Keiland Pullen
DSC 441 – Fall 2021

Assignment 1

Problem 1.)

- **a.)** Yes, monitoring a patient's heart rate for abnormalities can be a data-mining task. There could be a dataset that contains recorded instances of the patient's normal heartbeat and possibly a dataset that contains abnormal heartbeats. A model can be built based on the normal heartbeat data to detect abnormal heartbeats.
- **b.)** Yes, integrating information from multiple stores into one source is a data-mining task. This task is known as "Data Integration".
- **c.)** Yes, sorting a customer database based on the amount spent in a store is not a data-mining task. This is an example of "Classification"
- **d.)** No, predicting the outcomes of tossing (fair) pair of dice is more about probability.
- **e.)** Yes, monitoring seismic waves for earthquake activities. Similar to the heart rate question, data can be collected on the seismic waves – allowing the possibility of models to be created that can predict any potential damage that the waves may cause.

Problem 2.)

For dataset X = {-5.0, 23.0, 17.6, 7.23, 1.11}

a.) Decimal scaling on interval [-1,1]
max|v| = 23.0 -> j=100

[-0.05, 0.23, 0.176, 0.0723, 0.0111]

b.) Min-Max normalization on interval [0,1]
$v' = \frac{v-min}{\max - min}$ * (new_max – new_min) + new_min

For -5.0:
$v' = \frac{-5 -(-5)}{23.0-(-5)} * (1 - 0) + 0$
v' = 0

For 23.0
$v' = \frac{23.0 -(-5)}{23.0-(-5)} * (1 - 0) + 0$
$v' = \frac{28}{28} * (1 - 0) + 0$
v' = 1

For 17.6
$v' = \frac{17.6 -(-5)}{23.0-(-5)} * (1 - 0) + 0$
$v' = \frac{22.6}{28} * (1 - 0) + 0$
v' = 0.807

For 7.23
v' = $\frac{7.23 - (-5)}{23.0 - (-5)}$ * (1 – 0) + 0
v' = $\frac{12.23}{28}$ * (1 – 0) + 0
v' = 0.437

For 1.11
v' = $\frac{1.11 - (-5)}{23.0 - (-5)}$ * (1 – 0) + 0
v' = $\frac{6.11}{28}$ * (1 – 0) + 0
v' = 0.218

c.) Min-Max normalization on interval [-1, -1]

v' = $\frac{v - min}{\max - min}$ * (new_max – new_min) + new_min

For -5.0:
v' = $\frac{-5 - (-5)}{23.0 - (-5)}$ * (-1 – (-1)) + 0
v' = 0

For 23.0
v' = $\frac{23.0 - (-5)}{23.0 - (-5)}$ * (-1 – (-1)) + 0
v' = $\frac{28}{28}$ * (-1 – -1) + 0
v' = 0

For 17.6
v' = $\frac{17.6 - (-5)}{23.0 - (-5)}$ * (-1 – -1) + 0
v' = $\frac{22.6}{28}$ * (-1 – (-1)) + 0
v' = 0

For 7.23
v' = $\frac{7.23 - (-5)}{23.0 - (-5)}$ * (-1 – (-1)) + 0
v' = $\frac{12.23}{28}$ * (-1 – -1) + 0
v' = 0

For 1.11

v’ = $\frac{1.11 - (-5)}{23.0 - (-5)}$ * (-1 – (-1)) + 0

v’ = $\frac{6.11}{28}$ * (-1 – -1) + 0

v’ = 0

d.) Z-score or standard deviation normalization

μ (mean) = 8.788
σ (sd) = 10.306

v’ = $\frac{v - \mu}{\sigma}$

For -5.0:

v’ = $\frac{-5.0 - 8.788}{10.306}$

v’ = $\frac{-13.788}{10.306}$

v’ = -1.34

For 23.0:

v’ = $\frac{23.0 - 8.788}{10.306}$

v’ = $\frac{14.212}{10.306}$

v’ = 1.37

For 17.6:

v’ = $\frac{17.6 - 8.788}{10.306}$

v’ = $\frac{8.812}{10.306}$

v’ = 0.855

For 7.23:

