Jeanne Reppert
5/1/2020

IAF 604, Machine Learning and Predictive Analytics, Assignment 2 Report

Introduction

The purpose of this exercise is to explore the KNN Algorithm (K-nearest neighbors) and its appropriate use for particular datasets. This supervised learning algorithm utilizes classification in order to understand data. The method is faster and easier to use than other machine learning techniques but is best used with data sets that have certain types of characteristics. First, KNN algorithms work best with data that is not highly dimensional (curse of dimensionality) and it does not work well with features that are categorical. There is some discussion that extremely large databases can present some issues for KNN due to the computation cost involved with a very large database. Thus care should be taken to chose data with a limited amount of features and continuous variables when possible.

Rationale for Dataset Choice

For this study, three datasets were investigated for possible use with the KNN Algorithm. The three datasets were all medical datasets since I am most interested in this topic. Variables for each dataset were either demographic data in numeric form and various numeric health measures. Target variables in each set were measures of a selected patient outcome. For all three datasets NaN values were a concern since the KNN algorithm would require the removal of these values before the algorithm could be utilized.

The first set was the Pima Indians Diabetes set containing 8 numeric variables and a 9th binary target variable with 769 observations. While there were no 'NaN' values in this set, it did have missing values (measures of '0' that do not seem logical) that would need to be removed (imputation did not appear to be a viable option with this data.) There were 5 observations of '0' for glucose, 35 observations of '0' for blood pressure, 227 measures of skin fold thickness with '0' and 2 observations of '0' for BMI. It is likely that the skin fold thickness variable would need to eliminated since there are so many missing observations. The target variable, 'Outcome' was a bit imbalanced.

The second dataset explored was the Heart Disease UCI database. This set had 304 observations with 13 variables with a 14th binary target variable. While this set was fairly small it had few problematic observations (a little more than 20 observations with '0' values that do not make sense.) Its target variable was fairly balanced as well. Finally, the third dataset explored was the Framingham Heart Study dataset. This dataset contained many more observations (4240) with 15 numeric variables. The target variable, which was a measure of ten-year heart disease risk outcome, was fairly imbalanced with 3596 observations of '0' and 644 observations of '1'. The dataset also has more 'NA' values than the other sets (although a lot more observations overall). The education variable had 105 missing values, blood glucose had 388 missing variables, 29

missing values for cigarettes per day, bpmeds had 53, total cholesterol had 50 and BMI had 19 and heart rate had one missing value.

In order to select the best set several considerations regarding the data were made. All three sets had a binary target with varying levels of imbalance. The diabetes and Heart Disease UCI databases had limited observations. Although a small amount of observations is not necessarily a factor that is a problem with KNN, after removing the observations with missing observations, I felt that a larger dataset would be preferable for my use. Coupled with the concern that some observations might have to be removed due to measures of '0' that seemed to indicate a missing measure and due to the fact that imputation may not be optimal in the cases of these two sets, the Framingham database was given consideration over the other two datasets.

While the Framingham database did have as many missing values as the other two sets, it was decided that since the amount of observations was fairly large, the observations could be deleted from the set using one of two methods. First, for education, the decision was made to delete this variable. The first reason for deletion was that this variable seemed to be a bit more biased since it was not a medical measurement and was assigned a rating which could result in some inconsistency. This issue, along with presence of 105 missing ratings, led to the decision to omit this variable. For the other variables, the dropna function was used to remove the observations with NaN values in the dataset. This resulted in a total of 3751 observations with which to perform the KNN algorithm.

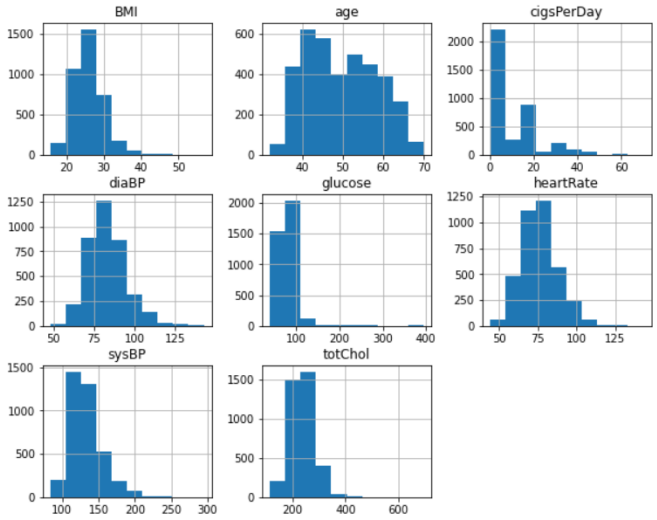
Framingham Database – Data Exploration and Cleaning

As stated above, the Framingham dataset is a two dimensional set with a total of 4240 observations, 6 total columns which were all numeric (one being a binary target variable) and 67840 total elements. Feature names included were gender, age, education, smoker, number of cigarettes, BP medications, stroke, hypertension, diabetes, cholesterol, systolic blood pressure, diastolic blood pressure, BMI, heart rate, and glucose. The target variable was the risk of developing heart disease within ten years. Once the education column was removed and the NaN values were removed from the dataset, there were 3751 total observations, 15 columns and 56265 total elements.

