**Classification Using Nearest Neighbors**

**Assignment 1**

Author: Shreeniket Subhash Deshmukh

ALY-6020 Predictive Analytics

Winter 2020 Quarter

Date: 02-23-2020

Submitted to – Professor Roseanna Hopper





#### Abstract

#### Classification using nearest neighbor is illustrated by the k-NN algorithm, in which k is a variable that denotes the number of neighbors that will be used for classification. In this assignment we are going to diagnose breast cancer with k-NN algorithm. This machine learning model will automate the identification cancerous cell to provide results that might be beneficial for heath systems. This analysis has been performed in five steps starting from data collection to improving models.

#### Analysis

#### Part A

#### Step 1: Data Collection

#### We will be using data which has been published by University of Wisconsin. This data includes measurements of different images of fine needle of an aspirate of a breast mass. Each observation in the data represents the characteristics of cell nuclei from the images. Structure of the data is 569 observations of the cancer biopsies along with 32 features. One of the features is “diagnosis” which can be benign or malignant. We are going to make predictions on this feature.

#### Step 2: Data Exploration and Preparation

#### In this step we are going to take an overview of the data and try to understand what changes needs to be done in the structure of the data to build a model. Other than id and diagnosis variables, remaining 30 variables are is num format. Including identifier column in the model may cause overfitting, as it is an identifier, we can remove the id column from the data. Amongst 569 observations 357 masses are benign and 212 are malignant, which tells 62.7% masses are benign and 37.3 masses are malignant. If we look at the remaining 30 features there are only 10 characteristics with three different measurements which are mean, standard error and worst.

#### If we summarize few variables of the data, we can see that ranges of each variables varies vastly. Whenever Euclidean distance has been calculated between neighbors it depends on the measurement scale. To avoid any causes to the accuracy of the classification model we need to normalize the data.

Figure 1: Summary of the data before normalization

#### 

Figure 2: Normalized data

Now all the variables range from 0 to 1, which we are going to use for building a classification model.

In the next part of the step 2, we have created two different data frames training and testing data. We have excluded target variable “diagnosis”. We will be using these data frames to train a model and test the accuracy of the model which will help us to figure out how dependent we can be on predictions of diagnosed patients.

#### Step 3: Training a model on the data

#### In this step we are going to train a model using training data created in the previous step which will be used to make predictions on the test data. We have stored all the predicted values by knn() function in the vector. Here we need to provide k parameter to the function and selecting a value of k is important. Usually it is considered as an odd number to avoid the tie between the classes if the number of class is even. In this situation we have taken nearest odd number of the square root of total number of observations. While classification of unlabeled items this model will consider 21 number of nearest neighbors.

#### As mentioned earlier, results will be stored in the vector to compare it with the actual results to find the accuracy of the predictions made by the model when k=21.

#### 

Figure 3: Predicted vector of the target variable

#### Step 4: Evaluating model performance

#### In this step we are going to check the predictions with actual observations. To perform this we have two vectors, first one contains predicted results by the model and second one contains the actual results. We have created cross table to compare these two columns as shown in the screenshot below.

#### 

Figure : Cross table 1

As per our model we have predicted correctly that 77 times mass was benign out of 100 when compared it to the actual results. Also, our classifier correctly predicted that mass was malignant 21 times when compared it to the actual results. But our model says that 2 of the masses are malignant which are benign in the actual results. These kind of errors by the model can be very dangerous as patient can think of not having cancer. These kinds of errors can affect the health systems.

In this model 2 out of 100 observations were incorrect, we have achieved 98 percent of accuracy.

Now, we will try different number of k to understand what difference it makes if we consider less or a greater number of neighbors while using nearest neighbor method.

