**Data Preparation**

1. Created 1000 data points and assigned it to either an x- or y-coordinate vector.
2. Created a data frame combine the coordinates and initialize a quadrant column.
3. Assigned quadrant letter values relative to the data point’s position.
4. Created separate 200 noise points whose assigned quadrant were any (thru sample function) but its correct one.
5. Added noise data into the 1000-point dataset.
6. Created 100 data points, serving as the test data. Added this as well

To the 1,200 row dataset, now at 1,300.

1. Normalized the full 1,300 dataset.
2. Created train and test models from the normalized dataset.
3. Created label vectors from the train and test models.
4. Performed knn prediction using the test dataset (normalized).
   1. Started with k = 35 as k value should be the square root of the number of dataset rows.
5. Created cross-tabulation.
6. Performed knn predictions with k = 15, 25, 30, 40, 70.
7. Used scale function instead of normalize function. Predicted at k = 35, 10 & 70.

> str(dMod1300)

'data.frame': 1300 obs. of 3 variables:

$ x : num 0.9995 0.6126 0.7016 0.5409 0.0215 ...

$ y : num 0.642 0.439 0.305 0.405 0.978 ...

$ quadrant: Factor w/ 4 levels "A","B","C","D": 2 4 4 4 1 4 4 3 2 3 ...

> # confirm that normalization worked

> summary(ndMod1300$x)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2447 0.4947 0.4984 0.7449 1.0000

> summary(ndMod1300$y)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2631 0.4997 0.5095 0.7714 1.0000

> CrossTable(x = ktestl, y = kpred1, prop.chisq = FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| kpred1

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 20 | 0 | 0 | 0 | 20 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.200 |

| 1.000 | 0.000 | 0.000 | 0.000 | |

| 0.200 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 31 | 0 | 1 | 32 |

| 0.000 | 0.969 | 0.000 | 0.031 | 0.320 |

| 0.000 | 1.000 | 0.000 | 0.053 | |

| 0.000 | 0.310 | 0.000 | 0.010 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 0 | 0 | 30 | 0 | 30 |

| 0.000 | 0.000 | 1.000 | 0.000 | 0.300 |

| 0.000 | 0.000 | 1.000 | 0.000 | |

| 0.000 | 0.000 | 0.300 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 0 | 18 | 18 |

| 0.000 | 0.000 | 0.000 | 1.000 | 0.180 |

| 0.000 | 0.000 | 0.000 | 0.947 | |

| 0.000 | 0.000 | 0.000 | 0.180 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 20 | 31 | 30 | 19 | 100 |

| 0.200 | 0.310 | 0.300 | 0.190 | |

-------------|-----------|-----------|-----------|-----------|-----------|

**Per the results k = 35 incurred 1 error that coordinates on quadrant B was predicted lying on quadrant D.**

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| kpred2

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 24 | 0 | 1 | 0 | 25 |

| 0.960 | 0.000 | 0.040 | 0.000 | 0.250 |

| 0.960 | 0.000 | 0.036 | 0.000 | |

| 0.240 | 0.000 | 0.010 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 28 | 0 | 0 | 28 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.280 |

| 0.000 | 0.966 | 0.000 | 0.000 | |

| 0.000 | 0.280 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 1 | 0 | 26 | 1 | 28 |

| 0.036 | 0.000 | 0.929 | 0.036 | 0.280 |

| 0.040 | 0.000 | 0.929 | 0.056 | |

| 0.010 | 0.000 | 0.260 | 0.010 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 1 | 1 | 17 | 19 |

| 0.000 | 0.053 | 0.053 | 0.895 | 0.190 |

| 0.000 | 0.034 | 0.036 | 0.944 | |

| 0.000 | 0.010 | 0.010 | 0.170 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 25 | 29 | 28 | 18 | 100 |

| 0.250 | 0.290 | 0.280 | 0.180 | |

-------------|-----------|-----------|-----------|-----------|-----------|

**Per the results k = 15 incurred 5 errors.**

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| kpred3

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 25 | 0 | 0 | 0 | 25 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.250 |

| 0.962 | 0.000 | 0.000 | 0.000 | |

| 0.250 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 28 | 0 | 0 | 28 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.280 |

| 0.000 | 0.966 | 0.000 | 0.000 | |

| 0.000 | 0.280 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 1 | 0 | 26 | 1 | 28 |

