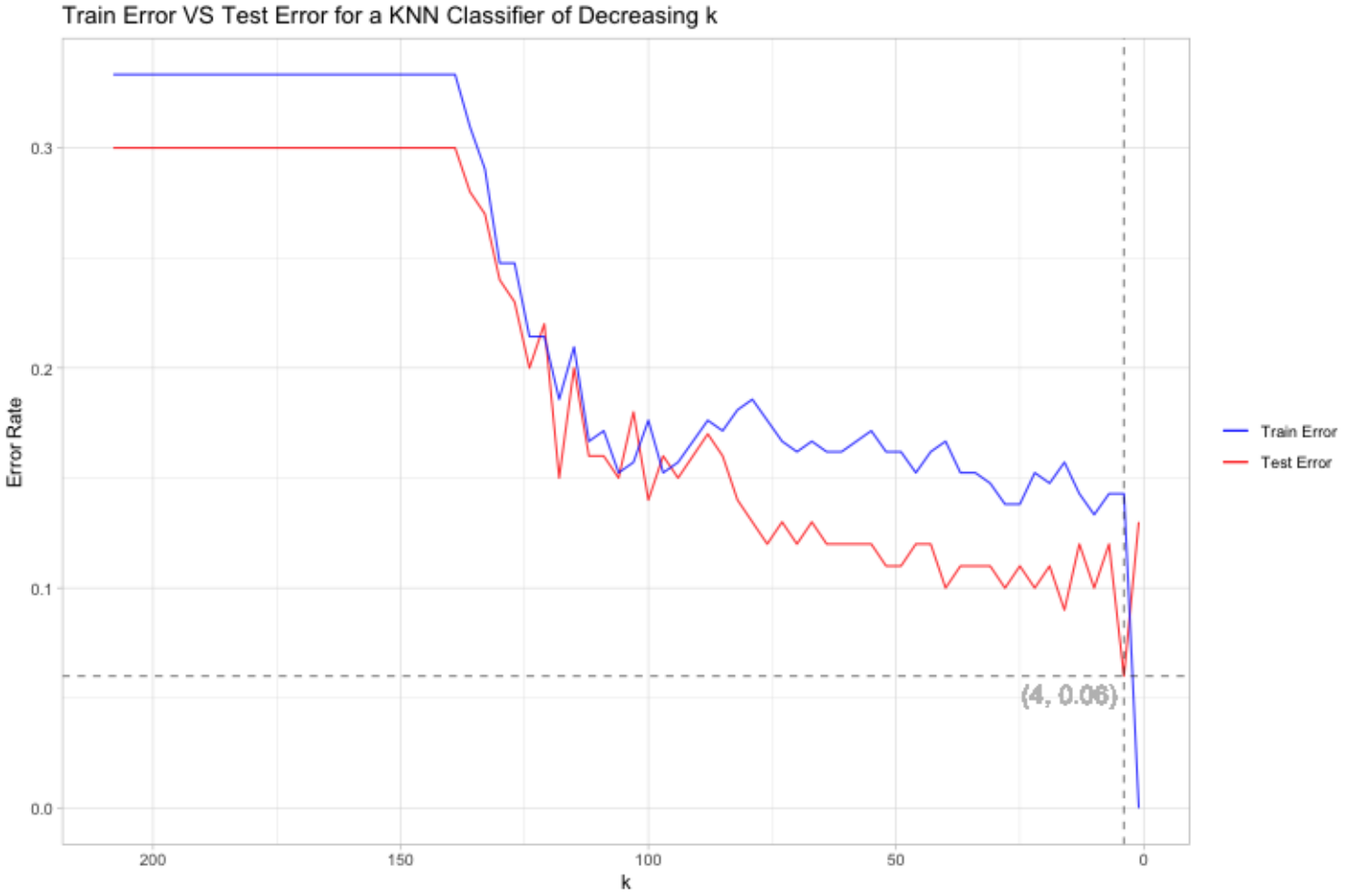
INF 552 Homework 1 Report
Author: Nelson Lam
Due Date: Sept 17, 2019

(b) i. + ii.

The above plot shows 15 scatterplots between 6 independent variables, and 6 boxplots for each independent variable, with blue indicating class 1 and red indicating class 0. Histograms, density plots and correlations are also shown. From the boxplots, we observe higher values for all class 1 over class 0 except for pelvic radius. There's also quite a bit of overlapping in the interquartile ranges. From the density plots, we can see that degree spondylolisthesis is heavily skewed for both class 0 and 1. This is fine though as KNN makes no assumptions about the underlying distribution of the variables. Other variables are approximately normal.

(c) ii.



