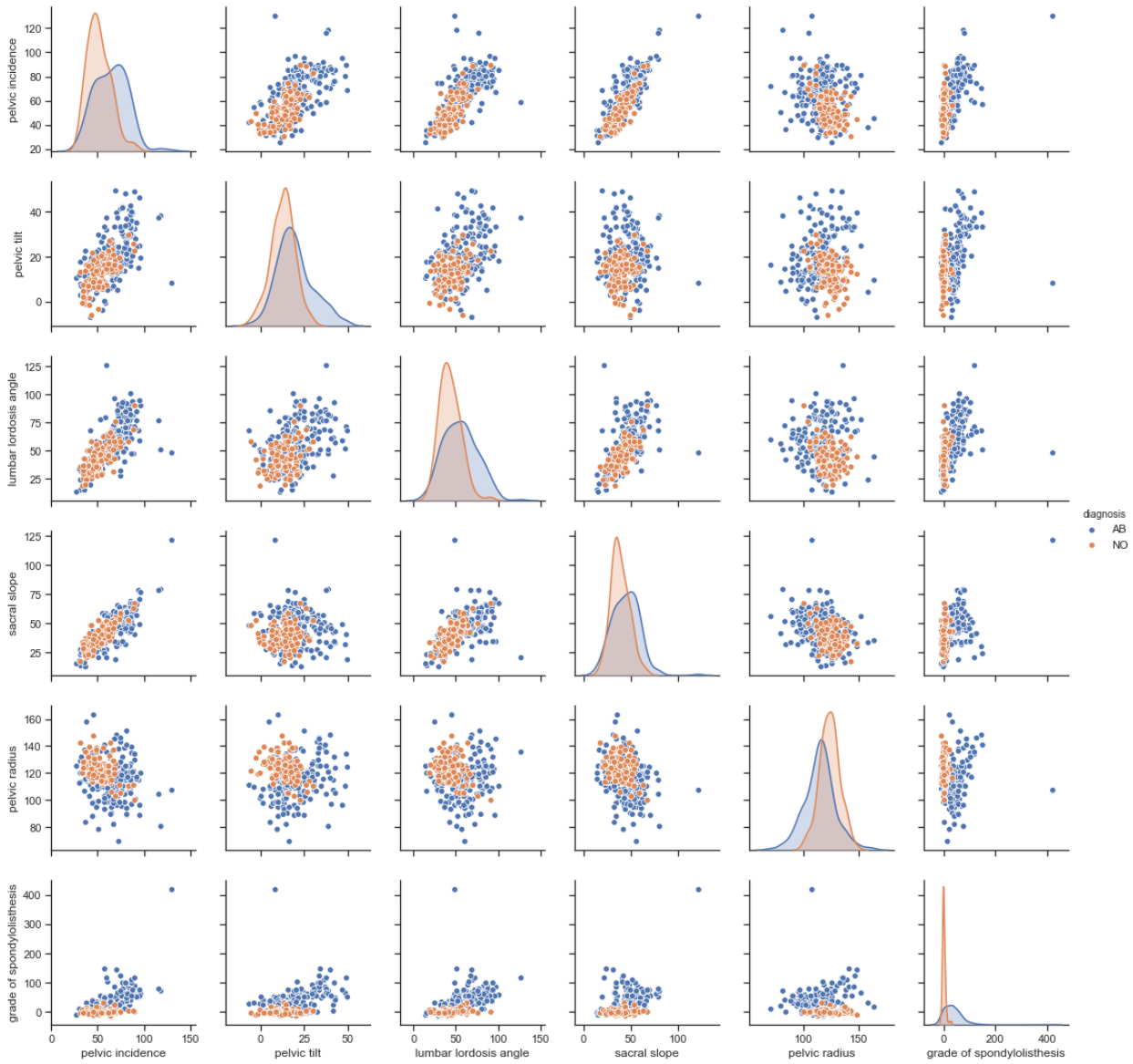
Rubina Iman Kabir
678-560-5414
Homework 1
INF 552