| 0.036 | 0.000 | 0.929 | 0.036 | 0.280 |

| 0.038 | 0.000 | 0.929 | 0.059 | |

| 0.010 | 0.000 | 0.260 | 0.010 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 1 | 2 | 16 | 19 |

| 0.000 | 0.053 | 0.105 | 0.842 | 0.190 |

| 0.000 | 0.034 | 0.071 | 0.941 | |

| 0.000 | 0.010 | 0.020 | 0.160 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 26 | 29 | 28 | 17 | 100 |

| 0.260 | 0.290 | 0.280 | 0.170 | |

-------------|-----------|-----------|-----------|-----------|-----------|

**Per the results k = 25 incurred 5 errors.**

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| kpred4

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 25 | 0 | 0 | 0 | 25 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.250 |

| 0.962 | 0.000 | 0.000 | 0.000 | |

| 0.250 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 28 | 0 | 0 | 28 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.280 |

| 0.000 | 1.000 | 0.000 | 0.000 | |

| 0.000 | 0.280 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 1 | 0 | 27 | 0 | 28 |

| 0.036 | 0.000 | 0.964 | 0.000 | 0.280 |

| 0.038 | 0.000 | 0.931 | 0.000 | |

| 0.010 | 0.000 | 0.270 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 2 | 17 | 19 |

| 0.000 | 0.000 | 0.105 | 0.895 | 0.190 |

| 0.000 | 0.000 | 0.069 | 1.000 | |

| 0.000 | 0.000 | 0.020 | 0.170 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 26 | 28 | 29 | 17 | 100 |

| 0.260 | 0.280 | 0.290 | 0.170 | |

-------------|-----------|-----------|-----------|-----------|-----------|

**Per the results, k = 30 incurred 3 prediction errors.**

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| kpred5

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 24 | 0 | 1 | 0 | 25 |

| 0.960 | 0.000 | 0.040 | 0.000 | 0.250 |

| 0.960 | 0.000 | 0.036 | 0.000 | |

| 0.240 | 0.000 | 0.010 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 28 | 0 | 0 | 28 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.280 |

| 0.000 | 1.000 | 0.000 | 0.000 | |

| 0.000 | 0.280 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 1 | 0 | 27 | 0 | 28 |

| 0.036 | 0.000 | 0.964 | 0.000 | 0.280 |

| 0.040 | 0.000 | 0.964 | 0.000 | |

| 0.010 | 0.000 | 0.270 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 0 | 19 | 19 |

| 0.000 | 0.000 | 0.000 | 1.000 | 0.190 |

| 0.000 | 0.000 | 0.000 | 1.000 | |

| 0.000 | 0.000 | 0.000 | 0.190 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 25 | 28 | 28 | 19 | 100 |

| 0.250 | 0.280 | 0.280 | 0.190 | |

-------------|-----------|-----------|-----------|-----------|-----------|

**Per the results, k = 40 incurred 2 errors**

**In addition, k = 70 resulted the following prediction:**

Total Observations in Table: 100

| kpred6

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 25 | 0 | 0 | 0 | 25 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.250 |

| 0.962 | 0.000 | 0.000 | 0.000 | |

| 0.250 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 28 | 0 | 0 | 28 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.280 |

| 0.000 | 1.000 | 0.000 | 0.000 | |

| 0.000 | 0.280 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 1 | 0 | 27 | 0 | 28 |

| 0.036 | 0.000 | 0.964 | 0.000 | 0.280 |

| 0.038 | 0.000 | 1.000 | 0.000 | |

| 0.010 | 0.000 | 0.270 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 0 | 19 | 19 |

| 0.000 | 0.000 | 0.000 | 1.000 | 0.190 |

| 0.000 | 0.000 | 0.000 | 1.000 | |

| 0.000 | 0.000 | 0.000 | 0.190 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 26 | 28 | 27 | 19 | 100 |

| 0.260 | 0.280 | 0.270 | 0.190 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Only one incurred an error.