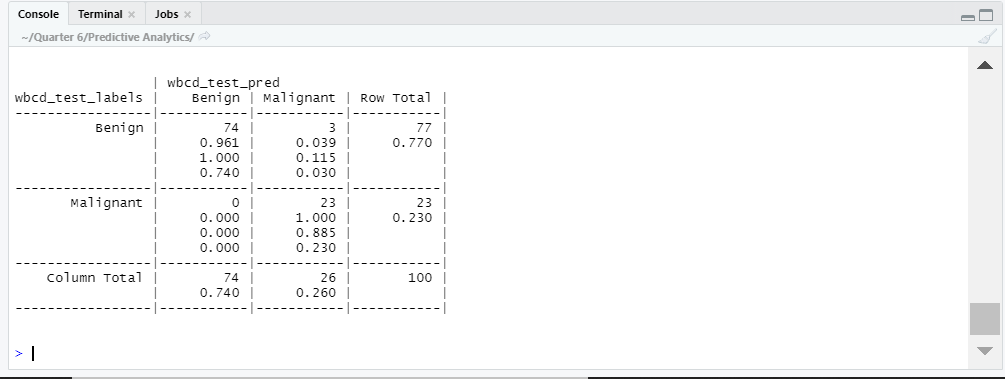


Figure :Cross table where k=5

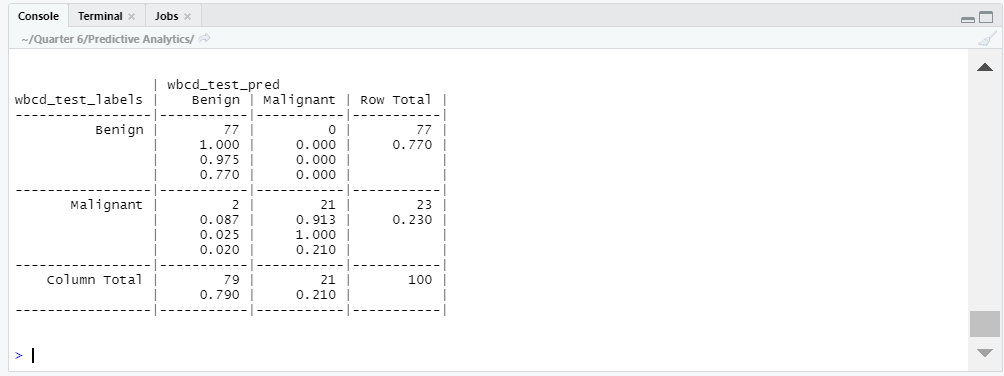


Figure :Cross table where k=15

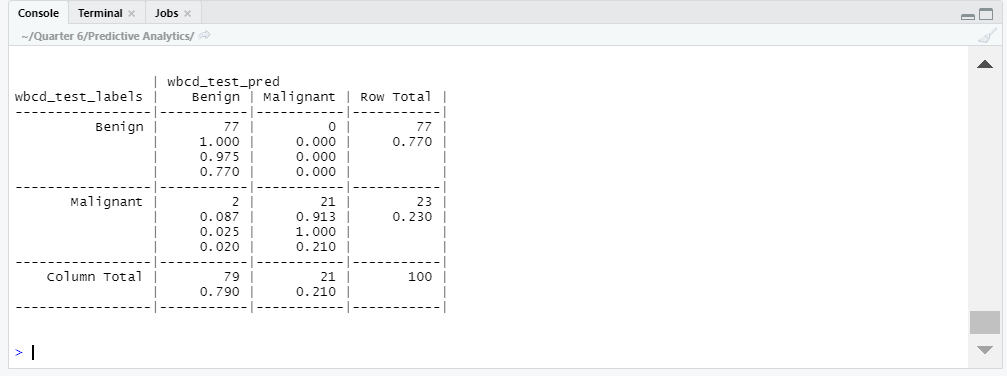


Figure :Cross table where k=27

As we can see in figure 6 and 7, we are getting results same as k=21, which 98 percent accuracy. While in figure 5 accuracy got reduced to 97 percent because model incorrectly predicted that 3 of the masses are benign instead of malignant.

Step 5: Improving model performance

In final step we will allow outliers to be weighted more heavily in case of distance calculations which can be achieved by z-score standardization. And we will compare it to the best model we used in the previous step.

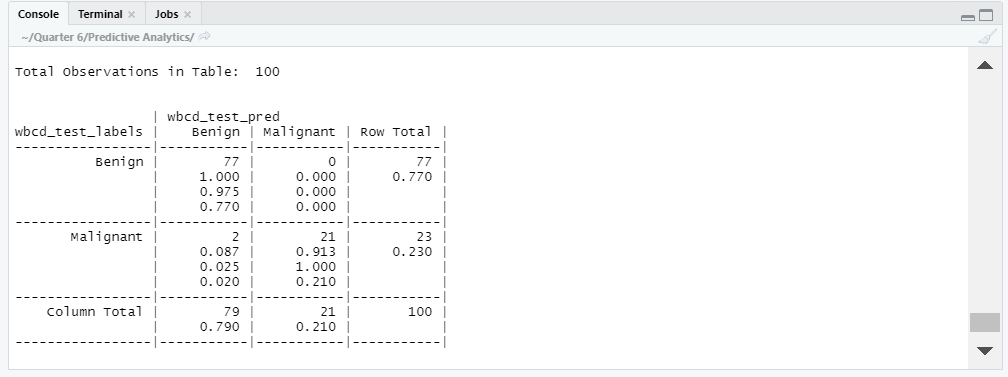


Figure :Cross table using z-score standardization

As we can see in figure 8, there is no change in the accuracy of the classifier we build using z-score standardization.

