## Homework2-INF552-Rahul Ethiraj

**1.) Time Series Classification**

**a)** **Importing data**

**b) Train test split data :**

x\_train\_df

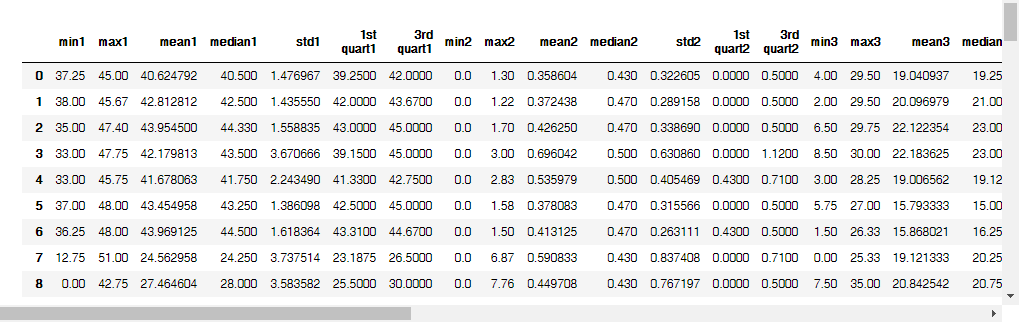
x\_test\_df

**c)**

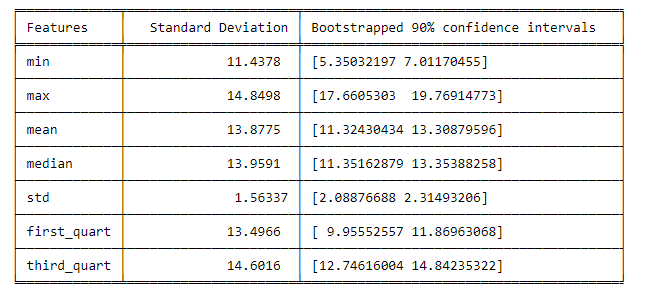
**(i) Feature Extraction:**

Minimum, maximum, mean, standard deviation, skewness, Fourier transforms of a range, periodicity, serial correlation, chaos, nonlinearity of time series are various methods of time series classification.

**(ii) Time-domain features:**



**(iii) Bootsrap confidence interval:**

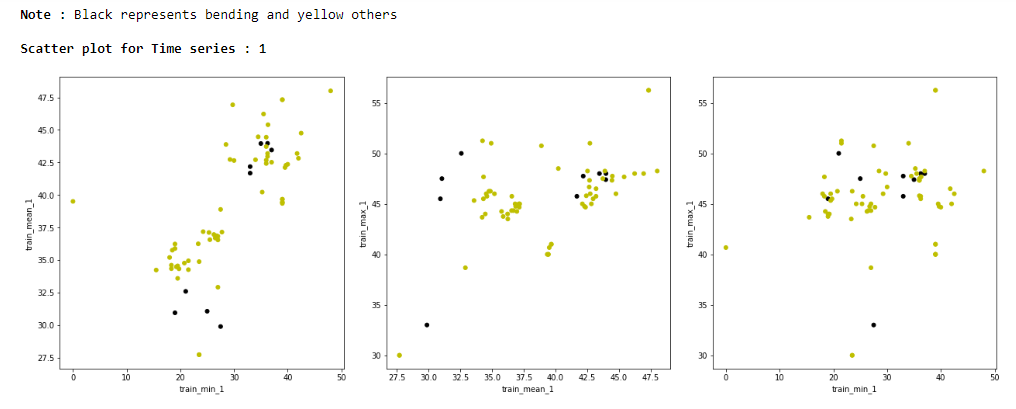


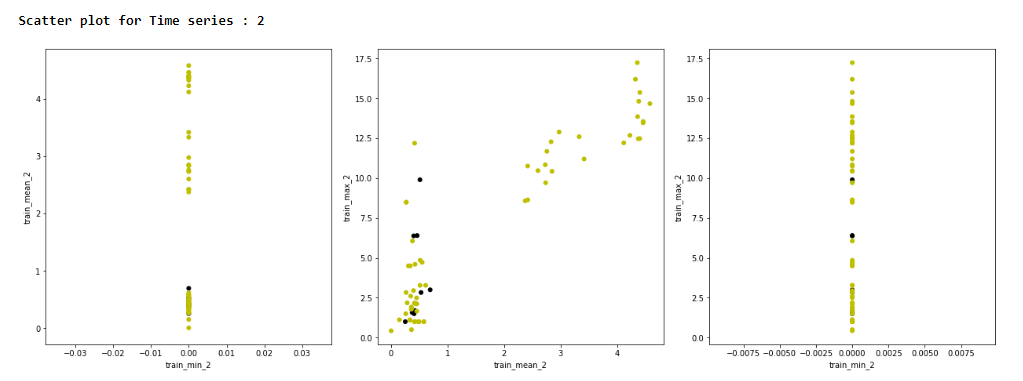
**(iv)**

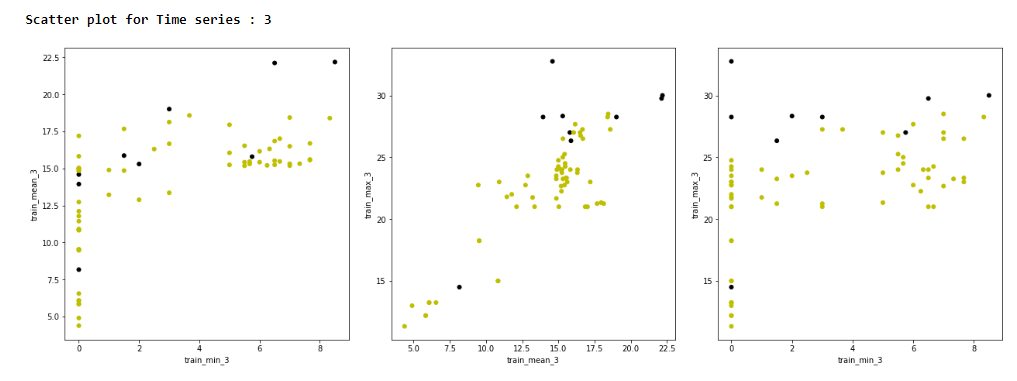
The time-domain feature ‘max’ contributed more to the target variable (classification index: bending vs others). I chose min, max and mean as the time-domain features and moved forward.

