

Predicting heart disease in seniors using fruit and vegetable intake



May 13, 2017

Rushabh RAMBHIA

# Problem Statement

Can we predict chances of Heart Disease in seniors based on their daily fruit and vegetable intake?

# Hypothesis

Seniors who eat more fruits and vegetables on a daily basis have lesser chances of Heart Disease.

# Previous Studies

On initial research, I found two previous studies on this topic

* Study 2: [Published in 1996 in JAMA](http://jamanetwork.com/journals/jama/article-abstract/396216) 
  + This study concluded that there was an inverse association between **fiber** intake and **Myocardial Infarction** in **male health professionals**
* Study 1: [Published in 2002 in American Society of Clinical Nutrition](http://ajcn.nutrition.org/content/76/1/93)
  + This study concluded that there was an inverse association of fruit and **vegetable** intake with the risk of cardiovascular disease in **general** US population
  + Conducted on **time series** data for 9,608 adults and data was collected from **1971 to 1975**

These studies confirm that there is existing scientific basis for my hypothesis. At this point, I think it is important for me to mention that my analysis is different from the above two in the following ways:

* Focus on Seniors (65+)
* Cross Sectional Data
* Data Source: BRFSS
* Feature of interest is daily fruit and vegetable intake
* Include other variables that could impact either Fruit Intake or Heart Disease in seniors
* Outcome is Categorical ie Heart Disease = Y or N

# Data Set

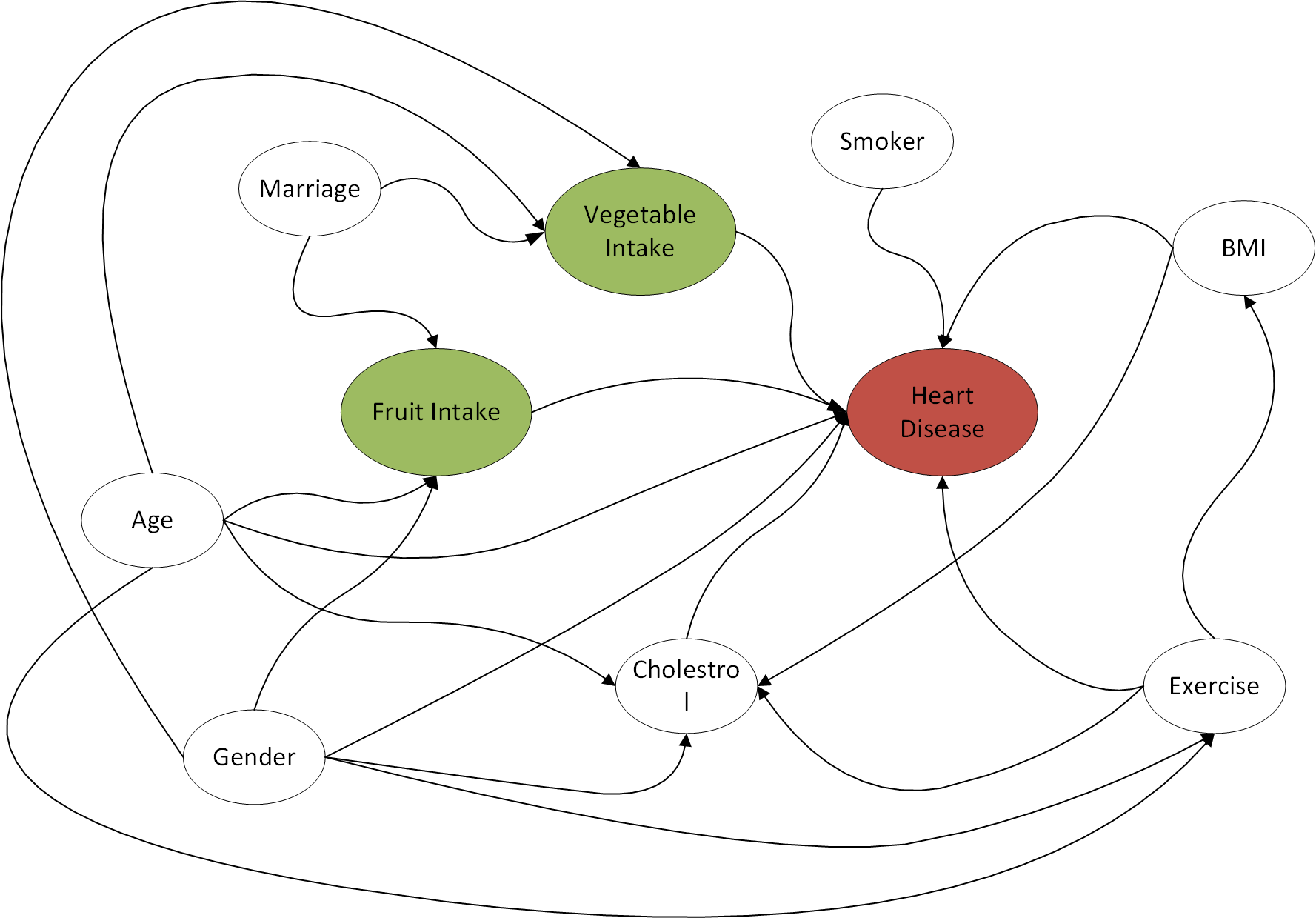
I am using 2015 BRFSS data set. This data is publicly available and can be downloaded from [CDC website](https://www.cdc.gov/brfss/).