v’ = $\frac{7.23 - 8.788}{10.306}$

v’ = $\frac{-1.558}{10.306}$

v’ = -0.151

For 1.11:

v’ = $\frac{1.11 - 8.788}{10.306}$

v’ = $\frac{-7.678}{10.306}$

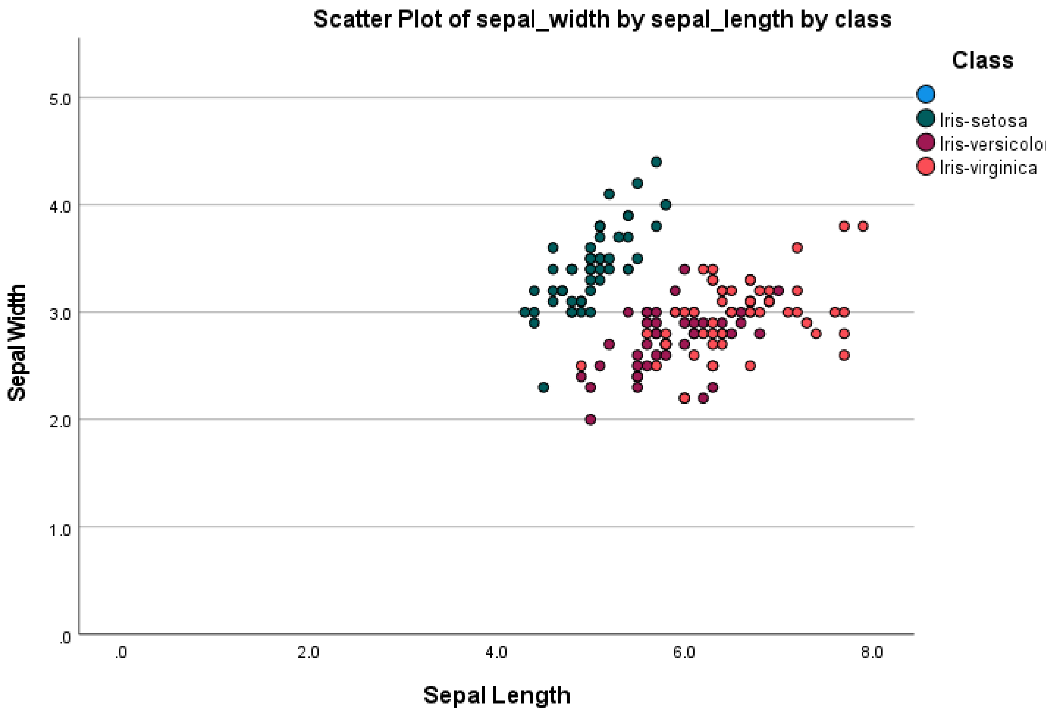
v’ = -0.745

Keiland Pullen Assignment 1
DSC 441 – Fall 2021

e.) From the notes, we wouldn't be sure of which normalization method to use until we address the data problem. Its stated that Min-Max is the default. Min-Max can promise that all features will have the same scale, but doesn't address outliers. Z-score is important if distribution and mean are a priority. However, Z-score does not produce normalized data with the same scale, but it does handle outliers. Decimal scaling is preferred if zero (0) or negative values matter.

Problem 3.)

a.)