**(d) Binary Classification Using Logistic Regression**

**(i) Scatter plots of the features:**



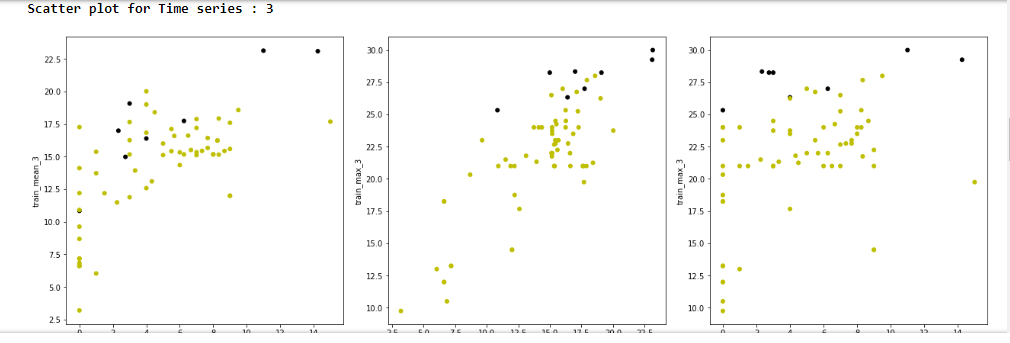
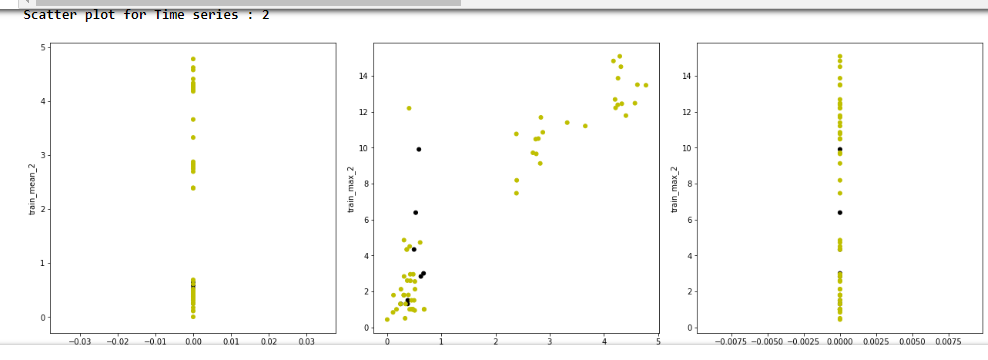
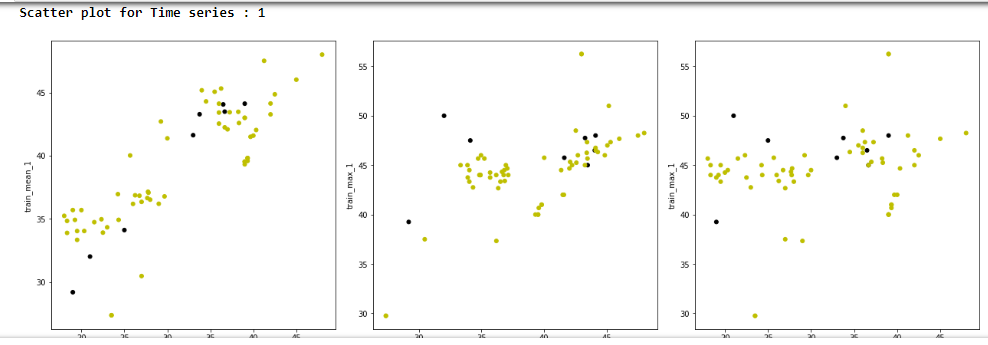




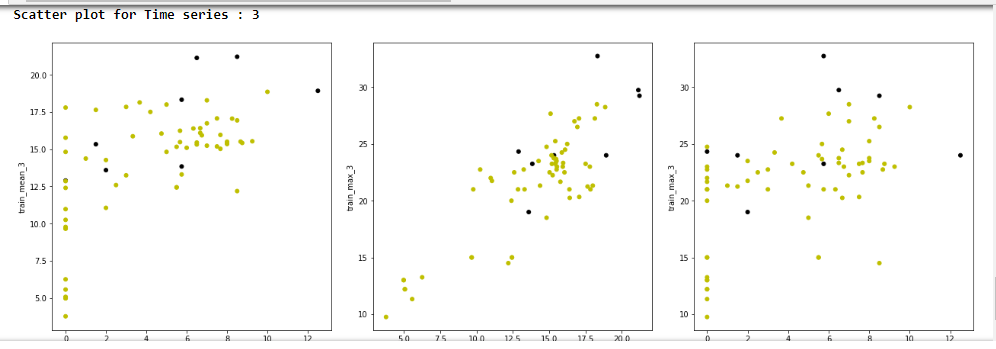
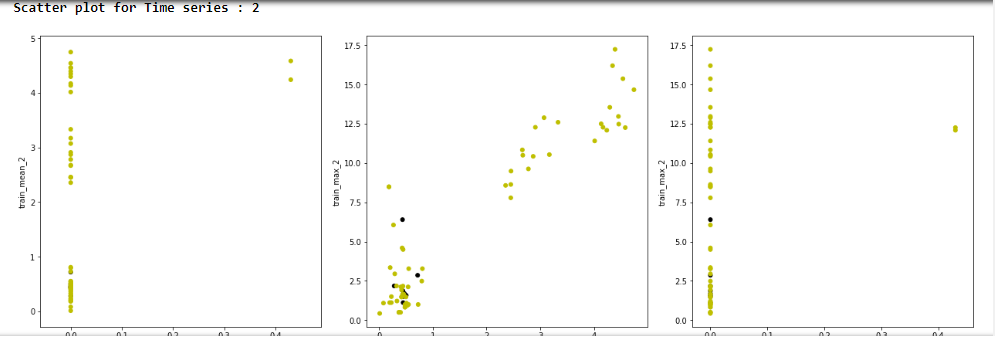
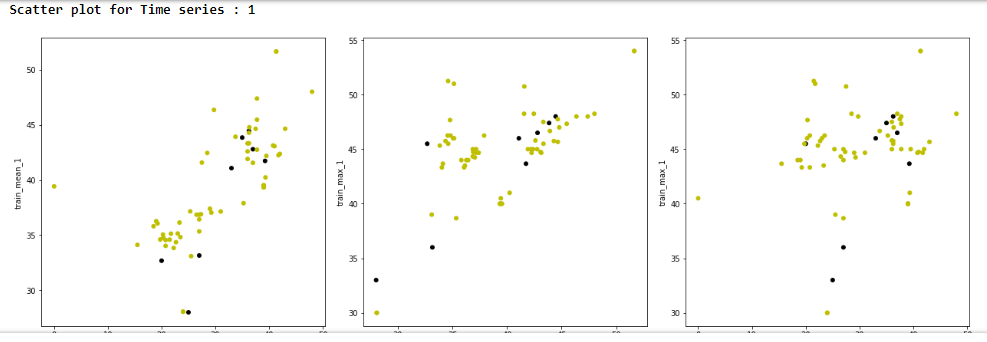
**(ii)**

**Dividing the train dataset into two parts, L=2**

**L=1:**



**L=2:**



Splitting the data is recommended here for this dataset as that modifies the features treating the whole data as two new datasets, which can lead to better training of the model.

This is evident from the following the results using the model.score() function.

**(iii) Breaking each time series at different sizes**

For l = {1,2, 3,…,20} time series.