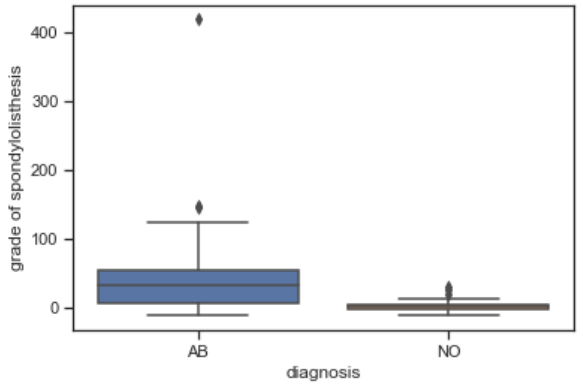
1. (b) (i)



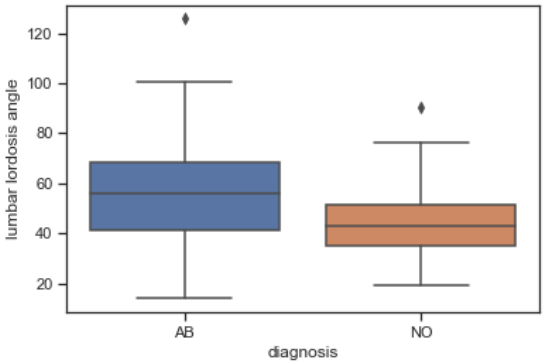
Correlation Values	Pelvic incidence	Pelvic tilt	Lumbar lordosis angle	Sacral scope	Pelvic radius	Grade of spondylolisthesis
Pelvic incidence		0.629	0.717	0.815	-0.248	0.639
Pelvic tilt			0.433	0.062	0.033	0.398
Lumbar lordosis angle				0.598	-0.080	0.533
Sacral scope					-0.342	0.524
Pelvic radius						-0.026

- Values in green indicate variables that exhibit a weak correlation. Future work would consist of sampling out the correlated features from the data that exhibit moderate to high correlation to examine the performance of KNN (e.g. computation time).

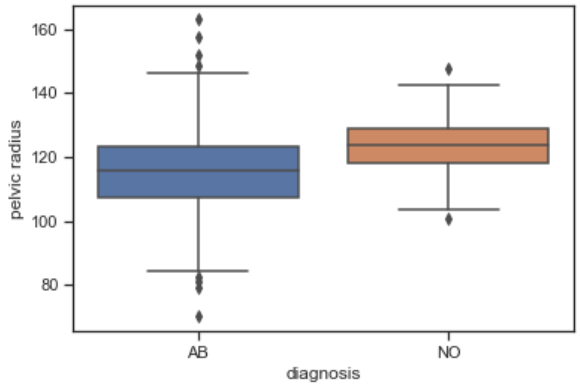
(b) (ii)



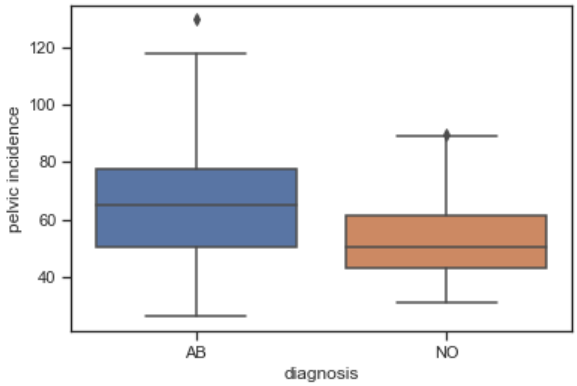
AB: larger spread than NO, higher average,
Outliers far from max
NO: very small spread



