Report

To run, open the predictor.ipynb, run bokeh serve --show predictor.ipynb on terminal.

3 approaches have been attempted

1. Stack bars- For all 27 categorical attributes, we can have two categories i.e. whether a particular category has disease (cancer, heart, diabetes) or not.
   For ex: If a person has a cat, he can either have diabetes or not.

   Prediction: If one has a cat, he has a very low chance of being diabetic.

2. Clustering- KMeans has been used for clustering the dataset.
   Feature engineering can be applied here in sense that we can select which attribute to choose for clustering.

3. KNN classification- In this, we first find a correlation matrix. Then we take the correlation array for the a particular disease (i.e. cancer, etc.). We then have choice to select top N correlated attributes(that have correlation coefficients closer to 1 or -1). We can also choose the number of rows to train and test.
   For ex: I choose r=5 rows, the results were poor. As can be seen in Last graph.
   If I choose r=40 rows, the results are better. Almost all predictions on whole dataset matches to ground truth.
   We can also choose 2 attributes among N attributes to show visualization.

   Prediction: We can see that if Group_peanuts_other_nuts_seed__total_grams >12, one has a very high chance of cancer.