IN L= 1

In part: 1

(69, 43)

Starting : 0

Ending : 480

Fold: 0, Score: 1.0

**p values for X\_train\_K :**

[8.50600808e-01 7.21466227e-01 3.38076083e-01 6.97191503e-01

1.06793874e-01 4.72624481e-01 7.46107796e-01 nan

2.57166794e-02 1.11614701e-01 4.28838813e-02 2.25191771e-01

2.71620102e-02 9.19249113e-03 3.94769109e-01 1.66112842e-02

2.13167491e-04 1.74284208e-02 1.09156965e-01 4.36611947e-02

7.79064203e-03 nan 1.24483775e-01 6.09470586e-01

6.09981461e-02 6.97956688e-01 1.23017963e-01 8.60829278e-02

6.34846753e-31 1.12626733e-10 1.18576815e-12 4.49172603e-10

4.04163452e-01 7.90099117e-12 7.39335177e-09 8.02265410e-01

6.83797960e-02 1.69673886e-01 5.90469738e-02 4.33482249e-01

5.86908338e-02 3.56414347e-02]

Fold: 1, Score: 1.0

**p values for X\_train\_K :**

[5.61643838e-01 8.93587320e-01 8.92314287e-01 7.43788750e-01

6.74439422e-01 8.71427760e-01 8.07071376e-01 nan

2.55660879e-02 1.01210198e-03 3.55548048e-02 1.37643358e-01

5.19466052e-02 1.02614038e-02 4.37909670e-01 5.86476646e-02

2.36426176e-02 6.18554711e-02 5.57474143e-01 7.49382105e-02

5.64079836e-02 nan 1.16856313e-01 6.35934313e-01

7.41741457e-02 6.03420558e-01 1.42802306e-01 7.73013447e-02

4.63281722e-47 5.70187249e-13 1.84516780e-08 9.14592233e-13

4.67350566e-01 2.55382756e-16 5.44996491e-10 8.02265410e-01

4.26656681e-02 1.02602109e-01 4.56038526e-02 2.81732824e-01

4.19379352e-02 2.55628787e-02]

Fold: 2, Score: 1.0

**p values for X\_train\_K :**

[8.89681752e-01 2.62979509e-01 8.78639428e-01 2.87454990e-01

4.25141950e-01 1.63043371e-01 1.92581005e-01 nan

6.27842250e-03 1.93137654e-03 1.00210185e-02 1.05265970e-01

1.15897675e-02 1.60158515e-03 2.84971175e-01 1.68961762e-01

1.33325428e-02 1.31703387e-01 1.12272400e-01 2.73614921e-01

8.40849746e-02 nan 4.74204717e-02 5.17158856e-01

3.02225347e-02 4.77759842e-01 3.02362456e-02 2.62503441e-02

1.27310593e-22 6.52854551e-08 1.36949592e-09 2.95815730e-07

3.30813319e-01 1.70667955e-08 1.56687498e-06 8.02265410e-01

4.11077377e-02 1.70806940e-01 3.80312888e-02 3.35961277e-01

3.85617885e-02 2.02754272e-02]

Fold: 3, Score: 1.0

**p values for X\_train\_K :**

[7.34863443e-01 5.67661637e-01 9.83285393e-01 5.44475162e-01

8.72003636e-01 6.36480762e-01 4.43187857e-01 nan

1.36136740e-02 3.07795012e-04 2.54393247e-02 1.07052713e-01

3.72490423e-02 3.32238148e-03 8.13566634e-01 1.09774392e-01

7.48756134e-03 1.12548215e-01 2.27890851e-01 1.99114876e-01

5.99090105e-02 nan 7.48459875e-02 9.61116321e-01

3.98339988e-02 5.59814442e-01 4.55118019e-02 4.81204514e-02

7.21386930e-29 1.91031023e-08 7.12069027e-10 9.88873224e-08

8.10441535e-01 4.57362630e-10 2.28507947e-06 8.02265410e-01

4.64123648e-02 1.95497077e-02 4.97780064e-02 2.72383878e-01

5.63427206e-02 2.37107143e-02]

Fold: 4, Score: 1.0

**p values for X\_train\_K :**

[9.15594845e-01 3.31906285e-01 9.55786346e-01 3.41852073e-01

3.62091804e-01 2.21721444e-01 2.68085243e-01 nan

7.67888706e-03 3.94198665e-03 1.30223749e-02 1.07127471e-01

1.96734941e-02 2.02241726e-03 2.62024422e-01 3.01977043e-01

1.00999486e-02 3.29513157e-01 9.37273484e-02 5.41602127e-01

1.40035490e-01 nan 6.62121987e-02 9.76627494e-01

3.17478736e-02 6.04854512e-01 5.33885342e-02 3.60199754e-02

9.22256044e-21 7.70861092e-08 5.64703737e-10 3.00298952e-07

4.12664527e-01 2.09612864e-08 1.80436965e-06 nan

3.58441417e-02 1.73869241e-01 3.33648574e-02 3.02241446e-01

3.94896862e-02 1.39319173e-02]

Final Score for L= 1 : 1.0

======================================================================

======================================================================

IN L= 2

In part: 1

(69, 43)

Starting : 0

Ending : 240

Fold: 0, Score: 1.0

**p values for X\_train\_K :**

[9.06655449e-01 7.57295333e-01 8.15301500e-01 4.49291632e-01

3.21144067e-01 5.64848252e-01 8.74419349e-01 nan

1.80750537e-02 9.67663108e-03 2.67035095e-02 1.82260970e-01

1.43092575e-02 6.73220066e-03 1.43677531e-01 3.55619290e-02

1.07130992e-02 3.52254271e-02 3.17449719e-01 5.79419803e-02

3.13377119e-02 8.02265410e-01 1.05101218e-01 6.68637084e-01

6.91098107e-02 6.04872028e-01 1.10683351e-01 7.00443255e-02

9.66173621e-23 1.73267538e-11 1.37374694e-10 5.93428862e-11

2.99685078e-01 1.53922297e-11 5.06072199e-11 8.02265410e-01

7.50205675e-02 2.24114643e-01 8.32587970e-02 4.20447944e-01

5.03444919e-02 6.26987373e-02]

Fold: 1, Score: 1.0

**p values for X\_train\_K :**

[9.86657167e-01 8.58220294e-01 9.50331243e-01 6.89307646e-01

5.17882873e-01 8.25661343e-01 8.64501020e-01 nan

1.25480467e-02 5.87772636e-04 1.74292877e-02 1.41841654e-01

1.70421842e-02 5.89467242e-03 4.19463298e-01 4.08601824e-02

1.03791721e-02 3.74529998e-02 5.45700068e-02 2.33799739e-01

4.53790852e-03 7.23232351e-01 9.80476880e-02 4.02145605e-01

7.02862000e-02 5.78336566e-01 9.36779324e-02 7.54987237e-02

1.51014206e-29 2.13243076e-12 5.57132492e-11 5.30130391e-12

7.58139278e-01 8.65305772e-14 5.99588527e-11 6.59642350e-01

4.32752947e-02 1.48011691e-02 5.75710990e-02 2.86738179e-01

4.29580270e-02 2.75684179e-02]

Fold: 2, Score: 1.0

**p values for X\_train\_K :**

[4.80384028e-01 4.71307035e-01 8.19991983e-01 2.70398931e-01

1.07889138e-01 2.94186324e-01 9.75182776e-01 nan

2.23791852e-02 1.13160231e-02 2.66120303e-02 2.32005394e-01

2.20447569e-02 9.96301129e-03 4.78938789e-01 2.83231881e-01

1.65555215e-02 3.11843872e-01 2.92881863e-02 9.99173227e-01

6.26438792e-02 7.23232351e-01 1.71147783e-01 9.65549418e-01

9.97710605e-02 8.36210936e-01 1.86416338e-01 1.28830523e-01

1.24468347e-24 6.03023138e-11 6.01518774e-09 2.06829775e-10

7.44998285e-01 1.98384754e-12 3.12384170e-09 6.59642350e-01

9.04160448e-02 4.35196938e-01 9.57611407e-02 5.22603901e-01

6.28019457e-02 7.07344521e-02]