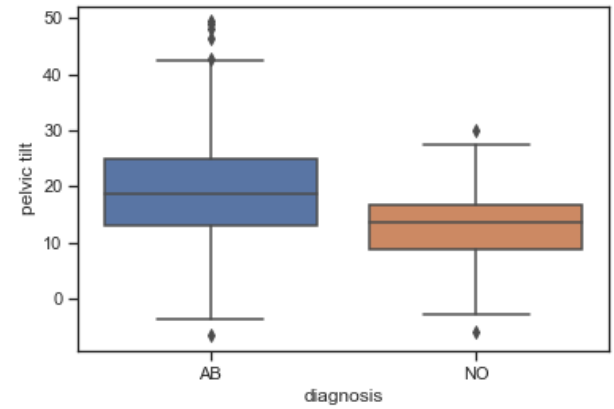
AB: larger spread
NO: small spread, slightly lower mean than AB



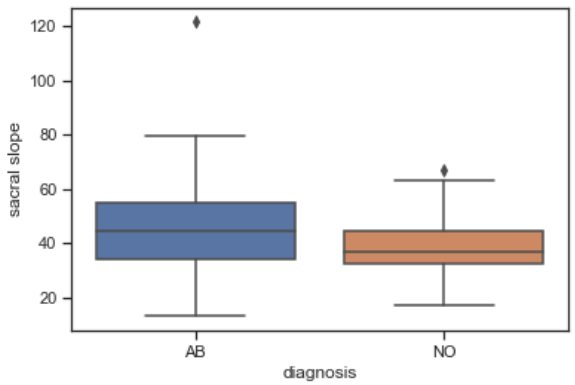
AB: lots of outliers, smaller mean
NO: higher mean, smaller spread



AB: larger spread, higher mean, more outliers
NO: lower mean

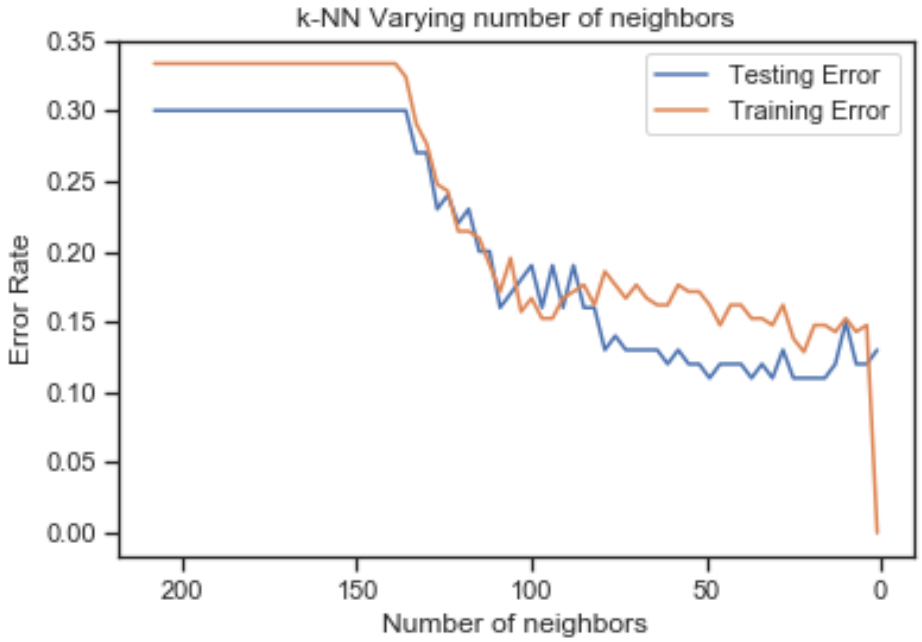


AB: lots of outliers, higher mean
NO: small spread



AB: outliers farther from maximum
NO: small spread

(c) (i)

