1. Consider the toy dataset below which shows if 4 subjects have diabetes or not, along with two diagnostic measurements.

Preg	BP	HasDiabetes	Preg.Norm	BP.Norm
2	74	No	0.5	1.0
3	58	Yes	1.0	0.2
2	58	Yes	0.5	0.2
1	54	No	0.0	0.0
p2	70	?	0.5	0.8

- a. Which variable is the "Class" variable?
 - The Class variable is HasDiabetes
- b. Normalize the Preg and BP values by scaling the minimum-maximum range of each column to 0-1. Fill in the empty columns in the table.
- c. Predict whether a subject with Preg=2, BP=70 will have diabetes using the 1-NN algorithm and
- i.Using Euclidean distance on the original variables

Nearest distance: 0 Nearest neighbor: Row 5

ii.Using Euclidean distance on the normalized variables

Nearest distance: 69.0163 Nearest neighbor: Row 1

For each of these cases, give the nearest distance, nearest neighbor (e.g., "Row 1" or "Row 2"), and prediction.

```
1 library(class)
 2 library(tidyverse)
 4 first \langle -c(2,3,2,1,2) \rangle
 5 second \leftarrow c(74,58,58,54,70)
 7 df <- data.frame(first, second)</pre>
 8
 9 normalized<-function(y) {</pre>
10
     x<-y[!is.na(y)]
11
12
13
     x \leftarrow (x - \min(x)) / (\max(x) - \min(x))
14
15
     y[!is.na(y)]<-x
16
17
     return(y)
18 }
19
20 apply(df[,c(1,2)],2,normalized)
21
22 a0 <- c(2, 70)
23 a1 <- c(2, 74)
24 | a2 < - c(3, 58)
25 a3 <- c(2, 58)
26 | a4 < - c(1, 54)
27 | a5 < -c(2, 70)
28
29 d1 <- sqrt(sum((a0-a1)^2))
30 d1
31 d2 <- sqrt(sum((a0-a2)^2))
32 d2
33 d3 <- sqrt(sum((a0-a3)^2))
34 d3
35 d4 <- sqrt(sum((a0-a4)^2))
36 d4
37 d5 <- sqrt(sum((a0-a5)^2))
38 d5
39
40 a0 <- c(2, 70)
41 a1 <- c(0.5, 1.0)
42 a2 <- c(1.0, 0.2)
43 a3 \leftarrow c(0.5, 0.2)
44 a4 <- c(0.0, 0.0)
45 | a5 < -c(0.5, 0.8)
46
47 d1 <- sqrt(sum((a0-a1)^2))
48 d1
49 d2 <- sqrt(sum((a0-a2)^2))
50 d2
51 d3 <- sqrt(sum((a0-a3)^2))
52 d3
53 d4 <- sqrt(sum((a0-a4)^2))
54 d4
55 d5 <- sqrt(sum((a0-a5)^2))
56 d5
57
```

localhost:40937 1/2

```
4/8/22, 10:11 PM
                                         /home/wesleyc/Schoolwork/CPSC 375/Homework#7/main.r
  58 pima <- read csv("pima-indians-diabetes-resampled.csv")
  59 pima
  60
  61
  62
  63 pima <- filter(pima, Glucose > 0)
  64
  65 normalize <- function(x) { return ((x-min(x)) / (max(x)-min(x)))}
  66 pimaNorm <- pima %>%
  67
     mutate(Preg.norm=normalize(Preg),Pedigree.norm=normalize(Pedigree),Glucose.norm=normalize(Glucos
  68
  69
  70 trainIndex <- sample(1:500)</pre>
  71
  72 | trainfeatures <- pimaNorm[trainIndex, c(1,2,3,4,5,6,7,8,9,10,11,12)]
  73 trainlabels <- pimaNorm[trainIndex, c(1,2,3,4,5,6,7,8,9,10,11,12)]
  74
  75 testIndex <- setdiff(1:nrow(pimaNorm), trainIndex)
  76
  77 | testfeatures <- pimaNorm[testIndex, c(1,2,3,4,5,6,7,8,9,10,11,12)]
  78 testlabels <- pimaNorm[testIndex, c(1,2,3,4,5,6,7,8,9,10,11,12)]
  79
  80 trainfeatures
  81 trainlabels
  82 testfeatures
  83 testlabels
  84
  85 trainfeatures <- pimaNorm[trainIndex, c(10,11,12)]
  86 trainlabels <- pimaNorm[trainIndex, c(10,11,12)]
  87
  88 testfeatures <- pimaNorm[testIndex, c(10,11,12)]
  89 testlabels <- pimaNorm[testIndex, c(10,11,12)]
  90
  91 trainfeatures
  92 trainlabels
  93 testfeatures
```

94 testlabels

98 table(testlabels, k1)

102 table(testlabels, k5)

106 table(testlabels, k11)

96 k1 <- knn(train = trainfeatures, test = testfeatures,

100 k5 <- knn(train = trainfeatures, test = testfeatures,

104 k11 <- knn(train = trainfeatures, test = testfeatures,

cl = trainlabels, k=5)

cl = trainlabels, k=11)

cl = trainlabels, k=1)

95

97

99

101

103

105

localhost:40937