Fold: 3, Score: 1.0

**p values for X\_train\_K :**

[7.33954838e-01 8.83695615e-01 4.01517983e-01 4.35416400e-01

4.31364431e-02 5.41428551e-01 5.12632886e-01 nan

1.22780177e-02 3.08690688e-02 1.72115119e-02 1.73965165e-01

1.73370226e-02 3.70702432e-03 8.73385349e-01 6.67017047e-02

9.16523666e-04 6.78167718e-02 8.56655114e-03 4.95957192e-01

5.45129683e-03 7.23232351e-01 1.17217294e-01 8.87904994e-01

6.56532371e-02 7.69961326e-01 7.33391612e-02 8.31280771e-02

5.86133697e-17 1.33024143e-10 5.20442922e-12 2.57676463e-10

1.04374938e-01 3.17852763e-10 1.90801928e-10 7.17139501e-01

7.14819460e-02 3.36357729e-01 5.75205842e-02 5.28408090e-01

4.39723958e-02 4.38575479e-02]

Fold: 4, Score: 1.0

**p values for X\_train\_K :**

[3.87892400e-01 5.46515791e-01 9.65591501e-01 2.52743601e-01

6.87198191e-02 3.10192901e-01 8.63355523e-01 nan

1.06113296e-02 1.07231732e-02 1.96442551e-02 1.59790263e-01

1.12401858e-02 3.13525797e-03 8.44538179e-01 4.94189721e-02

3.03935708e-03 4.86483144e-02 2.94200089e-02 2.84114595e-01

6.04356087e-03 7.88925395e-01 6.03372583e-02 2.86584636e-01

3.50074438e-02 5.89098356e-01 5.10128067e-02 4.20061585e-02

8.31493402e-29 1.02953317e-12 1.39799316e-11 3.48062001e-12

2.17797482e-01 1.31071506e-12 2.09250607e-12 6.37772926e-01

4.31024966e-02 1.70367150e-01 4.53232265e-02 4.11178434e-01

2.45193686e-02 2.81385612e-02]

IN L= 2

In part: 2

(69, 43)

Starting : 240

Ending : 480

Fold: 0, Score: 0.9285714285714286

**p values for X\_train\_K :**

[7.57585286e-01 1.30282872e-01 9.90416137e-02 8.59149788e-02

7.51591963e-01 1.10978404e-01 1.73947427e-01 8.02265410e-01

1.95275342e-02 3.24596874e-04 2.17968216e-02 1.49460407e-01

3.81699108e-02 1.22271932e-02 2.58654925e-01 1.99231044e-01

1.06480872e-01 1.50053899e-01 5.64155042e-01 3.13239831e-01

1.52466478e-01 nan 1.02366977e-01 4.43913241e-01

6.20688367e-02 5.49901860e-01 1.33614630e-01 6.28051421e-02

4.63666725e-12 7.67540850e-06 1.22039626e-04 2.34281630e-05

9.45016542e-01 1.07082899e-06 3.75754067e-05 6.59642350e-01

6.03501766e-02 1.95371670e-02 6.77705910e-02 2.67995312e-01

1.10845581e-01 2.77637582e-02]

Fold: 1, Score: 1.0

**p values for X\_train\_K :**

[3.75546669e-01 5.14963366e-01 1.33039786e-01 4.68825669e-01

3.98762090e-02 7.32858644e-01 2.84615797e-01 8.02265410e-01

7.45417350e-03 1.04715855e-06 1.02276367e-02 5.42148984e-02

3.36217996e-02 3.24913916e-03 1.65887071e-01 4.95866031e-01

4.51204709e-01 3.50706323e-01 6.37362752e-01 6.31356463e-01

3.96398383e-01 8.02265410e-01 3.91117568e-02 3.96070973e-02

2.69754866e-02 3.62654992e-01 4.49844537e-02 1.69770768e-02

2.48175822e-16 2.03441369e-06 3.71060791e-03 2.39334470e-06

3.02910975e-01 4.62438545e-08 5.03062944e-05 6.12419504e-01

2.11737224e-02 2.09180482e-03 1.90699940e-02 1.53354448e-01

6.62930905e-02 9.04107895e-03]

Fold: 2, Score: 0.9285714285714286

**p values for X\_train\_K :**

[6.64957224e-01 5.30372019e-01 3.59196965e-01 3.51054067e-01

4.10011005e-01 5.86822723e-01 4.97558279e-01 7.23232351e-01

1.40529402e-02 1.29654482e-04 2.85713612e-02 8.28116673e-02

7.93733808e-02 3.37010858e-03 4.27930710e-01 1.53760575e-01

1.24465148e-01 1.50788550e-01 9.30804305e-01 1.20781503e-01

1.75733205e-01 8.02265410e-01 5.44238353e-02 2.76959019e-01

4.03118338e-02 4.09009564e-01 5.18471041e-02 2.88096295e-02

1.48491187e-21 5.78701221e-07 1.29915468e-04 1.77853816e-06

4.89384414e-01 1.39493028e-08 7.91220469e-06 7.23232351e-01

4.50092915e-02 2.47744922e-03 4.27573941e-02 2.13479071e-01

1.06234406e-01 2.17752928e-02]

Fold: 3, Score: 0.9285714285714286

**p values for X\_train\_K :**

[9.17280241e-01 4.72857013e-01 3.67459924e-01 3.19902245e-01

8.15567024e-01 4.38994998e-01 5.67380541e-01 7.23232351e-01

1.23088747e-02 1.04727915e-04 1.91750283e-02 1.07449698e-01

3.79841780e-02 5.87491519e-03 2.93044855e-01 2.65816421e-02

1.35795717e-02 2.26966764e-02 4.02987128e-01 3.45149589e-02

1.68907673e-02 8.02265410e-01 7.57617648e-02 3.41352066e-01

4.12185183e-02 4.98684270e-01 9.23819596e-02 3.96907114e-02

4.15752921e-23 2.49563718e-08 1.26925366e-06 1.39322735e-07

9.21658503e-01 2.10628080e-09 1.68511797e-07 6.59642350e-01

3.51012906e-02 5.98980110e-03 2.80085392e-02 2.55999638e-01

5.31129428e-02 1.65165563e-02]

Fold: 4, Score: 0.9230769230769231

**p values for X\_train\_K :**

[7.43084557e-01 7.27702970e-01 4.80362188e-01 5.45620319e-01

5.75010937e-01 7.16394172e-01 7.52237102e-01 7.04989460e-01

1.19323507e-02 1.45762030e-05 1.44226506e-02 8.89040600e-02

4.99342933e-02 5.01723805e-03 2.99269301e-01 5.64586012e-02

2.71675509e-02 4.66859795e-02 5.63581870e-01 9.72811710e-02

3.37680016e-02 7.88925395e-01 5.50478571e-02 2.04560286e-01

3.57539536e-02 4.10414355e-01 5.16605651e-02 3.00740908e-02

1.07529015e-26 4.40713376e-10 1.48479790e-06 4.27250018e-09

4.79194561e-01 2.78354679e-12 2.57379348e-08 5.88092328e-01

2.09241397e-02 1.54025009e-03 2.09775032e-02 1.69423982e-01

4.57883959e-02 9.90754586e-03]

Final Score for L= 2 : 0.9417582417582417

Similarly, all the values for L=3,4…,20 were observed, by iterating the loops for different sizes.