Here are key highlights of [2015 data set](https://www.cdc.gov/brfss/annual_data/annual_2015.html)

* Cross Sectional data
* Over 441,000 people interviewed across 50 states in US
* Each interviewee answered over 120 mandatory and optional questions about their habits and chronic conditions
* There are over a total of 330 columns that include original responses to the interview questions along with calculated variables based on those responses
* Data feed available in TEXT as well as SAS format
* Data size is little less than 1 GB

# Identify Key Columns

There are 330 columns in original dataset. I created a web of causation to minimize the list of columns to the ones that are relevant to this study.



# Data Extraction

I followed below steps to prepare preliminary data set for my analysis.

1. Download SAS file from [CDC website](https://www.cdc.gov/brfss/annual_data/annual_2015.html)
2. Import SAS file using read\_sas function
3. Extract columns relevant to my analysis. See Appendix A for list of columns selected from original data set
4. Save this data into a CSV which can be used in later steps. This way I do not have to load the original dataset which is huge and can take a lot of memory

# Preliminary Data Preparation & Cleaning

In this phase I reviewed documentation of all fields I have selected for my analysis and create either calculated variables or dummy variables depending on how I thought I would use those fields in my analysis. See Appendix B for calculated and dummy variable list and logic.

Additionally I excluded records that met the following criteria:

* No response or NaN for response variable (\_MICHD)
* No response to age (\_AGEG5YR >= 14)
* Less than 65 years of age (AGEG5YR)
* Null values for cholesterol question (\_RFCHOL)
* Null values for exercise (EXERANY2)
* Null values for dailyFruit (FRUIT1)
* Null value for dailyVeggie (FVGREEN)

I was left with over 127K records out of original 441K records.

# Data Exploration

## Response Variable:

My model will have one response variable (HD). It is a binary response variable with two possible values: 0 = doesn’t have heart disease; 1 = has heart disease.

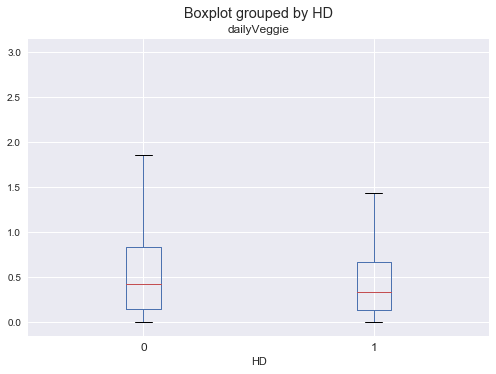
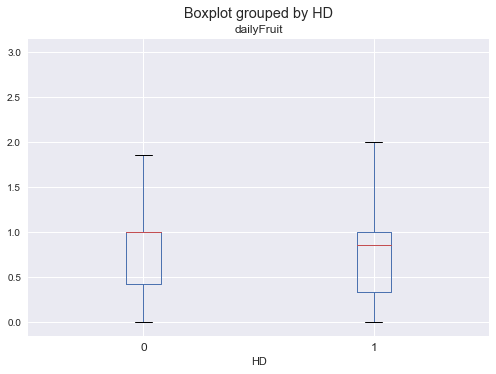
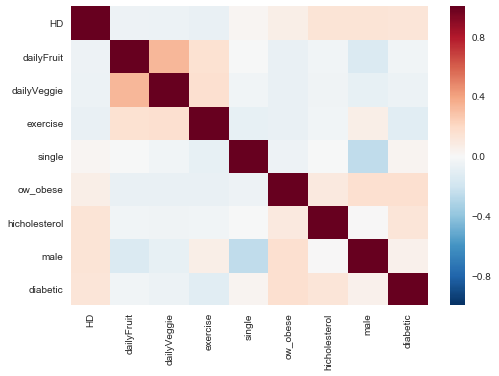
21,764 records out of 127,038 records have heart disease ie 17.13% have a positive result.

