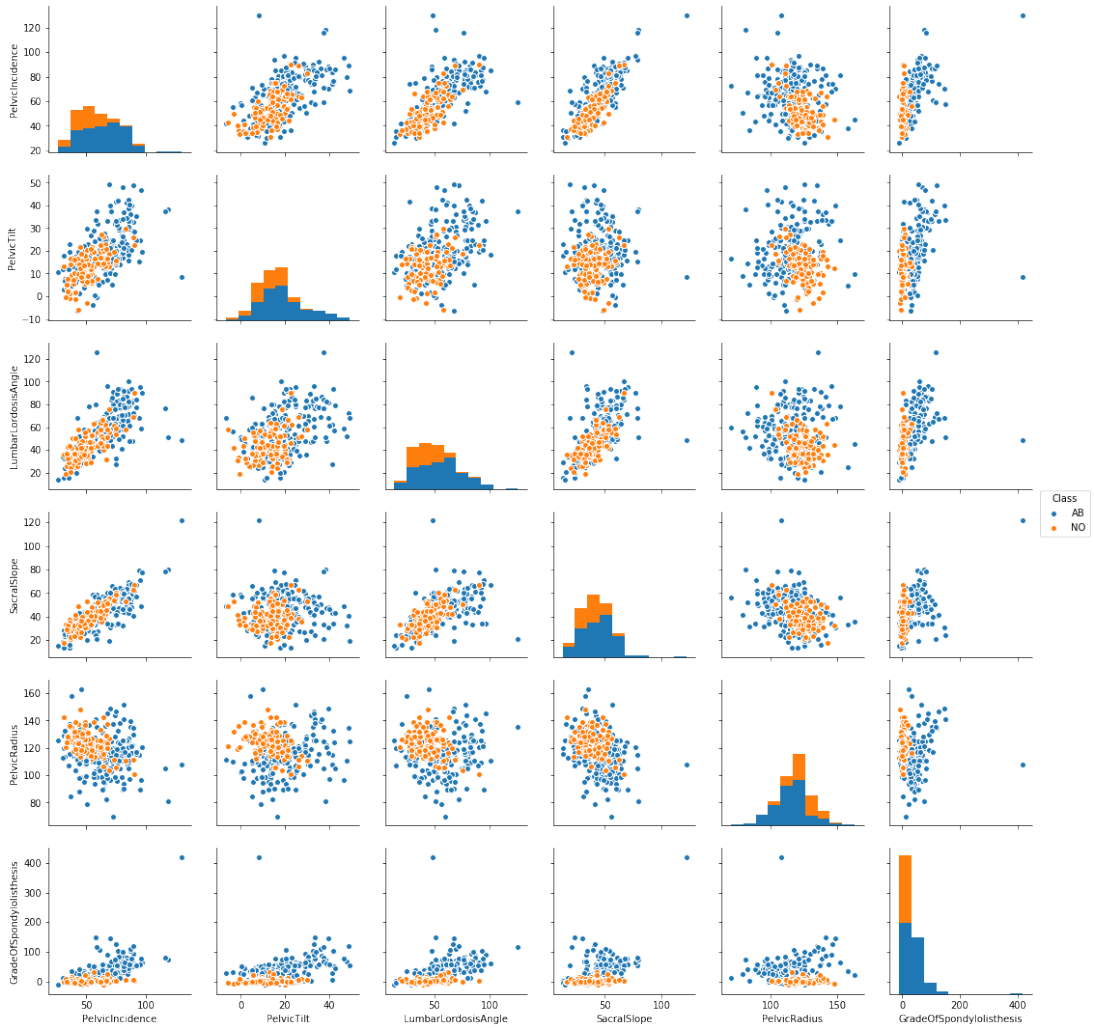
Report

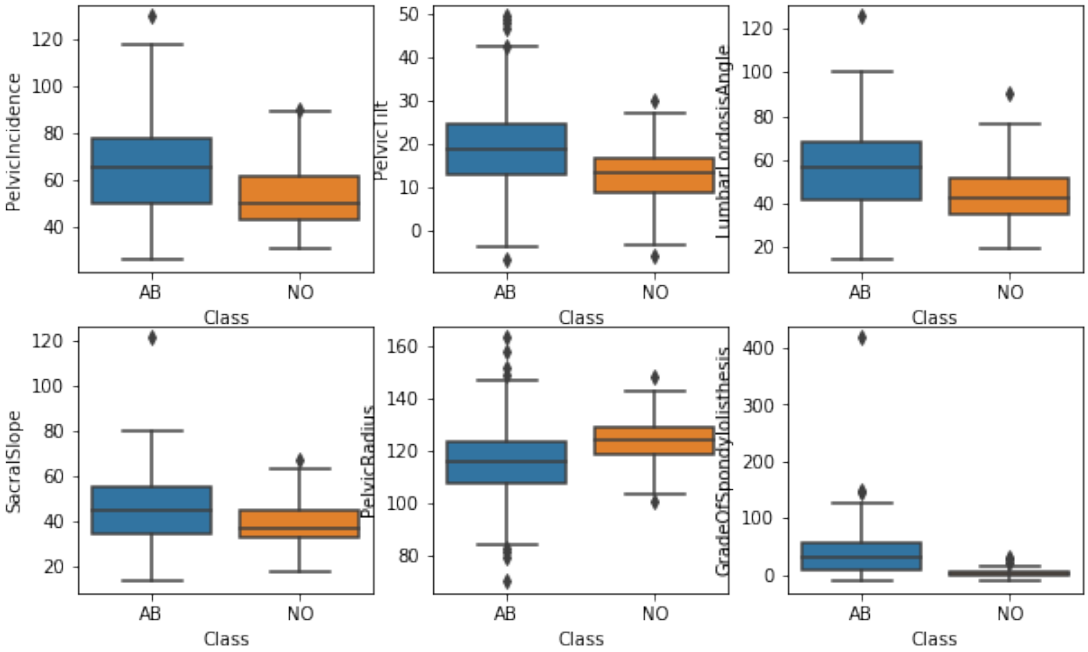
Jialin shi
UID: 7948542502

(a) Use pandas to load data. Rename the columns.

(b-i) Use seaborn to plot scatter plot between each pair of variables.