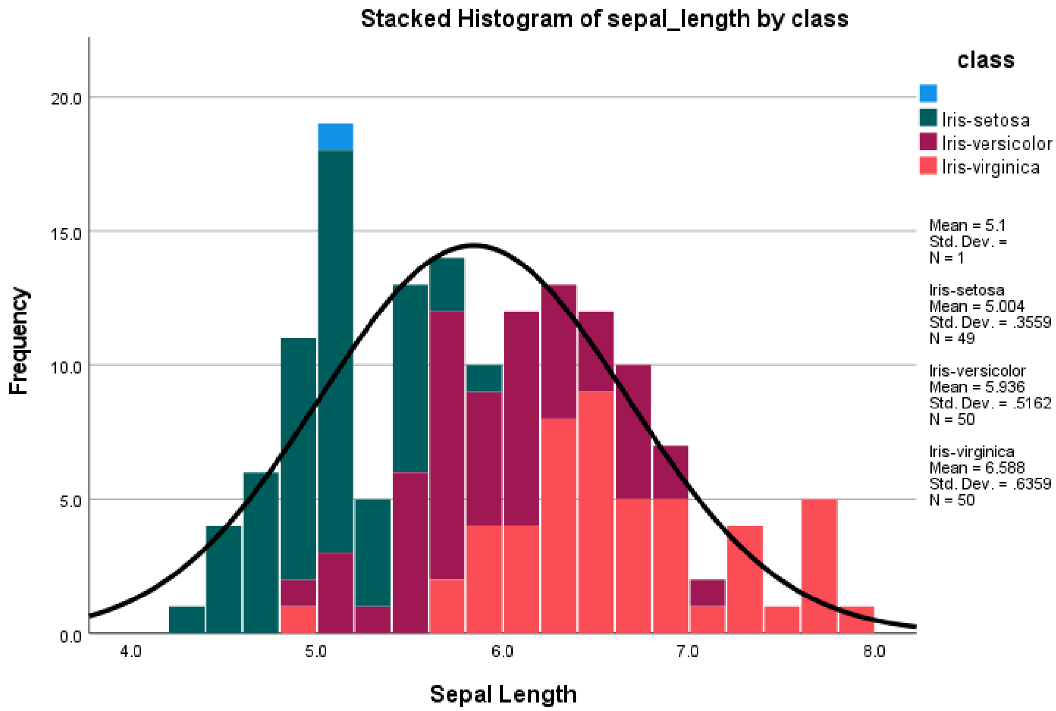
Could a classification algorithm be successful in classifying the data with respect to the Sepal Length and Sepal Width variables? I would say "No" and that the above plot demonstrates this. If there were only 2 Classes, then it might be possible, but in this case the cluster of Iris-versicolor and Iris-virginica are mixed together.

Keiland Pullen Assignment 1
DSC 441 – Fall 2021

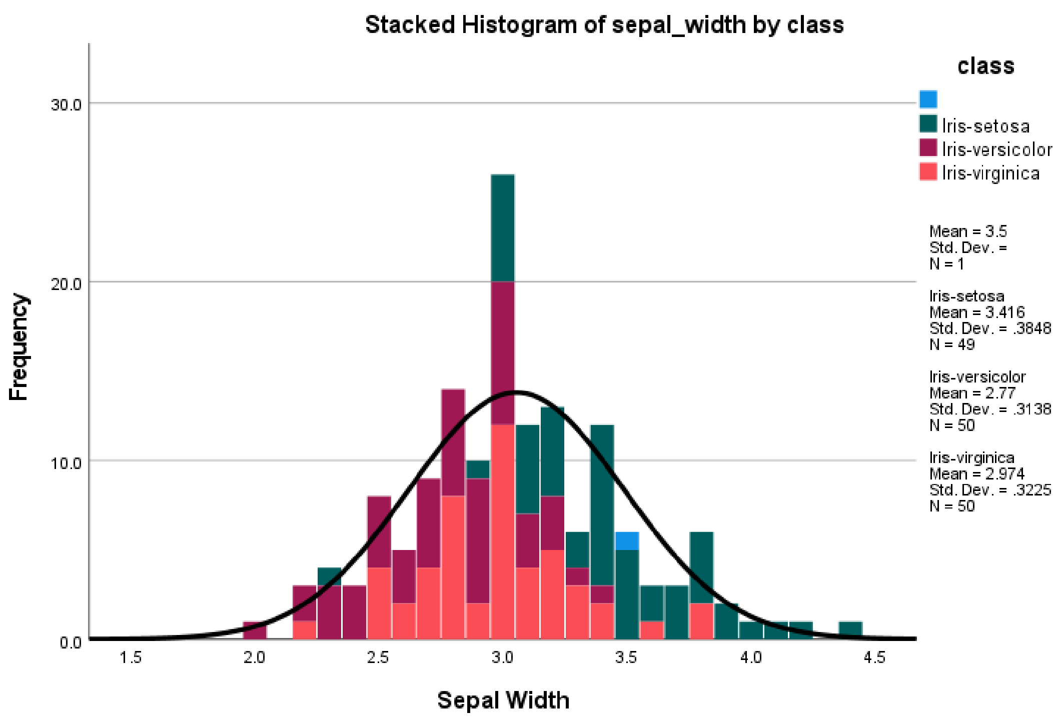
b.)