Framingham Database – Statistic Analysis and Visualizations

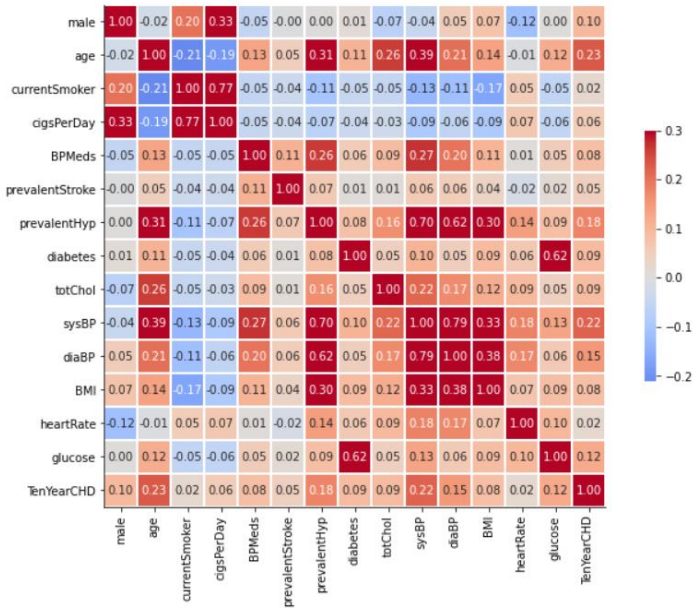
Since the KNN algorithm is non-parametric, the distribution of the data is not quite as much of a concern. Nonetheless some exploration was done to become familiar with the statistical measures of the data. For the gender (female=0, male=1) variable, there were 2081 females and 1670 males. The mean participant age given the observations was 49.57 and the median age was 49 indicating a relatively normal distribution of age. Overall, participants smoked an average of 9 cigarettes a day, however, a little more than half of the participants (1919) were non-smokers. A relatively small amount of the participants (114) took blood pressure medication. For the prevalent stroke category there were only 21 participants that indicated a stroke. 1170 of the participants reported hypertension and 2581 of the participants did not report hypertension. Only 102 of the participants reported diabetes. Other distributions are visualized below:

Select variable distributions:



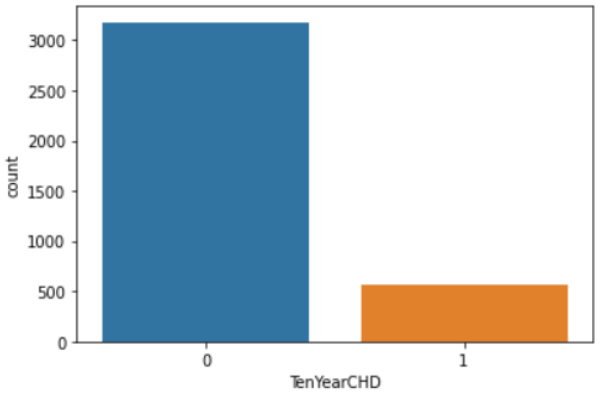
The mean BMI of the participants was 25.81 and the mean heart rate was 75.70. The mean diastolic blood pressure was 82.94 and the mean systolic blood pressure was 132.37. The mean glucose measure was 81.88 and the mean cholesterol was 236.93.

There appeared to be some redundancy in some of the measures that were included in the dataset. For example, the glucose measure and the indication of diabetes are possibly redundant. Likewise, the presence of hypertension and the blood pressure measurements also appeared to be redundant. In the same way, the smoker/non-smoker category and the number of cigarettes smoked appeared to be redundant. It is possible that the elimination of the hypertension, current smoker, and diabetes variable could be explored. To further explore this possibility, a correlation plot was investigated:



Understandably, there was indeed a strong correlation between hypertension and blood pressure, as well as a strong correlation between diabetes and glucose. There also was predictably a strong correlation between the number of cigarettes smoked per day and the currentSmoker variable. For this reason, the variables for diabetes, hypertension and current smoker will be eliminated since the other measures give more specific information about the same topic.

For the target variable there were 3179 participants that were not at risk for heart disease and 572 participants that were at risk for heart disease. This represents a fairly imbalanced target variable. In researching this issue, it was not completely clear whether or not this would present an issue with the KNN algorithm. For this reason, for the purposes of exploration, the model will be run without adjusting for the imbalance of the target variable and then it will be run again after adjusting for the imbalance of the target variable.