(b-ii) Use matplotlib and Seaborn to plot boxplot.



(b-iii) Use pandas to split data into train set and test set.

Replace 'AB'and 'NO' with '1'and'0'.

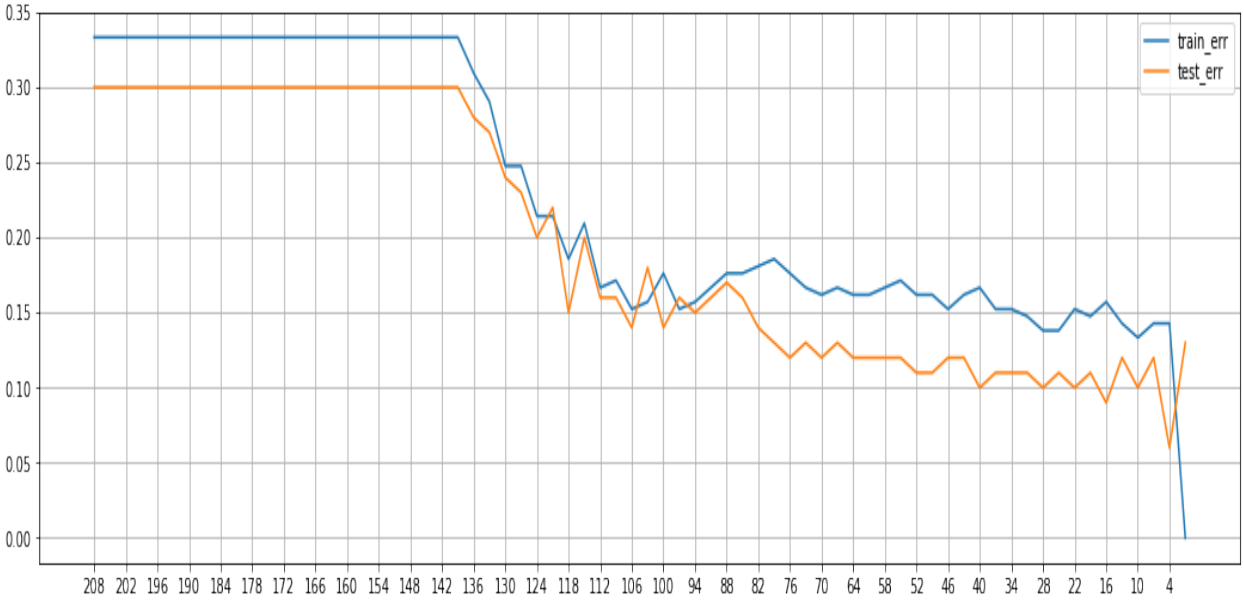
(c-i) Classification using KNN on Vertebral Column Data Set