**Part B**

Step 1: Data collection

This dataset contains different measurements of abalone and we are going to use those to predict the age of abalone. This contains nine different measurements of abalones. One of the measurements is number of rings that I have been categorized in three different parts named as young, adult and old. And in this problem age will the target variable. This data has been collected from UCI machine learning repository.

#### Step 2: Data Exploration and Preparation

#### In this step I have categorized number of rings into 3 parts:

#### Number of rings: 1 to 5 => Young

#### Number of rings: 6 to 13 => Adult

#### Number of rings: 14 to 30 => Old

#### This categorization tells us that we have 3498 adult, 490 old and 189 young abalones. That makes approximately 83 percent of adult, 12 percent old and 5 percent of young abalones. If we summarize the data, range varies vastly. So, we need to normalize it before we feed it to the model.

#### 

Figure : Summary of the data before normalization

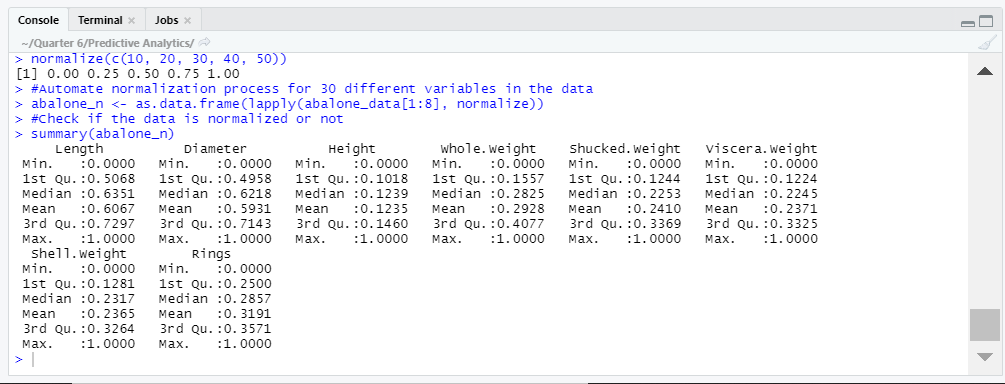


Figure : Normalized data

As we have normalized data ready to feed it to the model, in the next part of the step 2, we have created two different data frames training and testing data. We have excluded target variable “age”. We will be using these data frames to train a model and test the accuracy of the model which will help us to figure out the age of abalones using nine different measurements.

#### Step 3: Training a model on the data

#### In this step we are going to train a model using training data created in the previous step which will be used to make predictions on the test data. We have stored all the predicted values by knn() function in the vector.Here we need to provide k parameter to the function and selecting a value of k is important.

#### Similarly, from the Part A section of this assignment I have decided to take the nearest odd number of the square root of total number of observations which is 65.

#### As mentioned earlier, results will be stored in the vector to compare it with the actual results to find the accuracy of the predictions made by the model when k=65

#### 

Figure : Predicted vector

#### Step 4: Evaluating model performance

In this step we are comparing the results of the model with the actual results and will try to find out how accurately we have recognized the age of abalone using nearest neighbor method.

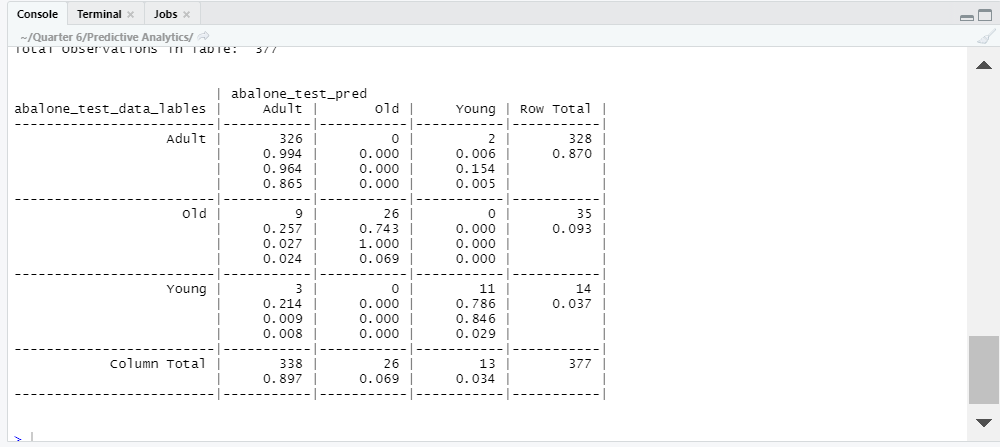


Figure : Cross Table

In figure 12, we can see that 326 adult abalone were predicted correctly by the model out of 328, while 26 of the old abalone were predicted correctly out of 35 and lastly 11 young abalone were predicted correctly out of 14. In total 14 abalone were predicted incorrectly out of 377. This means our model has predicted age of abalone with the accuracy of 96.28 percent.