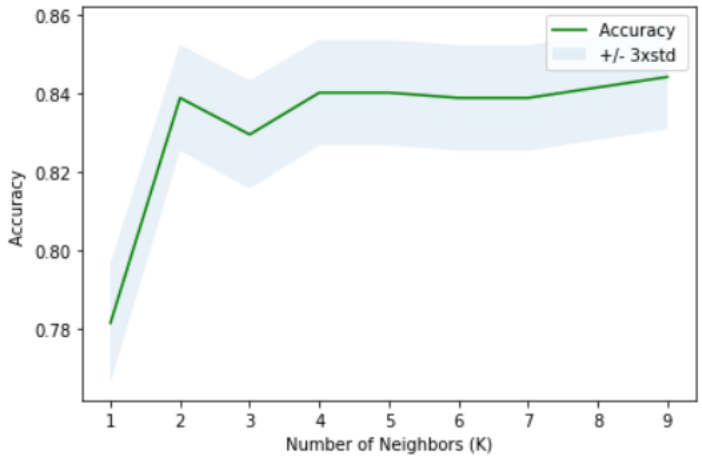


Framingham Database – The KNN Algorithm

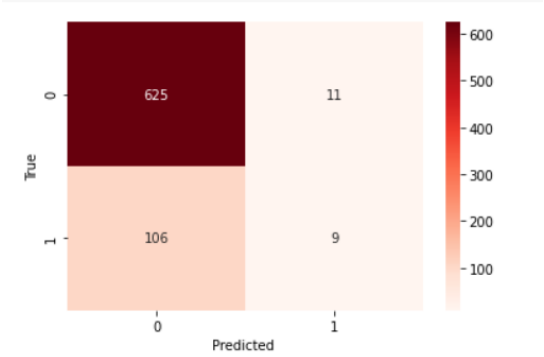
Once the data was cleaned and the columns mentioned above were dropped, the remaining features were converted into a numpy array named X and the target was converted to a numpy array named y. Once this process was completed a standard scaler function was applied to X to assign data a zero mean in order to enhance KNN distance calculations. The arrays were then split into a training and test sets (80/20). The final training set has 3000 observations and the final test set has 751 observations. The KNeighborsClassifier was then applied to the training dataset with a K value of 4 assigned for nearest neighbors. Following this process, a predicted y (yhat) was created using the resulting model and the X test set (X_test).

Framingham Database – Quantitative/Qualitative Measures

The metrics package was imported from SKLearn and used to compute accuracy measures for the training and test sets. The accuracy measure for the training set was 0.8677 and the accuracy measure for the test set was 0.8402. A loop was applied to the kmeans algorithm to attempt to find which K value rendered the best accuracy values. The resulting accuracy was .8442 when K = 9.



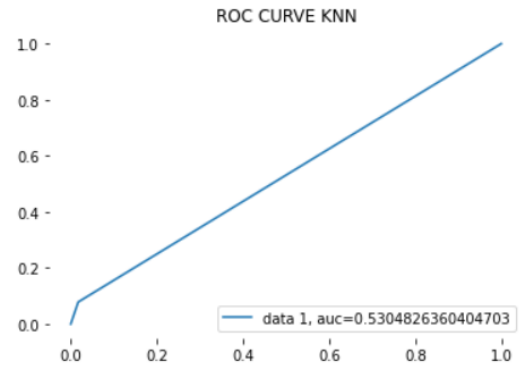
A confusion matrix was also plotted in order to further investigate predicted values. Precision, sensitivity, specificity, F-1 and the Youden's Index were also calculated with results as follows:



	precision	recall	f1-score	support
0	0.85	0.98	0.91	636
1	0.45	0.08	0.13	115
accuracy			0.84	751
macro avg	0.65	0.53	0.52	751
weighted avg	0.79	0.84	0.79	751

```
Accuracy :  0.844207723035952
Sensitivity :  0.9827044025157232
Specificity :  0.0782608695652174
Youden Index :  0.0609652720809406
```

A roc curve was plotted and an AUC score was also computed:




```
The score for the AUC ROC Curve is: 53.0%
```