In the case of the "Petal Width" and "Petal Length" variables, the plot does support the use of a classification algorithm. Each of the Classes appear to be clustered together.

c.)

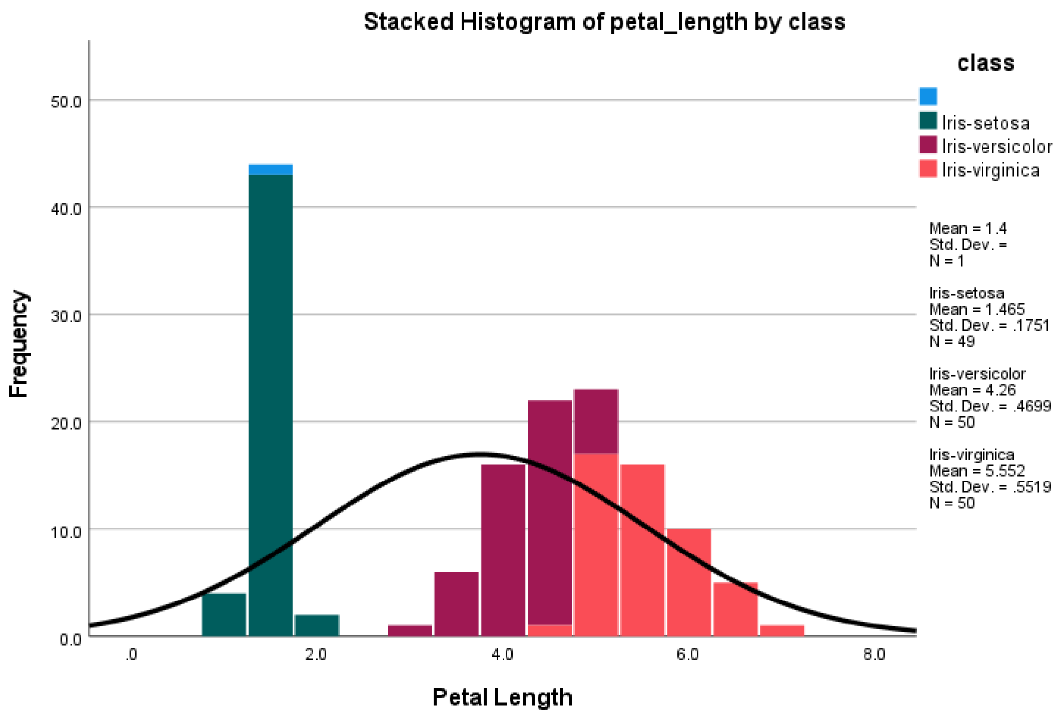


For the Sepal Length histogram the collective data appears to be normal, with what appears to be a couple outliers.