## 8.2 Features:

The two main features of my analysis are dailyFruit and dailyVeggie. I have calculated these features from original columns FRUIT1 and FVGREEN respectively. The functions used to calculate these features can be found in Appendix B.

These two features are continuous numbers and they range from 0 to 99. I have decided to exclude Nan values as well as outliers (values > 3) for both these features.

Below box plots show data ranges for dailyFruit and dailyVeggie features for HD=0 and HD=1 populations.



## 8.3 Correlation between variables:

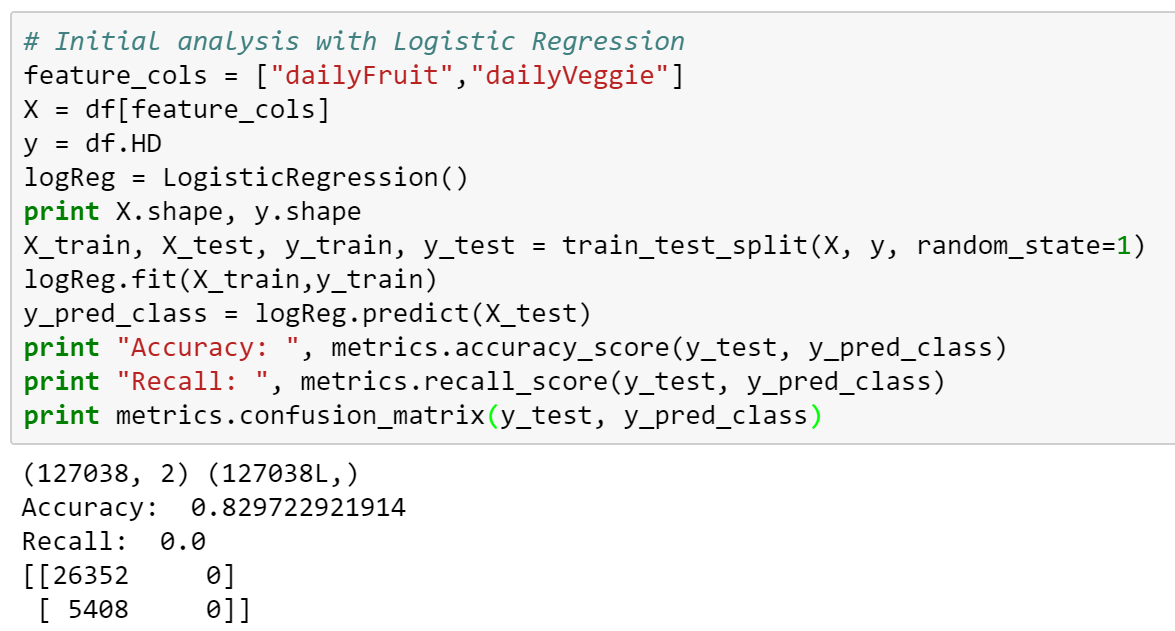
Observations from attached correlation heat map:

* + Columns dailyFruit and dailyVeggie are highly correlated.
  + Columns diabetic, male and hicholesterol are correlated with response variable HD
  + Columns ow\_obese is correlated with male and diabetic