Now, we will try different number of k to understand what difference it makes if we consider less or a greater number of neighbors while using nearest neighbor method.

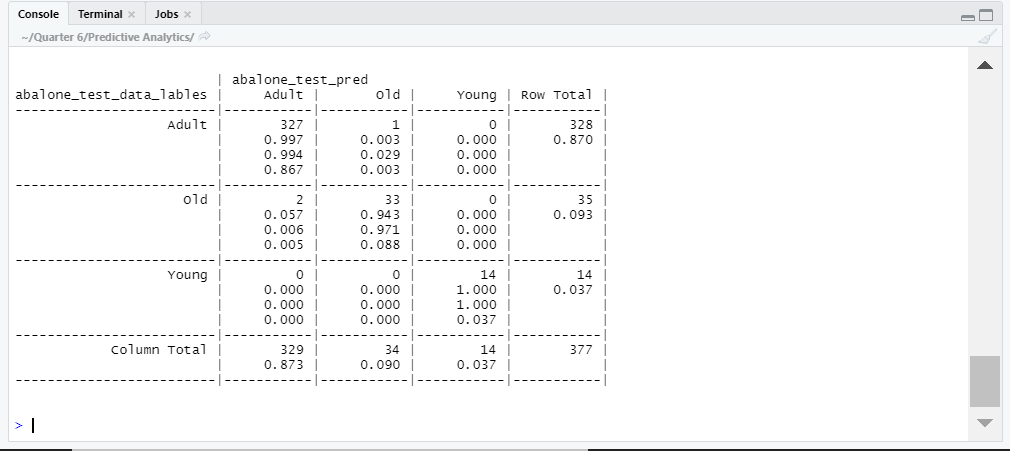


Figure : Cross table with k=3

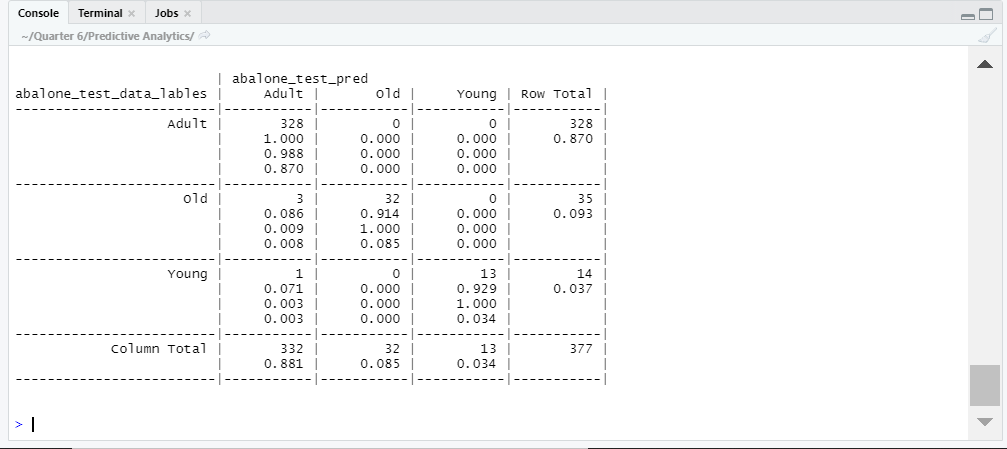


Figure : Cross table with k=15

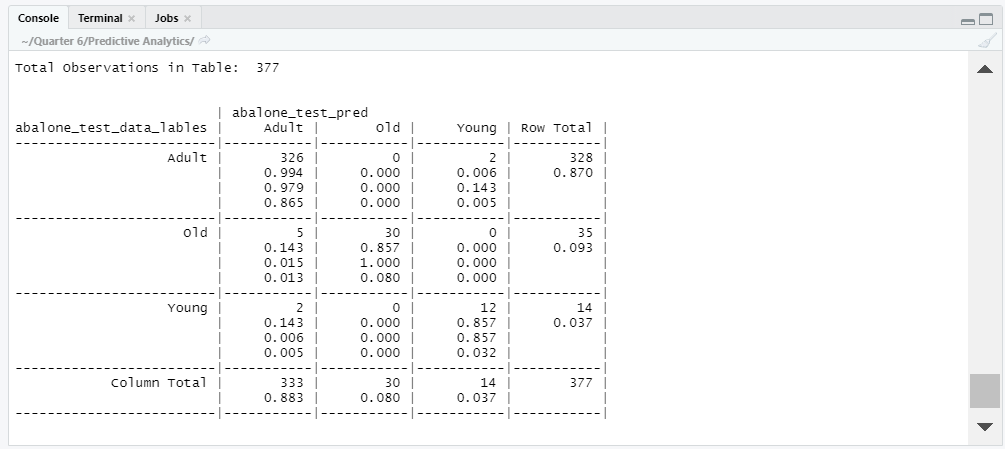


Figure : Cross table with k=27

Among figure 13, 14 and 15 we can see that if we reduce the number of neighbors while using knn we got only 3 incorrect predictions. With the value of k=3, this model has achieved 99.20 percent of accuracy for the prediction of the age of abalone.

Step 5: Improving model performance

In final step we will allow outliers to be weighted more heavily in case of distance calculations which can be achieved by z-score standardization. And we will compare it to the best model we used in the previous step.

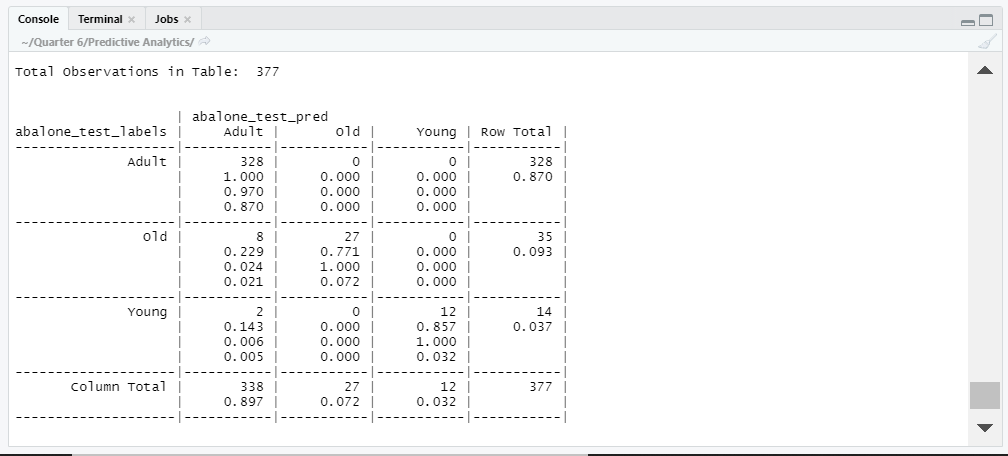


Figure 16: Cross table using z-score standardization

As we can see in figure 16, accuracy of the model got reduced to the 97 percent. Unfortunately, there is a slight decline in the accuracy of the model if we z-score standardization method.

