[7]



The degrees range from 1-7 from left to right. I feel that the red and orange graphs which correspond to degrees 1 and 2 respectively, do not fit the data well enough; the fits are too general. The yellow graph which corresponds to degree 3 fits each data set more accurately, but the green and blue graphs, showing degrees 4 and 5 fit the datasets quite well. The green graphs corresponding to degree 4 have local minima and maxima in locations that mimic the data trends well. The blue graphs for degree 5 do as well, but for certain X features, such as X[4], they begin to hint at overfitting due to too many local minima and maxima. The purple and pink graphs for degrees 6 and 7 are much too overfit, as their curvature no longer demonstrates trends in the dataset, but rather begins to demonstrate an exact fit to the specific points generated in the train set. For these reasons, I feel that using a degree of 4 to generate predictions for the Pima dataset is the best option for all X features.