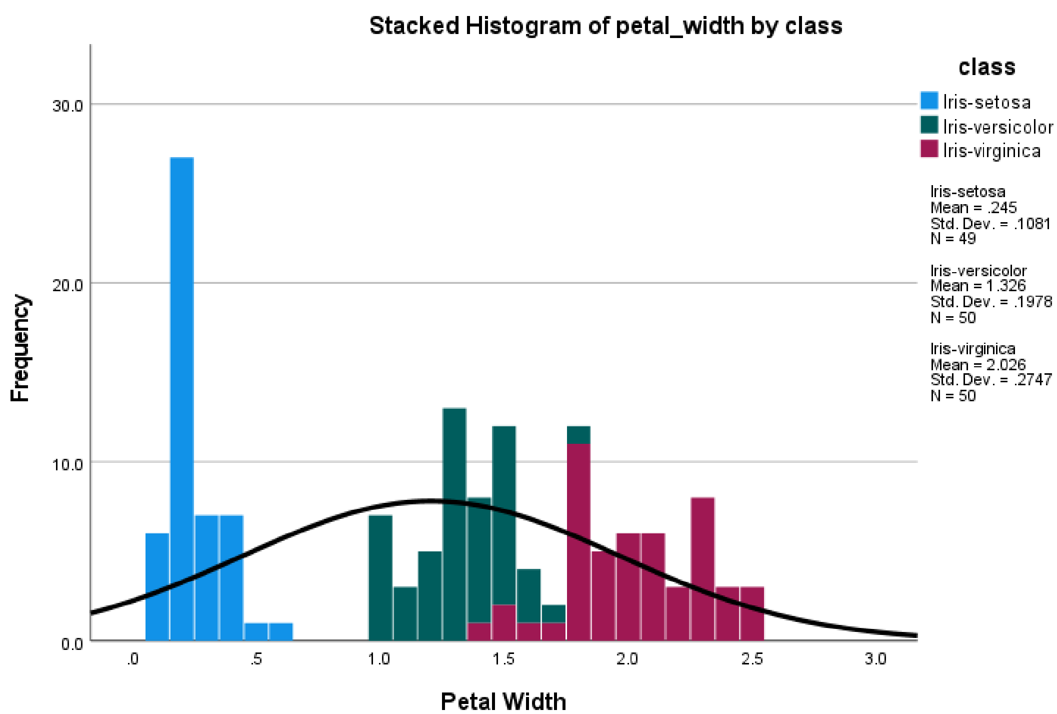


For the Sepal Width histogram, the collective data displays a normal distribution with several outliers.

Keiland Pullen Assignment 1
DSC 441 – Fall 2021



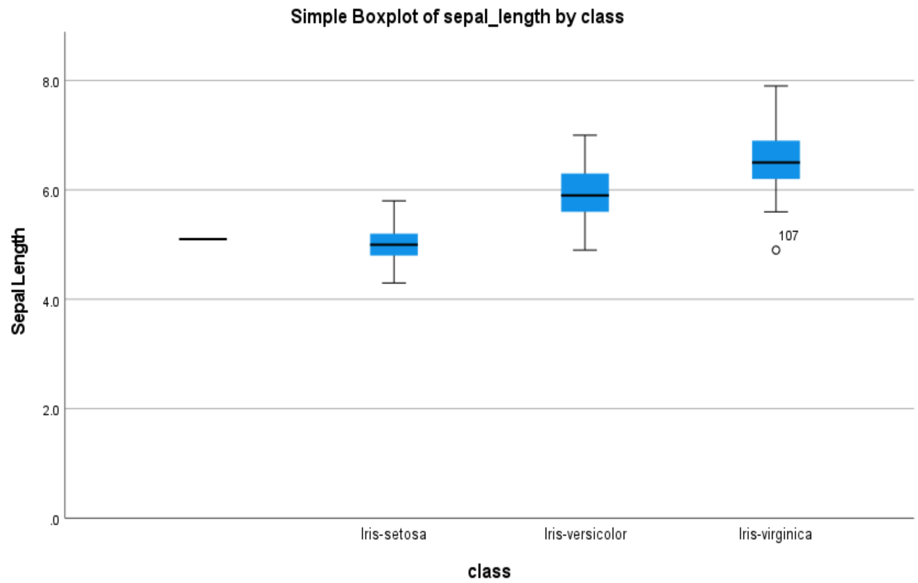
For the Petal Length histogram, could be bimodal as the gap indicates that this histogram is skewed to the left and there are two peaks.



For the Petal Width histogram, the gap indicates that the collective data is skewed to the left and could be bimodal.

Keiland Pullen Assignment 1
DSC 441 – Fall 2021

d.)



The above boxplot for Sepal length displays that there appears to be at least 1 outlier and there is a possibility that at least one empty data instance.

e.)

The above boxplot for Petal Length displays at least 2 outliers and what appears to be 1 empty data instance.

f.)

See uploaded SPSS documents:

- DSC_441_assignment1_KeilandPullen.sav
- KPullen_Hist_Boxplots.spv