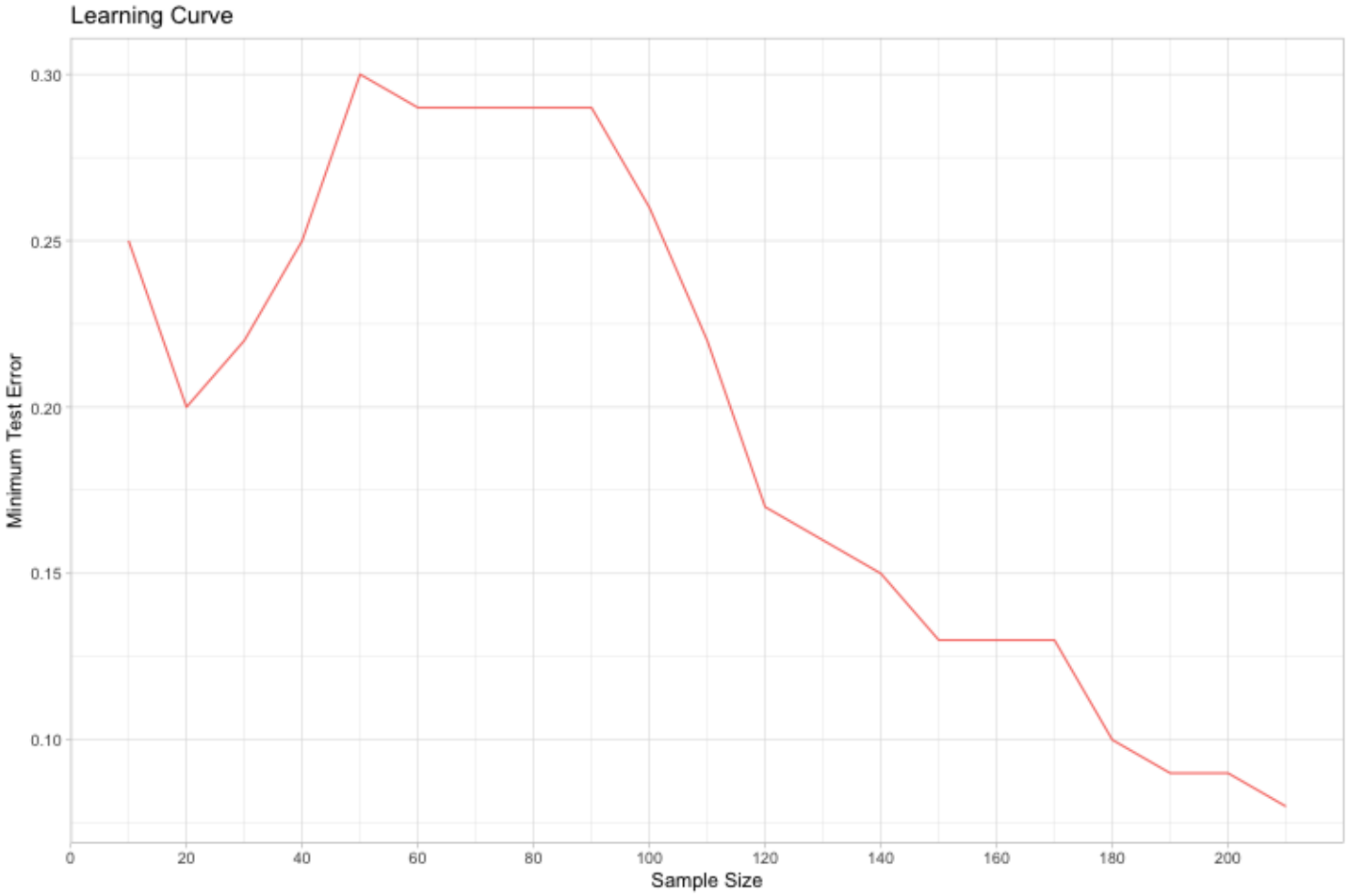
We can see that from the above graph, the lowest test error rate 0.06 is achieved when k is 4.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.83      0.89        30
           1       0.93      0.99      0.96        70

    accuracy                           0.94       100
   macro avg       0.95      0.91      0.93       100
weighted avg       0.94      0.94      0.94       100
```

For k = 4, precision = 0.94, recall = 0.94, f1-score = 0.94.

We have true negative count = 25 and false positive count = 5.
True negative rate is given by TN / (TN + FP), and thus it's equal to 0.833.

(c) iii.



From the above learning curve, we can see fluctuations in minimum test error when sample size < 90. However, the minimum test error decreases steadily thereafter, achieving the lowest at 0.08, when all sample are used.

(d)

Metric	Best test error
Manhattan Distance*	0.11
Minkowski p = 0.2**	0.11
Chebyshev Distance	**0.08**
Mahalanobis Distance	0.19

* k = 11 or k = 26 gives the lowest f1-score (with the same test error).
** The best $\log_{10}(p)$ are when p = 0.2, 0.8, 0.9 or 1, using k = 11.

The Chebyshev Distance produces the lowest best test error.

(e)

Metric	Best test error
Euclidean Distance	0.10
Manhattan Distance	0.10
Chebyshev Distance	0.11

The three distances give similar result for weighted decision.

(f)
The lowest training error rate obtained is 0. This is achieved whenever k = 1, in which the nearest neighbor of every point to be classified is itself, leading to an error rate of 0 naturally.