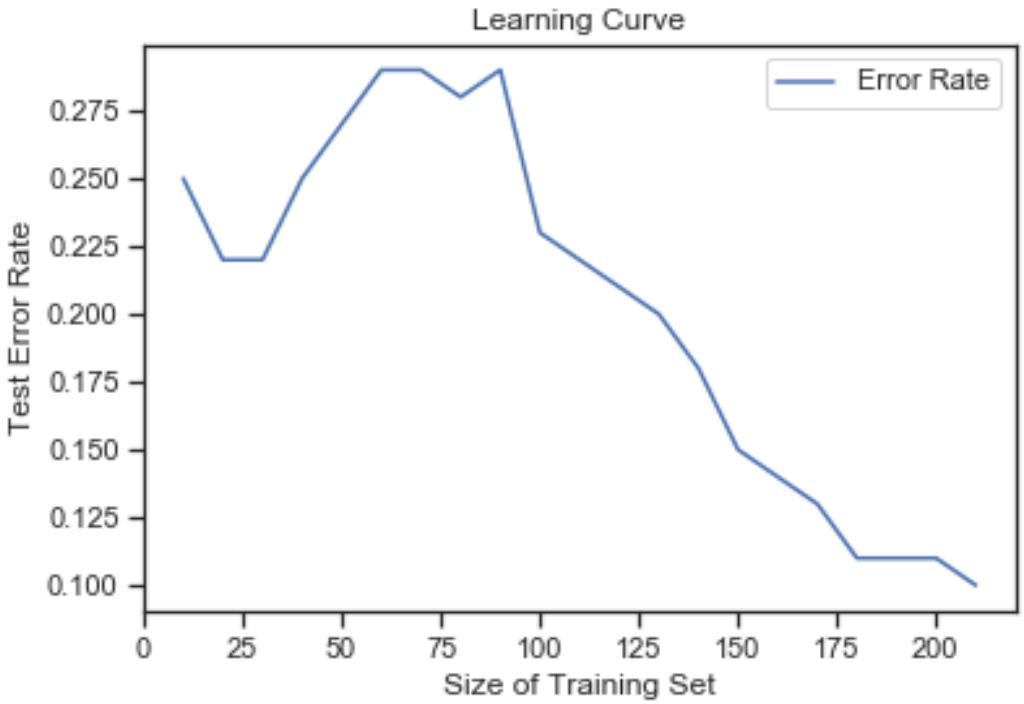


Based off the testing accuracies for tested k-values: **k* = 16 or k*=49**.

k*	Testing Error	Confusion Matrix	True Pos Rate	True Neg Rate	Precision	F-Score
16	0.11	[[70 0] [11 19]]	0.633	1.0	1.0	0.7755
49	0.11	[[69 1] [10 20]]	0.667	0.986	0.952	0.784

- Note, the F-score and TPR for k*=49 is slightly higher, though TNR and precision is lower. However, k*=49 is preferred to prevent overfitting. The training error is higher than the testing error for both k* values, hence both k* values provide a decent generalized model.
- Training error is higher than the testing error for k in between [4,118], so these values in this range provide a generalized model as well.
- The training and test error exhibit a decreasing yet jagged behavior as the number of neighbors decrease.
- Average training error: 0.2284353741496601
  Average testing error: 0.20542857142857163
- Increasing the number of neighbors becomes useless after k = 136 since both the test and train error begin to converge to a maximum error value of approximately 0.3 and 0.35, respectively.

(c)(iii)



- Achieve a low test error rate for training set sizes N=10,20,and 30, then the error spikes until N=100. After N = 100, the test error declines slightly linearly until N=210.
- The test error rate eventually declines as we increase the size of our test set.

(d) (i) Using Minkowski, **k*=21.** Below is a table of the reported errors for various p-values, and Chebychev distance metrics.

Metric	k*	Test Error
Minkowski	21	0.10
p = 0.100000	21	0.16
p = 0.200000	21	0.15
p = 0.300000	21	0.14
p = 0.400000	21	0.12
p = 0.500000	21	0.12
p = 0.600000	21	0.13
p = 0.700000	21	0.13
p = 0.800000	21	0.12
p = 0.900000	21	0.12
p = 1.000000	21	0.10
Chebychev	21	0.11
Mahalanobis	1	0.17

- As we increase p by 0.1, our test error declines; reaching a maximum when p=1 (aka using the Minkowski distance).
- BESTP = 1.00

(e) Using weighted voting, reported is the minimum test error found over k ∈ {1,6,11,16,...,196}

euclidean	0.10
manhattan	0.10
chebyshev	0.11

(f) The lowest testing error rate I achieved is **0.09999999999** using the Minkowski distance for k = 21.