**Final Score for L= 1 : 1.0**

Final Score for L= 2 : 0.9417582417582417

Final Score for L= 3 : 0.9417582417582417

Final Score for L= 4 : 0.9571428571428573

Final Score for L= 5 : 0.9560439560439562

Final Score for L= 6 : 0.9428571428571428

Final Score for L= 7 : 0.9571428571428571

Final Score for L= 8 : 0.9571428571428571

Final Score for L= 9 : 0.9417582417582417

Final Score for L= 10 : 0.9417582417582417

Final Score for L= 11 : 0.956043956043956

Final Score for L= 12 : 0.9571428571428571

Final Score for L= 13 : 0.956043956043956

Final Score for L= 14 : 0.9417582417582417

Final Score for L= 15 : 0.9417582417582417

Final Score for L= 16 : 0.9703296703296704

Final Score for L= 17 : 0.9417582417582417

Final Score for L= 18 : 0.9417582417582417

Final Score for L= 19 : 0.956043956043956

Final Score for L= 20 : 0.9714285714285715

**Best L : 1**

**Pruned number of features : 5**

Error inverse : 1.0

Features list : [False False False False False False True False False False False False

False False False False False True False False False False False False

False False False False False True True False False True False False

False False False False False False]

For each value of L the parts of data passed to the model were different, which resulted in different set of features to be pruned each time by the model.

I used Recursive Feature Elimination method for pruning the full 42 features.

Cross-validation can be done in two different ways: the wrong way and the right way. The only difference is that in the former, we perform the variable selection before cross validation using all the samples. In the latter, we perform the variable selection within a K-fold cross validation loop each and every time.

## Re-fitting the model using pruned set of features by RFE

**Regression Coef:**

[[-0.30272987 -0.22798283 0.78625013 0.42167481 -0.51325141]]

**p values for x\_train :**

[3.90271388e-01 2.84100194e-01 5.46927780e-08 8.13227556e-06

2.18370678e-08]

**Final Score for L = 1 : 0.927536231884058**

Printing all feature p-values :

=======================================================================

**Printing for third\_quart1**

OLS Regression Results

==============================================================================

Dep. Variable: target R-squared: 0.023

Model: OLS Adj. R-squared: 0.008

Method: Least Squares F-statistic: 1.581

Date: Mon, 02 Jul 2018 Prob (F-statistic): 0.213

Time: 20:17:48 Log-Likelihood: -22.008

No. Observations: 69 AIC: 48.02

Df Residuals: 67 BIC: 52.48

Df Model: 1

Covariance Type: nonrobust

===================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------

Intercept 0.6059 0.380 1.593 0.116 -0.153 1.365

I(third\_quart1) -0.0119 0.009 -1.257 0.213 -0.031 0.007

==============================================================================

Omnibus: 39.023 Durbin-Watson: 0.134

Prob(Omnibus): 0.000 Jarque-Bera (JB): 76.067

Skew: 2.150 Prob(JB): 3.04e-17

Kurtosis: 5.822 Cond. No. 377.

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

=======================================================================

**Printing for median3**

OLS Regression Results

==============================================================================

Dep. Variable: target R-squared: 0.011

Model: OLS Adj. R-squared: -0.004

Method: Least Squares F-statistic: 0.7142

Date: Mon, 02 Jul 2018 Prob (F-statistic): 0.401

Time: 20:17:48 Log-Likelihood: -22.447

No. Observations: 69 AIC: 48.89

Df Residuals: 67 BIC: 53.36

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept 0.0269 0.129 0.208 0.836 -0.231 0.285

I(median3) 0.0073 0.009 0.845 0.401 -0.010 0.025

==============================================================================

Omnibus: 39.247 Durbin-Watson: 0.145

Prob(Omnibus): 0.000 Jarque-Bera (JB): 76.573

Skew: 2.169 Prob(JB): 2.36e-17

Kurtosis: 5.795 Cond. No. 47.2

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

=======================================================================

**Printing for max5**

OLS Regression Results

==============================================================================

Dep. Variable: target R-squared: 0.206

Model: OLS Adj. R-squared: 0.194

Method: Least Squares F-statistic: 17.34

Date: Mon, 02 Jul 2018 Prob (F-statistic): 9.14e-05

Time: 20:17:48 Log-Likelihood: -14.873

No. Observations: 69 AIC: 33.75

Df Residuals: 67 BIC: 38.21

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept -0.4415 0.142 -3.106 0.003 -0.725 -0.158

I(max5) 0.0291 0.007 4.164 0.000 0.015 0.043

==============================================================================

Omnibus: 30.715 Durbin-Watson: 0.467

Prob(Omnibus): 0.000 Jarque-Bera (JB): 49.943

Skew: 1.773 Prob(JB): 1.43e-11

Kurtosis: 5.190 Cond. No. 79.2

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

=======================================================================

**Printing for first\_quart5**

OLS Regression Results

==============================================================================

Dep. Variable: target R-squared: 0.236

Model: OLS Adj. R-squared: 0.224

Method: Least Squares F-statistic: 20.68

Date: Mon, 02 Jul 2018 Prob (F-statistic): 2.34e-05

Time: 20:17:48 Log-Likelihood: -13.533

No. Observations: 69 AIC: 31.07

Df Residuals: 67 BIC: 35.53

Df Model: 1

Covariance Type: nonrobust

===================================================================================

coef std err t P>|t| [0.025 0.975]

-----------------------------------------------------------------------------------

Intercept -0.3143 0.104 -3.016 0.004 -0.522 -0.106

I(first\_quart5) 0.0312 0.007 4.547 0.000 0.018 0.045

==============================================================================

Omnibus: 42.084 Durbin-Watson: 0.594

Prob(Omnibus): 0.000 Jarque-Bera (JB): 98.497

Skew: 2.120 Prob(JB): 4.09e-22

Kurtosis: 7.034 Cond. No. 44.1

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

=======================================================================

**Printing for mean5**

OLS Regression Results

==============================================================================

Dep. Variable: target R-squared: 0.253

Model: OLS Adj. R-squared: 0.241

Method: Least Squares F-statistic: 22.64

Date: Mon, 02 Jul 2018 Prob (F-statistic): 1.08e-05

Time: 20:17:48 Log-Likelihood: -12.767

No. Observations: 69 AIC: 29.53

Df Residuals: 67 BIC: 34.00

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept -0.3839 0.114 -3.374 0.001 -0.611 -0.157

I(mean5) 0.0329 0.007 4.759 0.000 0.019 0.047

==============================================================================

Omnibus: 39.290 Durbin-Watson: 0.636

Prob(Omnibus): 0.000 Jarque-Bera (JB): 84.668

Skew: 2.019 Prob(JB): 4.12e-19

Kurtosis: 6.626 Cond. No. 52.9

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

**iv) Regression coefficients and p-values for L=1, Number of features=5 using 5-fold CV:**

Regression Coef:

[[-0.30272987 -0.22798283 0.78625013 0.42167481 -0.51325141]]