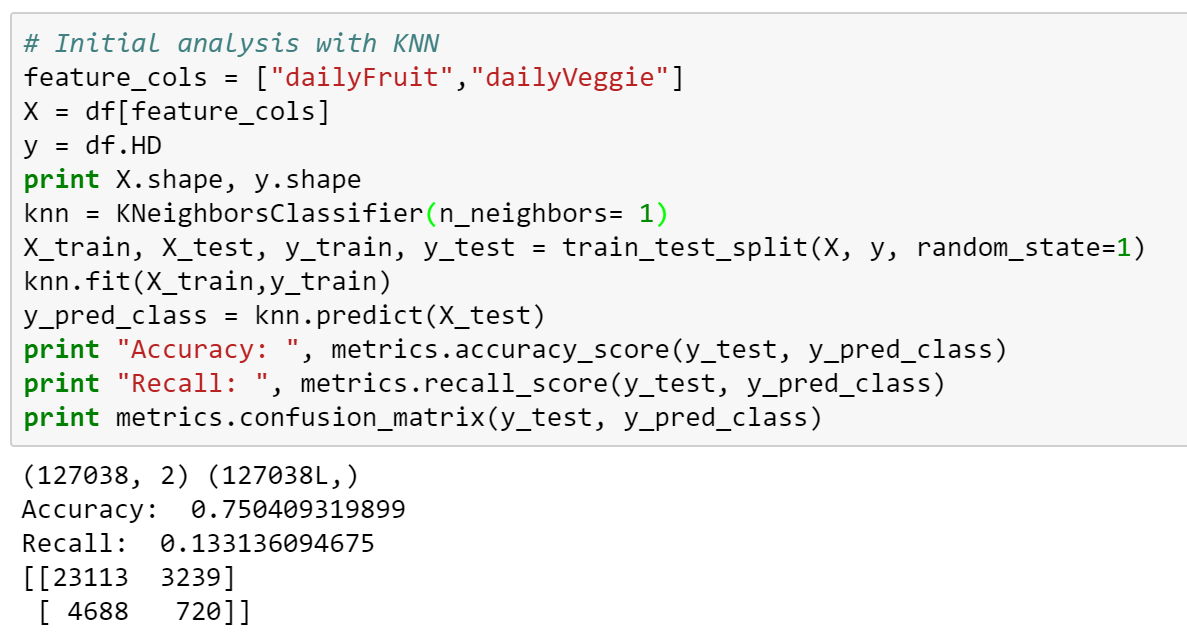
# Modelling Process

## Imbalanced data

On running Logistic Regression and using accuracy to validate the results, I observed that model’s accuracy was as good as Null accuracy (83%). I also checked Recall score of my initial model, and it was zero indicating that the model was unable to predict any HD = 1.



I used KNN on this data and observed that recall was the highest (13.3%) with accuracy (75%) lower than Null accuracy for K = 1.



Typically such problem arises due to imbalanced data. So I decided to apply following tactics:

### Under sampling

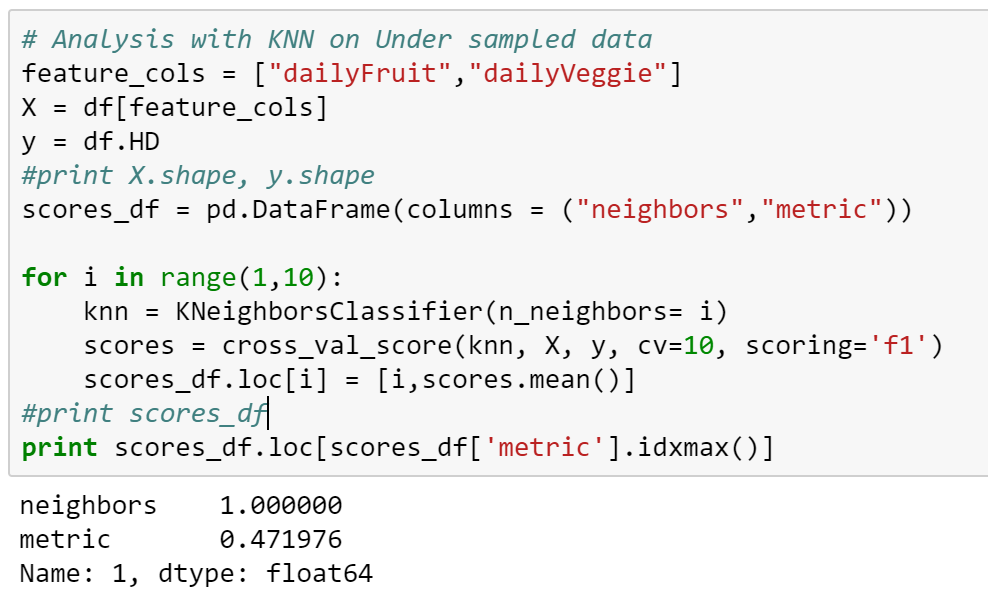
I decided to balance my data so that number of records with HD = 0 is the same as the number of records with HD = 1. This means the Null accuracy of this data set would be 50%.