Define the model: init K-NN

(c-ii) Plot train_error and test_error

For training size =210, x=k-value, y= train_error, test_error

when k =4, test_err = min(test_err)

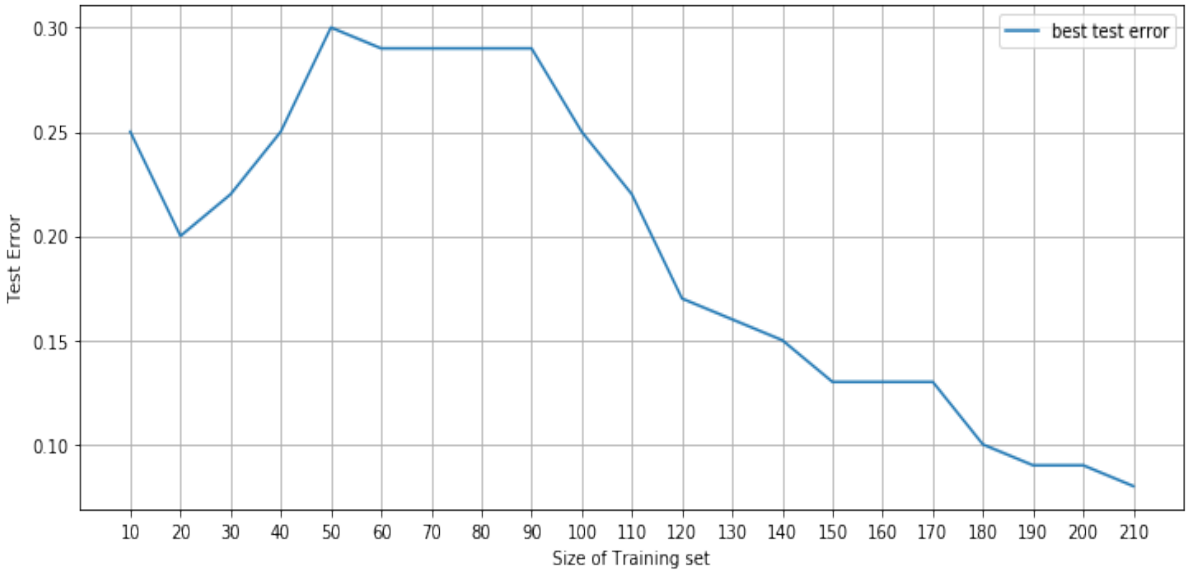


when k =4, Confusion Matrix is:
[[69 1]
[5 25]]