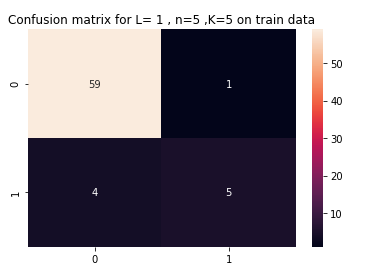
p values for x\_train :

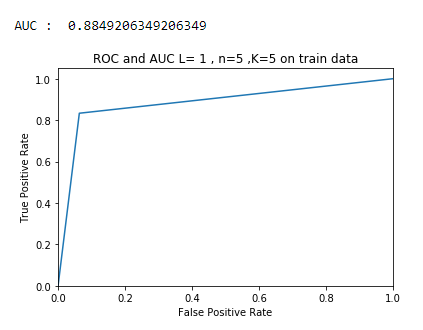
[3.90271388e-01 2.84100194e-01 5.46927780e-08 8.13227556e-06

2.18370678e-08]

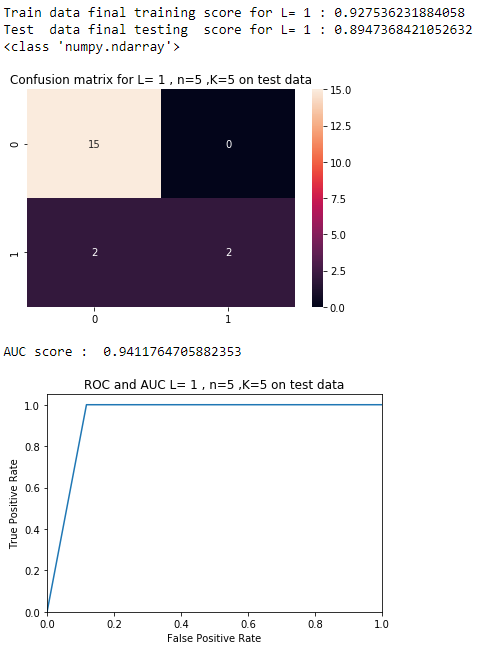
Final Score for L = 1 : 0.927536231884058

**Confusion matrix, ROC & AUC:**





**v)**



Test score is very much closer to the train score, which implies that the model has performed well on the testing data.

## vi)

## For training:

We had 5 misclassifications in total.

An AUC score of 1 is a result of 0% overlapping degree, and a score of about 0.89 signifies there is an

overlap in the data.

Classes seem to be not well-separated to cause instability in calculating logistic regression parameters.

**For testing:**

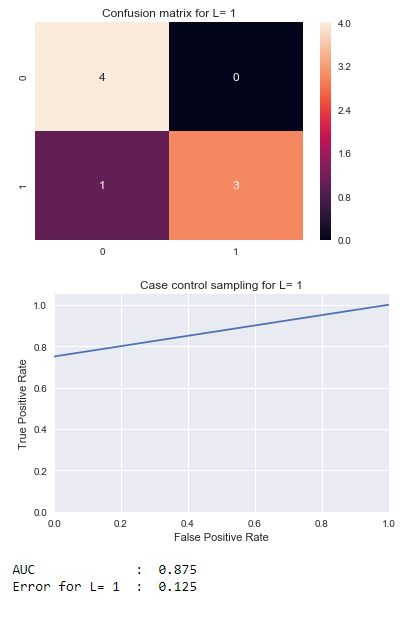
We have 2 misclassifications.

An AUC score of 1 is a result of 0% overlapping degree, and a score of about 0.94 signifies there is an

overlap in the data.

Classes seem to be not well-separated to cause instability in calculating logistic regression parameters.

**vii) Case control sampling:**



There exists imbalance in the data and therefore under-sampled the majority class(0).

Both the classes are now well-balanced (1:1 ratio)

**(e) Binary classification using L1-penalized logistic regression:**

The code tries 10 values of C and cross validates on them as well as on the train set. Following are the results:

IN L= 1

In part: 1

Starting : 0

Ending : 480

Fold: 0, Score: 0.9285714285714286

**Regression coef-values :**

[[ 0. 0. 0. 0. 0. -0.27275373

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

8.6146879 0. 0. 0. 0. 0.

0. 0. 0. 0. -2.94195997 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 1, Score: 0.9285714285714286

**Regression coef-values :**

[[ 0. -3.18615869 0. 0. 0. 0.

-0.15209024 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

-0.11191496 0. 0. 0. 2.25442505 0.

5.77754826 0. 0. 0. 0. 0.

0. 0. -3.47295174 0. 0. 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 2, Score: 0.8571428571428571

**Regression coef-values :**

[[-5.17993223 -1.23371418 0. 0. 0.59100634 -7.25808031

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 3.85567301 0.

0. 0. 0. 0. 0. 0.

-0.06565019 0. -5.69097591 0. 24.63381078 0.

6.3476103 0. 0. 0. 0. 0.

0. 0. 0. 0. -9.98324396 0. ]]

C : [166.81005372]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 3, Score: 1.0

**Regression coef-values :**

[[ 0. -3.65947558 0. 0. 0. 0.

-0.97927771 0. 0. -4.37417825 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

-2.49172622 0. 0. 0. 10.65660604 0.

14.82580117 0. -1.00592686 0. 0. 0.

0. 0. 0. -0.93694531 0. 0. ]]

C : [21.5443469]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 4, Score: 1.0

**Regression coef-values :**

[[ -4.69688181 0. 0. 0. 0.

0. -10.05972797 0. 0. -2.54805489

0. 0. 0. 0. 0.

0. 0. -0.11028776 0. 0.

-1.70744181 0. 0. 4.63845158 -1.50871159

0. 0. 0. 23.71731575 0.

18.22184272 0. 0. 0. 0.

0. 0. 0. 0. 0.

-9.12127452 0. ]]

C : [166.81005372]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

**Final Score for L= 1 : [0.9428571428571428]**

======================================================================

======================================================================

IN L= 2

In part: 1

Starting : 0

Ending : 240

Fold: 0, Score: 0.9285714285714286

**Regression coef-values** :

[[ 0. 0. 0. -3.2463717 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 2.33742012 0.

5.13832456 0. 0. 0. 0. 0.

0. 0. 0. 0. -4.23608703 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 1, Score: 0.9285714285714286

**Regression coef-values :**

[[ 0. 0. 0. -2.39728467 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 2.28244987 0.

5.61525796 0. 0. 0. 0. 0.

0. 0. 0. 0. -3.39501999 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 2, Score: 1.0

**Regression coef-values :**

[[ 0. 0. 0. -6.28830876 0. 0.

0. 0. 0. -1.944261 0. 0.

0. 0. 0. 0. 0. 0.

0. -0.74317906 0. 0. 0. 0.0820609

0. 0. 0. 0. 13.7467722 0.

12.0990617 0. 0. 0. 0. 0.

0. 0. 0. 0. -6.0920041 0. ]]

C : [21.5443469]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 3, Score: 1.0

**Regression coef-values :**

[[ 0. 0. 0. -0.69137297 0. -3.29486098

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. -2.66942252 0. 0. 0. 0.

-4.10828134 0. 0. 0. 11.2863585 0.

15.12561329 0. 0. 0. 0. 0.