### Use of Cost Sensitive Measure for Validation

I also decided to use f1 score as it is important for me to know what percentage of actual positives is the model able to predict correctly. F1 score is combination of precision and recall. The greater the value the better the model prediction.

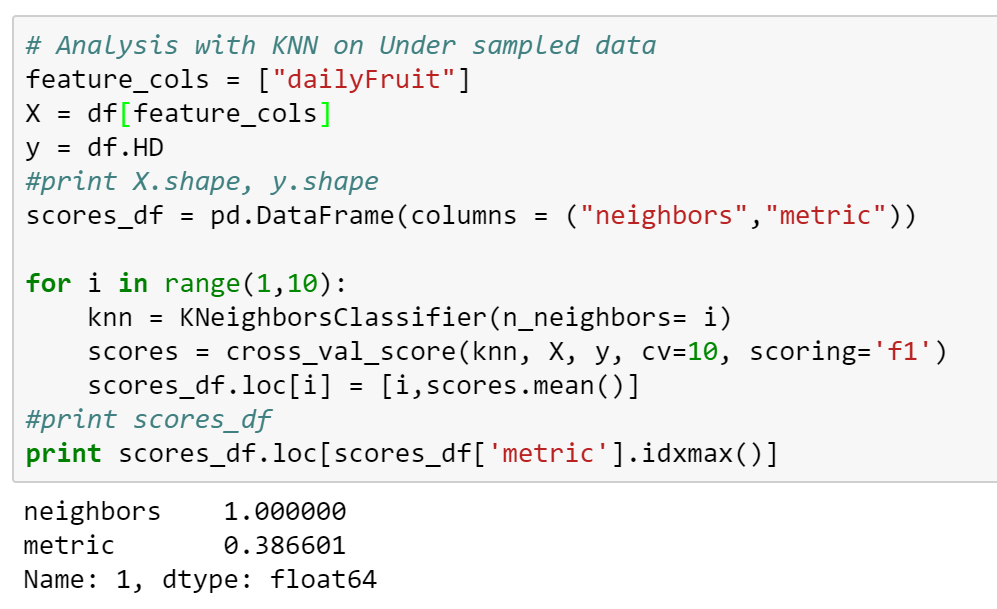
## Using KNN estimator with dailyFruit and dailyVeggie

I used cross validation on under sampled data with KNN estimator and got f1 score of 0.47 for K = 1.



## Using KNN estimator with dailyFruit only

Using the estimator with only one feature dailyFruit, I got a f1 score if 0.38. My conclusion was that it dailyFruit and dailyVeggie used together gave better prediction compared to using only dailyFruit.