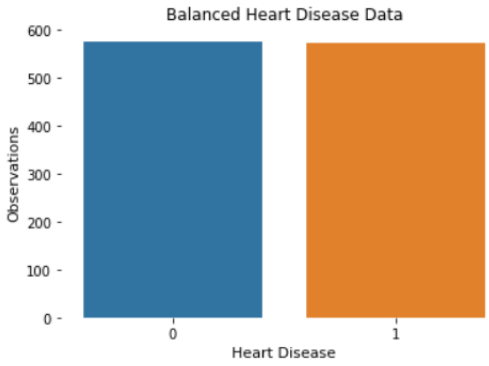
Jeanne Reppert
5/1/2020

Framingham Database – Analysis

The precision score for the model was 0.84 with a sensitivity score of .98 and a specificity score of .08. While the accuracy measures were reasonable and the test and train set accuracies were relatively close, the confusion matrix and sensitivity/specificity scores along with the Youden Index tell another story. Most of the accurate predictions that were made involved predicting those who were not at a 10 year risk for heart disease but only 9 correct predictions were made for those who actually might be at a 10 year risk for heart disease. There were also 106 false positives predicted (predicted as not at risk when the participant was at risk). Overall, given these measures the resulting Youden's Index was fairly low (0.061). Because of these measures, some effort was made improve the model by attempting to resample the dataset in order to create a target that was more balanced. This effort will be discussed in the next section.

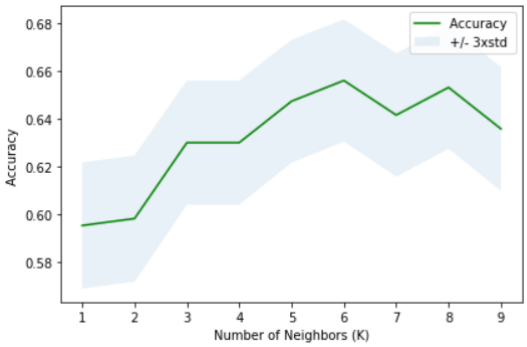
Framingham Database – Creating a Balanced Dataset and Results

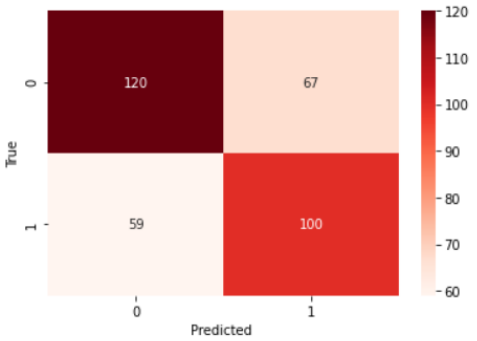
A sampling technique was used to create a balanced target in the dataset with 576 observations in each class. This resulted in a smaller set of data (df_norm) made up of 1152 observations.



The same process as above was performed including assigning the X_norm and y _norm groups to numpy arrays and applying the standard scaler for data standardization. The data was then split into train/test sets at a .8/.2 ratio resulting in 921 observations in the training set and 231 observations in the test set. The KNeighborsClassifier was then applied to the training dataset with a K value of 4 assigned for nearest neighbors. Following this process, a predicted y (yhat_norm) was created using the resulting model and the x set (nX_test).

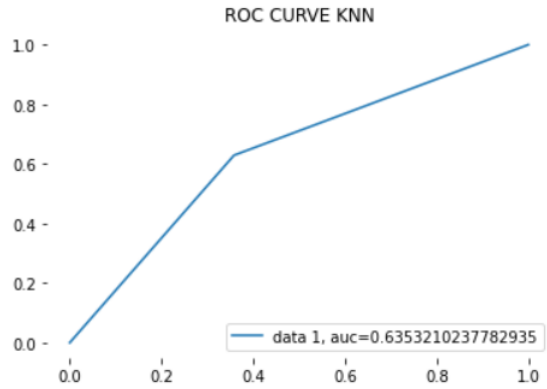
The accuracy measure for the training set was 0.7357 and the accuracy measure for the test set was 0.6301. A loop was applied to the kmeans algorithm to attempt to find which K value rendered the best accuracy values. The resulting accuracy was .6561 when K = 6. The same additional quantitative measures were taken with the following results:





	precision	recall	f1-score	support
0	0.67	0.64	0.66	187
1	0.60	0.63	0.61	159
accuracy			0.64	346
macro avg	0.63	0.64	0.63	346
weighted avg	0.64	0.64	0.64	346

```
Accuracy :  0.6358381502890174
Sensitivity :  0.6417112299465241
Specificity :  0.6289308176100629
Youden Index :  0.27064204755658694
```




```
The score for the AUC ROC Curve is: 63.5%
```

Framingham Database – Conclusions

The measures on the second dataset gave mixed results. The accuracy scores did decrease and the scores for training and testing were further apart pointing to some overfitting of the model. The AUC curve for the second model however is higher than the model with the original data with a score of 0.635 as opposed to a score of 0.53 in the original data. More observations with 10 year risk for heart disease were accurately identified, however this did come at a cost of more false negatives. This also resulted in a lower sensitivity score (0.64) but a much higher specificity (0.63) score. The score for precision also decreased to (0.67). Overall this change resulted in an improved Youden's Index score (although still fairly low since the ideal score is 1) of 0.27. With the exception of a possible overfitting problem it would appear that using the sampled data that was balanced did improve the model and results.