**Conclusion**

In this assignment we studied that k-NN is a lazy algorithm because it does not train model, it just stores training data. This makes it one of the simplest algorithms in machine learnings. Still it is capable of solving complex problems like cancer cell prediction as seen in the Part A. In the breast cancer data, we have achieved 98 percent of accuracy to make predictions on cancer masses with the help of 30 different features. And in the abalone data we have predicted age of abalone with the accuracy of 99 percent. In this example, we can conclude that age of the abalone varies to the number of rings in the shell. Weight and height of the abalone varies accordingly to the age of the abalones.

**References**

#### UCI Machine Learning Repository: Breast Cancer Data Set. (2020). Archive.ics.uci.edu. Retrieved 23 February 2020, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

#### UCI Machine Learning Repository: Abalone Data Set. (2020). Archive.ics.uci.edu. Retrieved 23 February 2020, from https://archive.ics.uci.edu/ml/datasets/Abalone

#### (2020). Ia801600.us.archive.org. Retrieved 23 February 2020, from https://ia801600.us.archive.org/11/items/Machine.Learning.with.R.2nd.Edition.2015.Lantz/Machine.Learning.with.R.2nd.Edition.2015.Lantz.pdf

#### K-nearest Neighbors Algorithm with Examples in R (Simply Explained knn). (2018). Medium. Retrieved 23 February 2020, from https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c