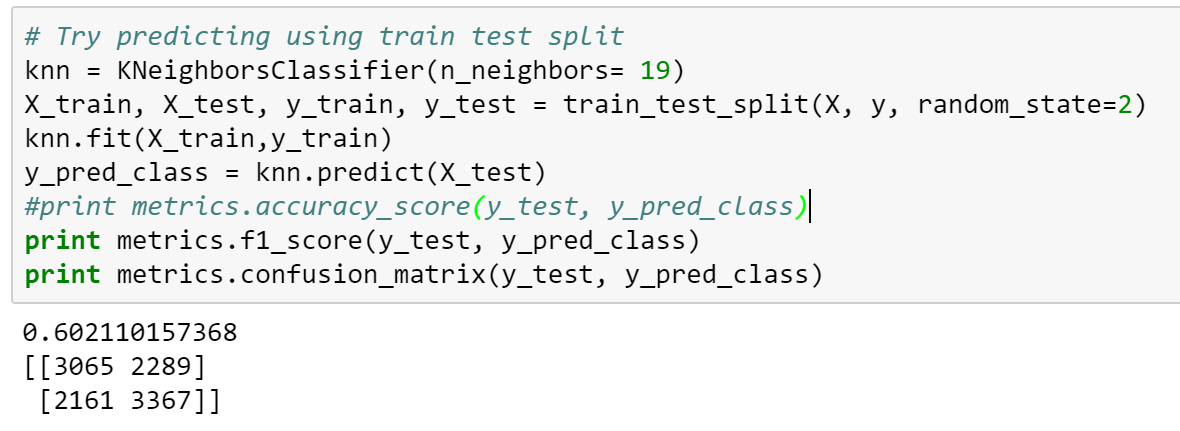


## Adding more features to KNN estimator and optimizing for better f1 score

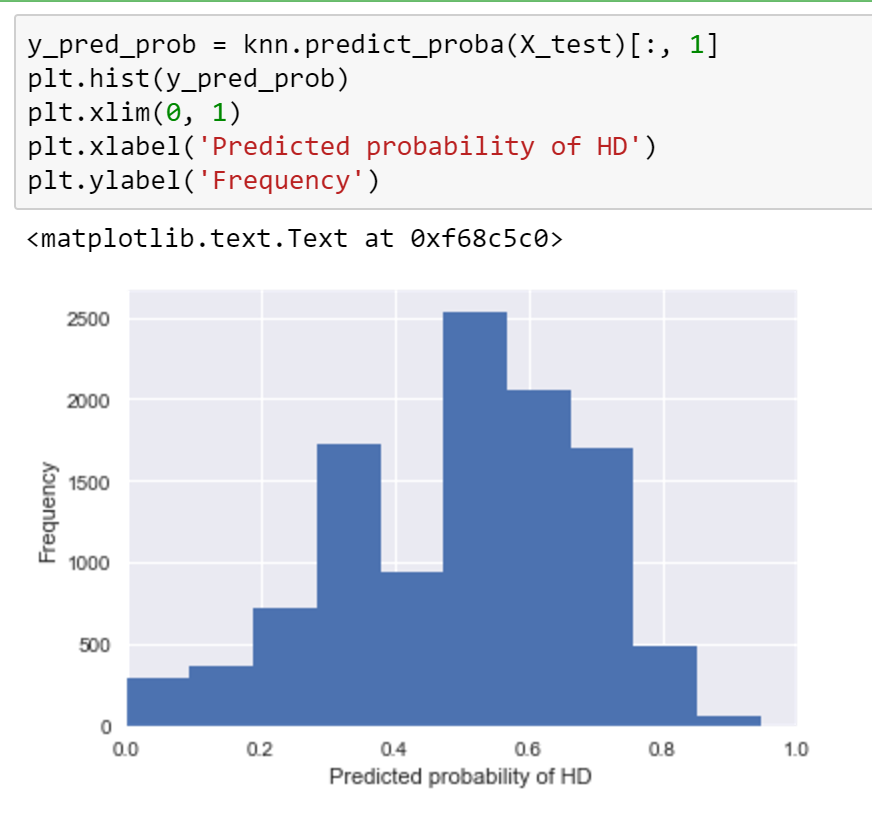
As I added more variables and tested for higher values of K, the f1 score got better. K = 19 gave f1 score of 0.559.



Using the same estimator with train test split method gives 0.60 f1 score.



The probability distribution of HD = 1 can be seen below.



# Conclusion

While dailyFruit and dailyVeggie aren’t necessarily the best predictors for heart disease in a person, we can still determine presence of heart disease in a person using these two features in combination of other features such as presence of diabetes, high cholesterol and exercise.

# Appendices A:

List of columns from original data source.

|  |  |
| --- | --- |
| **Column** | **Description** |
| \_MICHD |  |
| SEX |  |
| MARITAL |  |
| \_RFBMI5 |  |
| DIABETE3 |  |
| \_RFCHOL |  |
| EXERANY2 |  |
| FRUIT1 |  |
| FVGREEN |  |
| \_AGEG5YR |  |

# Appendices B:

List of calculated and dummy variables

|  |  |
| --- | --- |
| **Column** | **Description** |
| HD | This is response variable. All records with \_MICHD = 1 where set as 1 and all other values of \_MICHD were set to 0. |
| dailyFruit |  |
| dailyVeggie |  |
| single |  |
| diabetic |
| hicholesterol |
| exercise |
| male |
| ow\_obese |
| senior |