I utilized the scale instead of normalize, producing the following results:

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| zkpred

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 22 | 0 | 0 | 0 | 22 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.220 |

| 0.957 | 0.000 | 0.000 | 0.000 | |

| 0.220 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 1 | 25 | 0 | 0 | 26 |

| 0.038 | 0.962 | 0.000 | 0.000 | 0.260 |

| 0.043 | 1.000 | 0.000 | 0.000 | |

| 0.010 | 0.250 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 0 | 0 | 22 | 2 | 24 |

| 0.000 | 0.000 | 0.917 | 0.083 | 0.240 |

| 0.000 | 0.000 | 1.000 | 0.067 | |

| 0.000 | 0.000 | 0.220 | 0.020 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 0 | 28 | 28 |

| 0.000 | 0.000 | 0.000 | 1.000 | 0.280 |

| 0.000 | 0.000 | 0.000 | 0.933 | |

| 0.000 | 0.000 | 0.000 | 0.280 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 23 | 25 | 22 | 30 | 100 |

| 0.230 | 0.250 | 0.220 | 0.300 | |

-------------|-----------|-----------|-----------|-----------|-----------|

K = 35.

> zkpred2 <- knn(train = ztrain, test = ztest, cl = ktrainl, k = 10)

> CrossTable(x = ktestl, y = zkpred2, prop.chisq = FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| zkpred2

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 22 | 0 | 0 | 0 | 22 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.220 |

| 1.000 | 0.000 | 0.000 | 0.000 | |

| 0.220 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 0 | 26 | 0 | 0 | 26 |

| 0.000 | 1.000 | 0.000 | 0.000 | 0.260 |

| 0.000 | 0.963 | 0.000 | 0.000 | |

| 0.000 | 0.260 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 0 | 0 | 23 | 1 | 24 |

| 0.000 | 0.000 | 0.958 | 0.042 | 0.240 |

| 0.000 | 0.000 | 1.000 | 0.036 | |

| 0.000 | 0.000 | 0.230 | 0.010 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 1 | 0 | 27 | 28 |

| 0.000 | 0.036 | 0.000 | 0.964 | 0.280 |

| 0.000 | 0.037 | 0.000 | 0.964 | |

| 0.000 | 0.010 | 0.000 | 0.270 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 22 | 27 | 23 | 28 | 100 |

| 0.220 | 0.270 | 0.230 | 0.280 | |

-------------|-----------|-----------|-----------|-----------|-----------|

> zkpred3 <- knn(train = ztrain, test = ztest, cl = ktrainl, k = 70)

> CrossTable(x = ktestl, y = zkpred3, prop.chisq = FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 100

| zkpred3

ktestl | A | B | C | D | Row Total |

-------------|-----------|-----------|-----------|-----------|-----------|

A | 22 | 0 | 0 | 0 | 22 |

| 1.000 | 0.000 | 0.000 | 0.000 | 0.220 |

| 0.957 | 0.000 | 0.000 | 0.000 | |

| 0.220 | 0.000 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

B | 1 | 25 | 0 | 0 | 26 |

| 0.038 | 0.962 | 0.000 | 0.000 | 0.260 |

| 0.043 | 1.000 | 0.000 | 0.000 | |

| 0.010 | 0.250 | 0.000 | 0.000 | |

-------------|-----------|-----------|-----------|-----------|-----------|

C | 0 | 0 | 23 | 1 | 24 |

| 0.000 | 0.000 | 0.958 | 0.042 | 0.240 |

| 0.000 | 0.000 | 1.000 | 0.034 | |

| 0.000 | 0.000 | 0.230 | 0.010 | |

-------------|-----------|-----------|-----------|-----------|-----------|

D | 0 | 0 | 0 | 28 | 28 |

| 0.000 | 0.000 | 0.000 | 1.000 | 0.280 |

| 0.000 | 0.000 | 0.000 | 0.966 | |

| 0.000 | 0.000 | 0.000 | 0.280 | |

-------------|-----------|-----------|-----------|-----------|-----------|

Column Total | 23 | 25 | 23 | 29 | 100 |

| 0.230 | 0.250 | 0.230 | 0.290 | |

-------------|-----------|-----------|-----------|-----------|-----------|

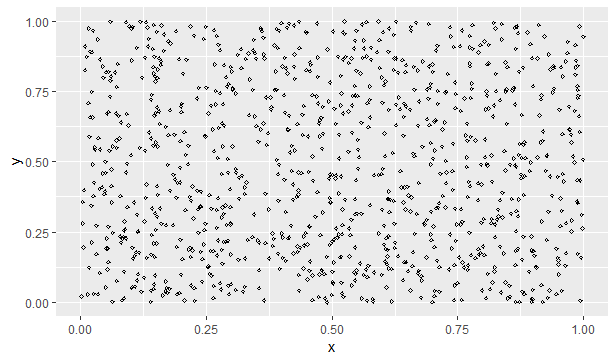


Figure 1.0. Starting 1000 plot.