0. 0. 0. 0. -2.84244273 0. ]]

C : [21.5443469]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 4, Score: 0.9230769230769231

**Regression coef-values :**

[[ 0. 0. 0. -1.91279038 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0.3881247 0.

7.95487675 0. 0. 0. 0. 0.

0. 0. 0. 0. -3.59724743 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

IN L= 2

In part: 2

Starting : 240

Ending : 480

Fold: 0, Score: 0.8571428571428571

**Regression coef-values :**

[[ -4.15478151 -4.25909276 0. 0. -22.50674674

0. -20.97308507 0. 0. 0.

0. -11.53023299 0. -1.25970788 -6.65882722

3.24874939 17.62642168 0. -9.84307051 1.83283607

0. 0. 0. -6.88174669 -9.73233496

23.58106466 0. 0. 3.19927128 5.85744322

12.45289477 2.16007223 -23.8365537 1.12149946 3.00484062

0. 0. 0. -11.06729863 0.

36.95331207 -22.56911476]]

C : [1291.54966501]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 1, Score: 0.9285714285714286

**Regression coef-values :**

[[-0.18691306 -0.34875786 0. -2.51811923 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 2.30999279 0.

3.19029898 0. 0. 1.86844195 0. 0.

0. 0. -0.34368656 0. 0. -4.02489707]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 2, Score: 0.8571428571428571

**Regression coef-values :**

[[ 0. 0. 0. -0.16472102 0.

0. -14.40755269 0. 0. 0.

0. 0. 0. -5.94692698 1.75875836

0. 1.36447388 0. -9.17500526 0.

0. 0. 0. 0. -0.14940421

0. -1.55173381 0. 0. 0.

16.17977678 0. 0. 2.34297872 0.

0. 0. 0. 0. -1.13876694

0. -0.8494328 ]]

C : [21.5443469]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 3, Score: 1.0

**Regression coef-values :**

[[ 0. 0. 0. 0. -2.64319536

0. -12.86437522 0. 0. 0.

0. 0. 0. 0. -2.69649671

0. 13.10929323 0. -6.43492774 0.

0. 0. 0. 0. -3.38680311

0. 0. 0. 0. 0.

13.58650063 0. -5.19641778 0. 0.

0. 0. 0. -7.55897776 0.

0. 0. ]]

C : [21.5443469]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

Fold: 4, Score: 0.8461538461538461

**Regression coef-values :**

[[ 0. 0. 0. -5.51950478 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 1.27924777 0.

0. 0. 0. 0. 0. 0.

-5.28792832 0. 0. 0. 6.13185612 0.

0.28815944 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]]

C : [2.7825594]

CS\_ : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

**Final Score for L= 2 : [0.9428571428571428, 0.8978021978021978]**

======================================================================

======================================================================

**Similarly, results were obtained for other values of L.**

**Final Score for L= 1 : [0.9428571428571428]**

Final Score for L= 2 : [0.8978021978021978]

Final Score for L= 3 : [0.9274725274725275]

Final Score for L= 4 : [0.856043956043956]

Final Score for L= 5 : [0.9142857142857143]

Final Score for L= 6 : [0.8703296703296702]

Final Score for L= 7 : [0.8846153846153847]

Final Score for L= 8 : [0.8846153846153847]

Final Score for L= 9 : [0.9142857142857143]

Final Score for L= 10 : [0.8846153846153847]

Final Score for L= 11 : [0.8703296703296705]

Final Score for L= 12 : [0.9285714285714285]

Final Score for L= 13 : [0.8549450549450549]

Final Score for L= 14 : [0.8857142857142858]

Final Score for L= 15 : [0.9]

Final Score for L= 16 : [0.8989010989010989]

Final Score for L= 17 : [0.856043956043956]

Final Score for L= 18 : [0.8714285714285716]

Final Score for L= 19 : [0.8989010989010989]

Final Score for L= 20 : [0.8857142857142858]

**Training errors:**

With p-values:

Final Score for L=1: **0.927536231884058**

With L-1 penalization:

Final Score for L=1: **0.9428571428571428**

Best L : **1**

**Regression coef-values :**

[[-3.41374304e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 -1.06322415e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 -1.11427729e-03 0.00000000e+00

0.00000000e+00 0.00000000e+00 2.21379049e+00 -4.34208948e+00

0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 1.27603509e+01 0.00000000e+00

0.00000000e+00 9.79191014e+00 0.00000000e+00 0.00000000e+00

0.00000000e+00 0.00000000e+00 0.00000000e+00 0.00000000e+00

-9.91925872e+00 0.00000000e+00]]

C : [21.5443469]

C\_s : [1.00000000e-04 7.74263683e-04 5.99484250e-03 4.64158883e-02

3.59381366e-01 2.78255940e+00 2.15443469e+01 1.66810054e+02

1.29154967e+03 1.00000000e+04]

**The L-1 penalization is slightly better in the accuracy and is easier to implement because of readymade libraries.**

**(f)**

**(i) Multi-class classification (The realistic case):**

**Final Score for L= 1 : 0.9056372549019608**

Final Score for L= 2 : 0.8487745098039217

Final Score for L= 3 : 0.7862745098039216

Final Score for L= 4 : 0.7571078431372549

Final Score for L= 5 : 0.732107843137255

Final Score for L= 6 : 0.7571078431372549

Final Score for L= 7 : 0.795343137254902

Final Score for L= 8 : 0.7848039215686274

Final Score for L= 9 : 0.757843137254902

Final Score for L= 10 : 0.7654411764705882

Final Score for L= 11 : 0.7688725490196078

Final Score for L= 12 : 0.7938725490196078

Final Score for L= 13 : 0.8041666666666668

Final Score for L= 14 : 0.7661764705882353

Final Score for L= 15 : 0.7446078431372549

Final Score for L= 16 : 0.7571078431372549

Final Score for L= 17 : 0.7661764705882353

Final Score for L= 18 : 0.6946078431372549

Final Score for L= 19 : 0.721078431372549

Final Score for L= 20 : 0.7370098039215687

**Best L : 1**

**Error inverse : 0.9056372549019608**

**Regression coef-values :**

[[ 0. 0. 0. 0. 0. -1.05255069

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 2.89333516 0. 0. 1.46603467 0.

2.9138691 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. -5.34272818 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. -0.04881261 0. 0. 0.

2.5877929 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 1.98994936 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]

[ 0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 3.23716979 0.

4.2369885 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0.

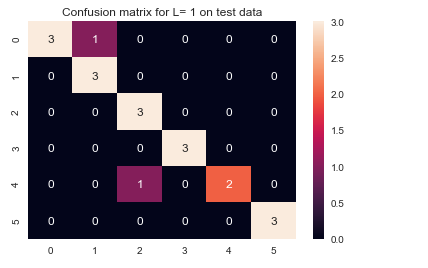
0. 0. 0. 0. 0. 0.

0. 0. 0. 0. 0. 0. ]]

**Train data final training score for L= 1 : 0.9056372549019608**

Test data final testing score for L= 1 : 0.8947368421052632

Test score is very much closer to the train score, which implies that the model has performed well on the testing data.