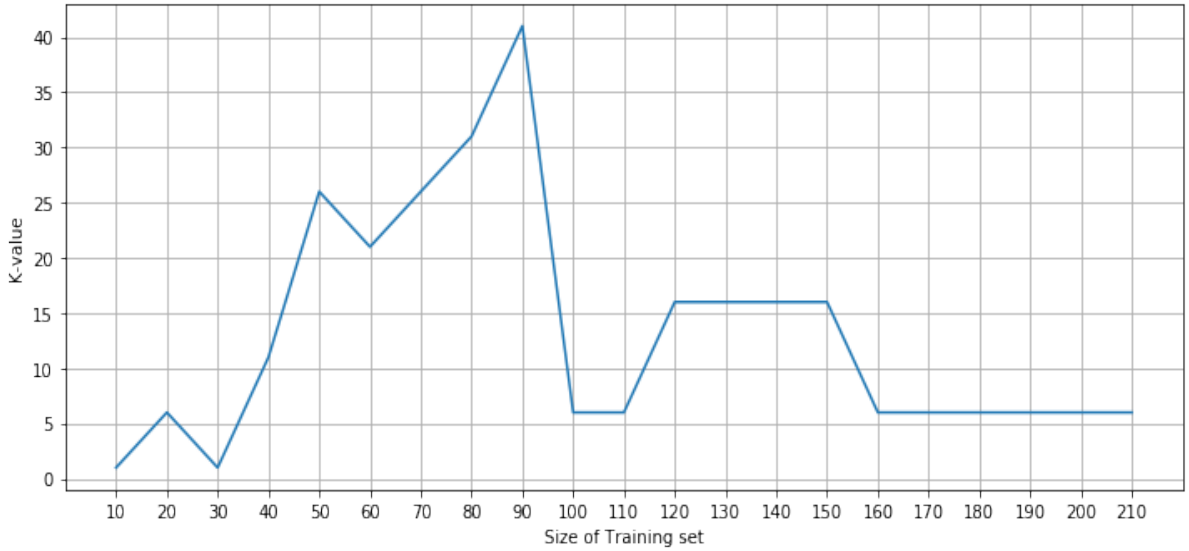
F1 score is: 0.9583333333333333
True Positive Rate is: 0.9857142857142858
True Negative Rate is: 0.8333333333333334
Precision is: 0.9324324324324325
Test Accuracy: 0.94

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	**0.96**	**0.83**	**0.89**	**30**
1	**0.93**	**0.99**	**0.96**	**70**
AVG / TOTAL	**0.94**	**0.94**	**0.94**	**100**

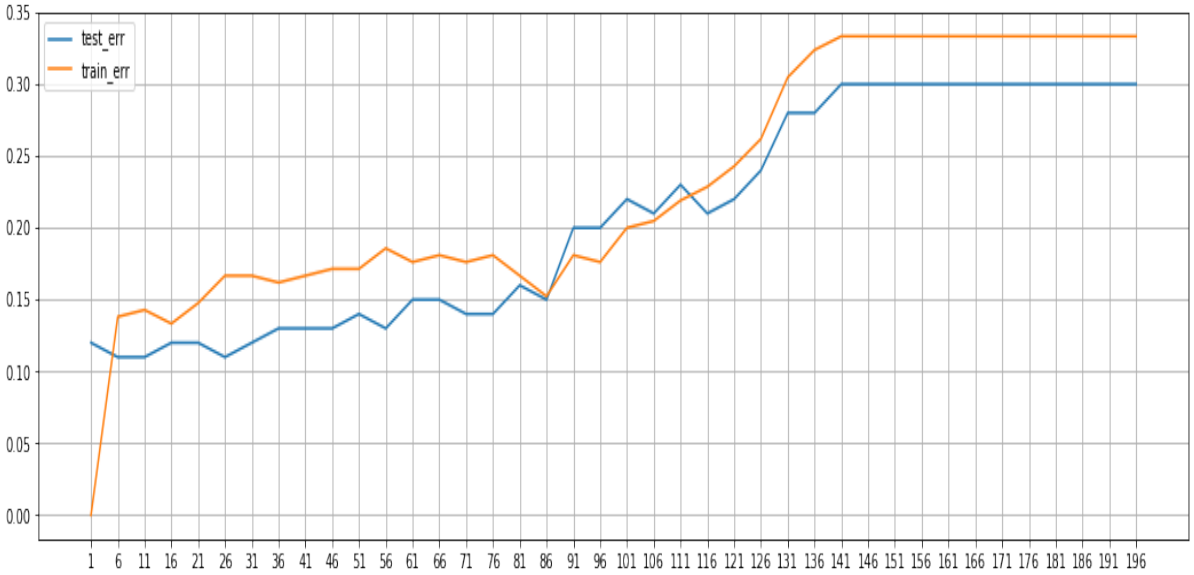
(c-iii) Plot learning curve
for each size of training set, the test error looks like:



best k-value for each training size

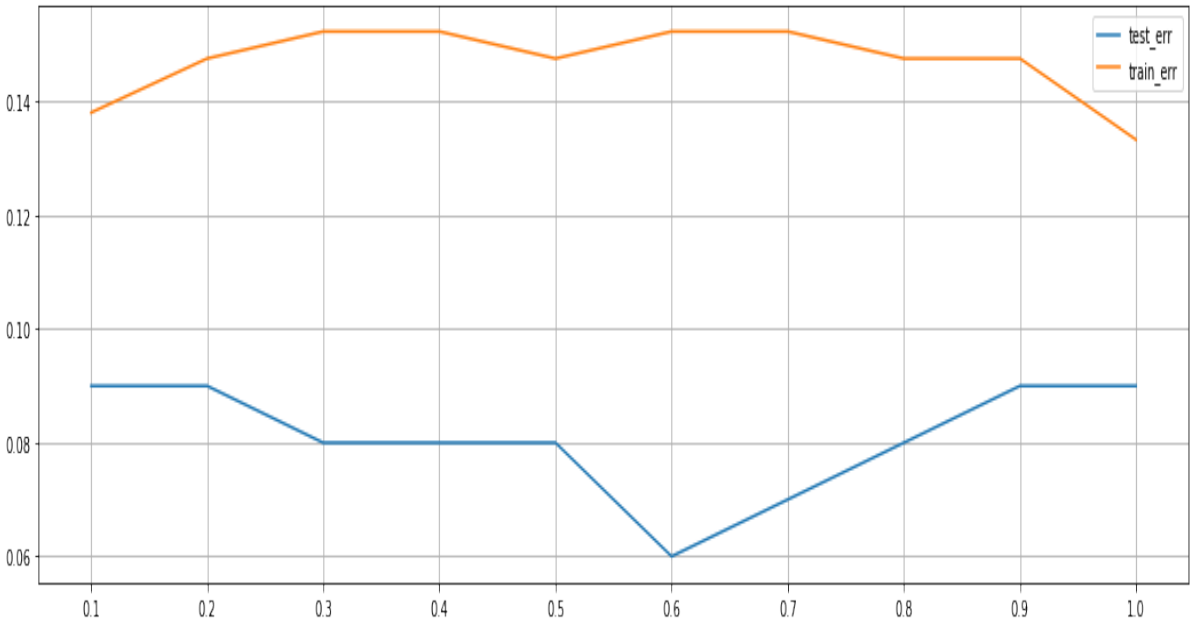


(d-i-A) find best k-value ［metric = 'minkowski'］manhattan



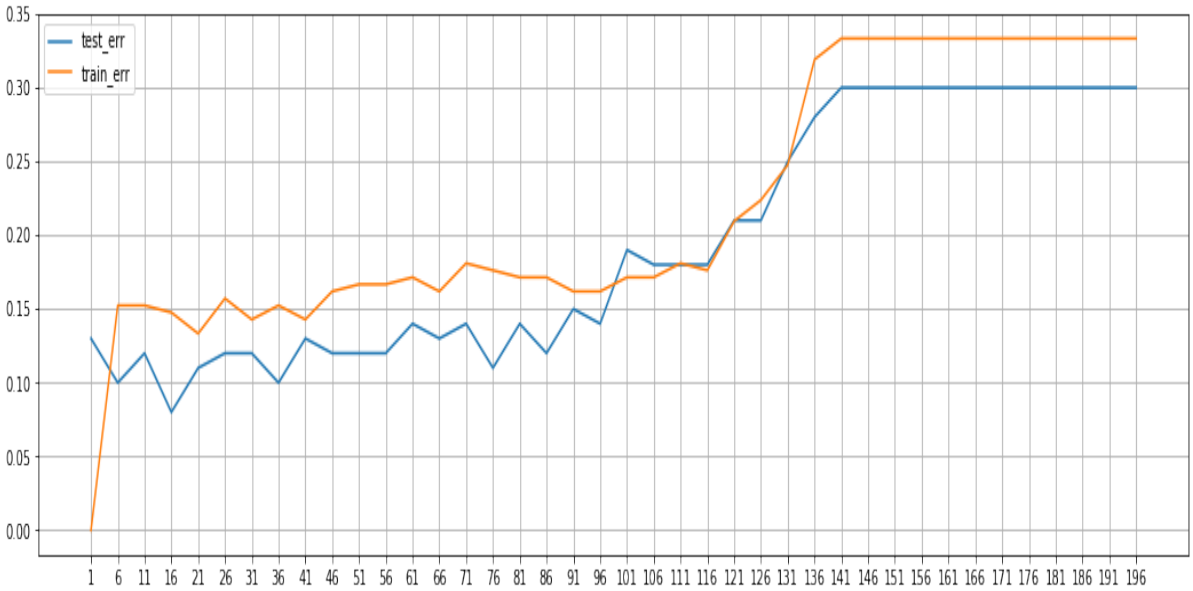
min test_err: 0.10999999999999999
best k-value: 6