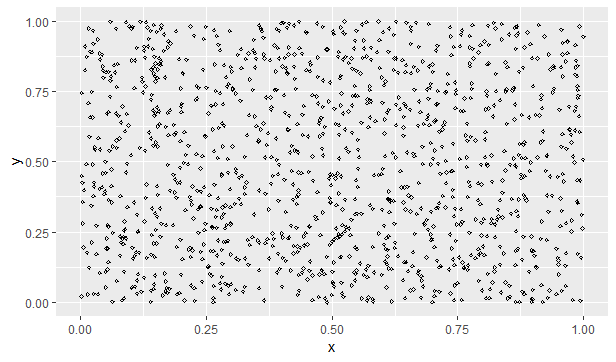


Figure 2.0. 1200 data plot including noise data

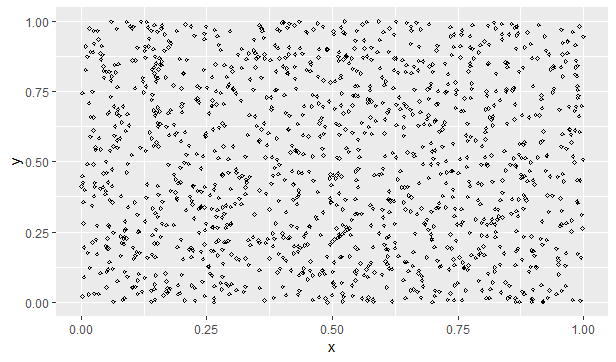


Figure 3.0. Non-normalized 1,300 data plot

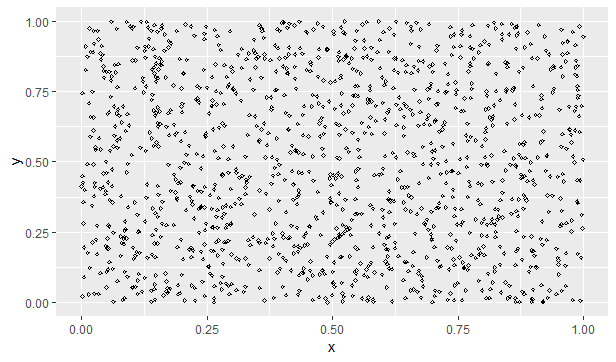


Figure 4.0. Normalized 1,300 data plot

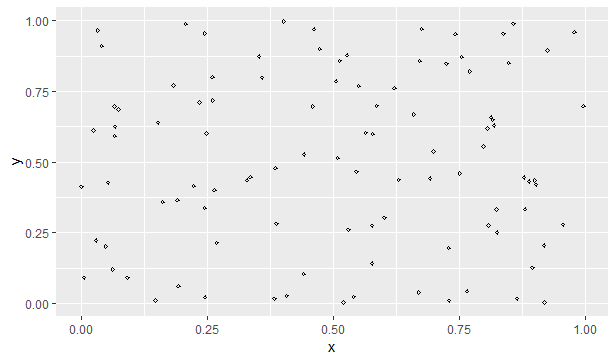


Figure 5.0. 100 Test data plot

**Analysis**

The model can correctly predict more than 90% of the test data in both cases (through normalization and the *scale* function). In practice, increasing the K would extend test data points’ range in classifying itself and risks misclassifying itself especially if the test value is located on the extremes between multiple categorical variables. Conversely, decreasing the k may pose ‘underfitting’ and potentially misclassifying the data point. It is the matter of where it will be located and to relative to other data points.

I expected that setting a high k would increase the error rate, which did not occur. This is likely to the percentage of noise data (15.37%) present in the 1,300-row dataset, which did not produce no to minimal difference. Also, R’s *runif* function generated ‘normally’ distributed numbers between 0 to 1. In other words, the data points were equally distributed.

The spread on the 1000, 1200, 1300 by eye test is uniform; non-normalized and normalized data remain nearly identical. The fact that both plots look the same tell that the non-normalized dataset was already normally distributed, despite not generating the *favstats* at this point of writing (I did generate the *favstats* of the normalized one). Figure 5.0 best displays the uniform coordinate point distribution. Skewing through adjusting the proportion of data points per quadrant and adding more noise data to the dataset would certainly incur more errors.

In all, adjusting the k value as low as 1 to as high as 100 would likely accurately classify the test data points with a quadrant value by those reasons mentioned above.