**ii) Naïve Bayes with Gaussian prior:**

**Final Score for L= 1 : 0.9132352941176471**

Final Score for L= 2 : 0.8563725490196077

Final Score for L= 3 : 0.7669117647058823

Final Score for L= 4 : 0.7377450980392156

Final Score for L= 5 : 0.758578431372549

Final Score for L= 6 : 0.7223039215686274

Final Score for L= 7 : 0.7397058823529411

Final Score for L= 8 : 0.6529411764705882

Final Score for L= 9 : 0.695343137254902

Final Score for L= 10 : 0.7017156862745099

Final Score for L= 11 : 0.6813725490196079

Final Score for L= 12 : 0.7071078431372548

Final Score for L= 13 : 0.6794117647058824

Final Score for L= 14 : 0.6502450980392156

Final Score for L= 15 : 0.6884803921568627

Final Score for L= 16 : 0.7036764705882353

Final Score for L= 17 : 0.6419117647058823

Final Score for L= 18 : 0.6115196078431373

Final Score for L= 19 : 0.5995098039215687

Final Score for L= 20 : 0.6024509803921569

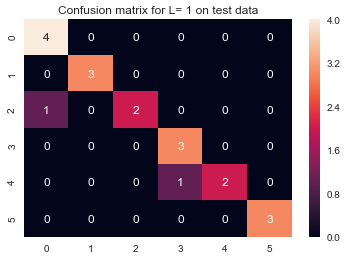
**Best L : 1**

**Error inverse : 0.9132352941176471**

**Train data final training score for L= 1 : 0.9132352941176471**

Test data final testing score for L= 1 : 0.8947368421052632

**Test score is very much closer to the train score, which implies that the model has performed well on the testing data.**



## Naïve Bayes with Multinomial prior:

**Final Score for L= 1 : 0.7134803921568629**

Final Score for L= 2 : 0.6980392156862745

Final Score for L= 3 : 0.6911764705882354

Final Score for L= 4 : 0.6294117647058823

Final Score for L= 5 : 0.6676470588235295

Final Score for L= 6 : 0.6642156862745099

Final Score for L= 7 : 0.5995098039215686

Final Score for L= 8 : 0.6127450980392156

Final Score for L= 9 : 0.5634803921568629

Final Score for L= 10 : 0.5482843137254902

Final Score for L= 11 : 0.6044117647058823

Final Score for L= 12 : 0.5987745098039216

Final Score for L= 13 : 0.6051470588235295

Final Score for L= 14 : 0.5468137254901961

Final Score for L= 15 : 0.5835784313725491

Final Score for L= 16 : 0.5218137254901961

Final Score for L= 17 : 0.6392156862745098

Final Score for L= 18 : 0.535049019607843

Final Score for L= 19 : 0.6169117647058824

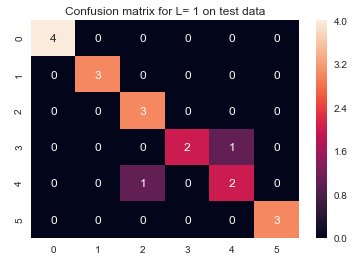
Final Score for L= 20 : 0.6002450980392157

Best L : 1

Error inverse : 0.7169117647058824

**Train data final training score for L= 1 : 0.7134803921568629**

Test data final testing score for L= 1 : 0.736842105263158

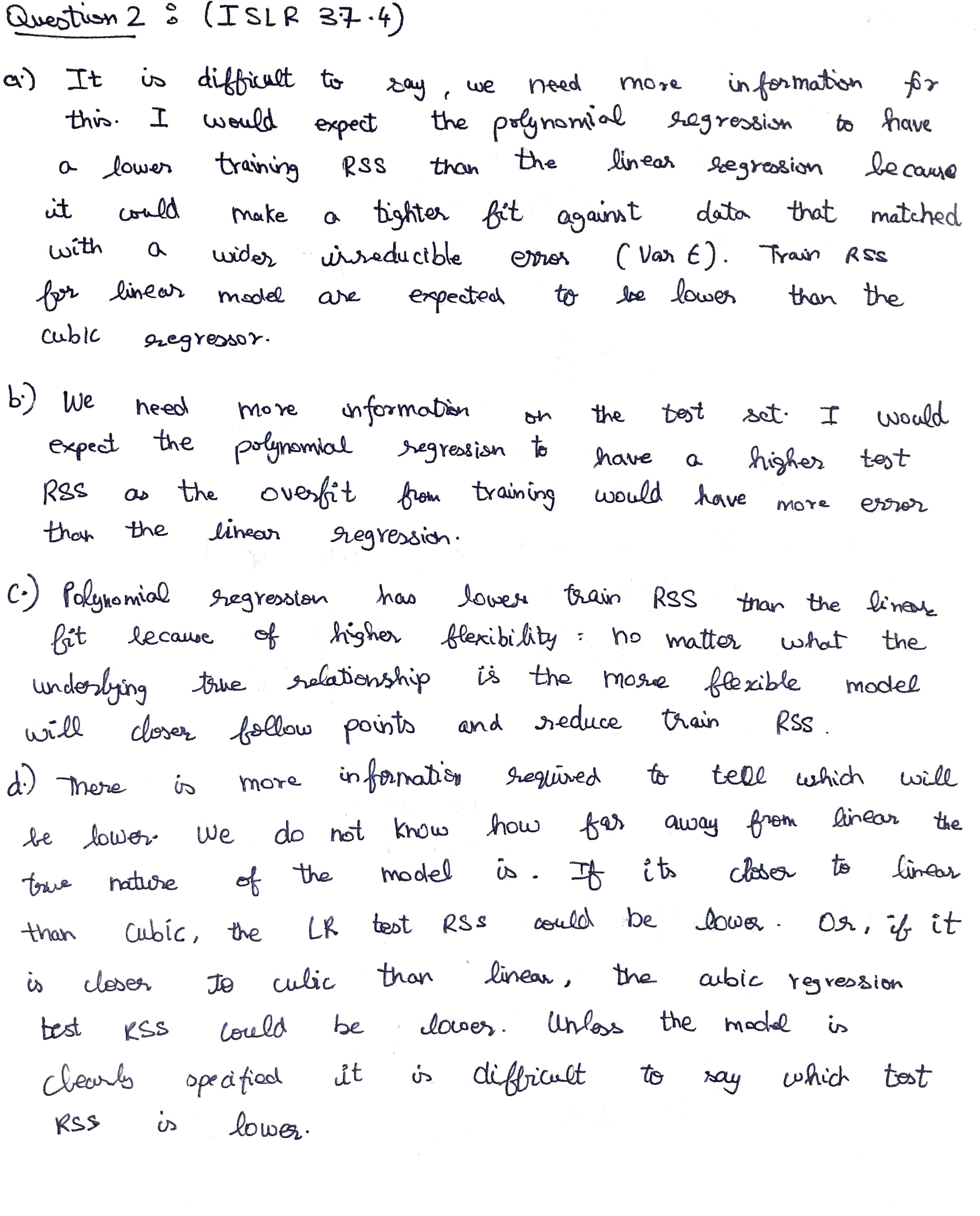


**iii)**

Naïve Bayes with Gaussian prior performed best with almost a 91% accuracy of predictions.

This method is better for multi-class classification in this problem.

## 2.) ISLR 3.7.4

****

## 3.) ISLR 4.7.3 & 4.) ISLR 4.7.7:

## 

## 