(d-i-B) find best log10(p)-value　[metric = 'minkowski']



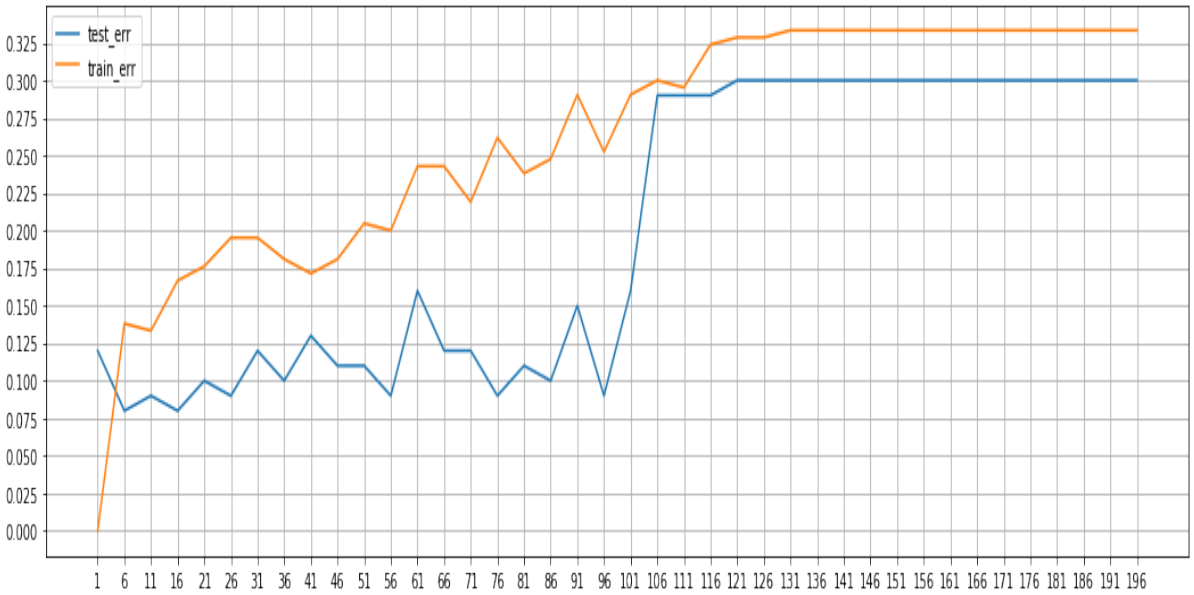
min test_err: 0.06000000000000005
best log10(p): 0.6

(d-i-c) find best k-value　[metric = 'chebyshev']



when k is: 16
min test_err is: 0.07999999999999996

(d-ii) find best k-value [metric = 'mahalanobis']



k = 6
min test_err 0.07999999999999996

	Metric	Best Test Error	Best K-value
d-i-A	Manhattan Distance	0.11	6
d-i-B	Manhattan Distance	0.06	log10(p)=0.6
d-i-C	Chebyshev	0.08	16
d-ii	mahalanobis	0.08	6

(e) change weight= 'distance'

euclidean	0.09999999999999998
minkowski	0.09999999999999998
chebyshev	0.10999999999999999

(f)
The lowest training error rate I got in this homework is 0 when k=1. However, when k=1, it is very likely to over fit to train set.
You can see the answer in the plot (c-ii).