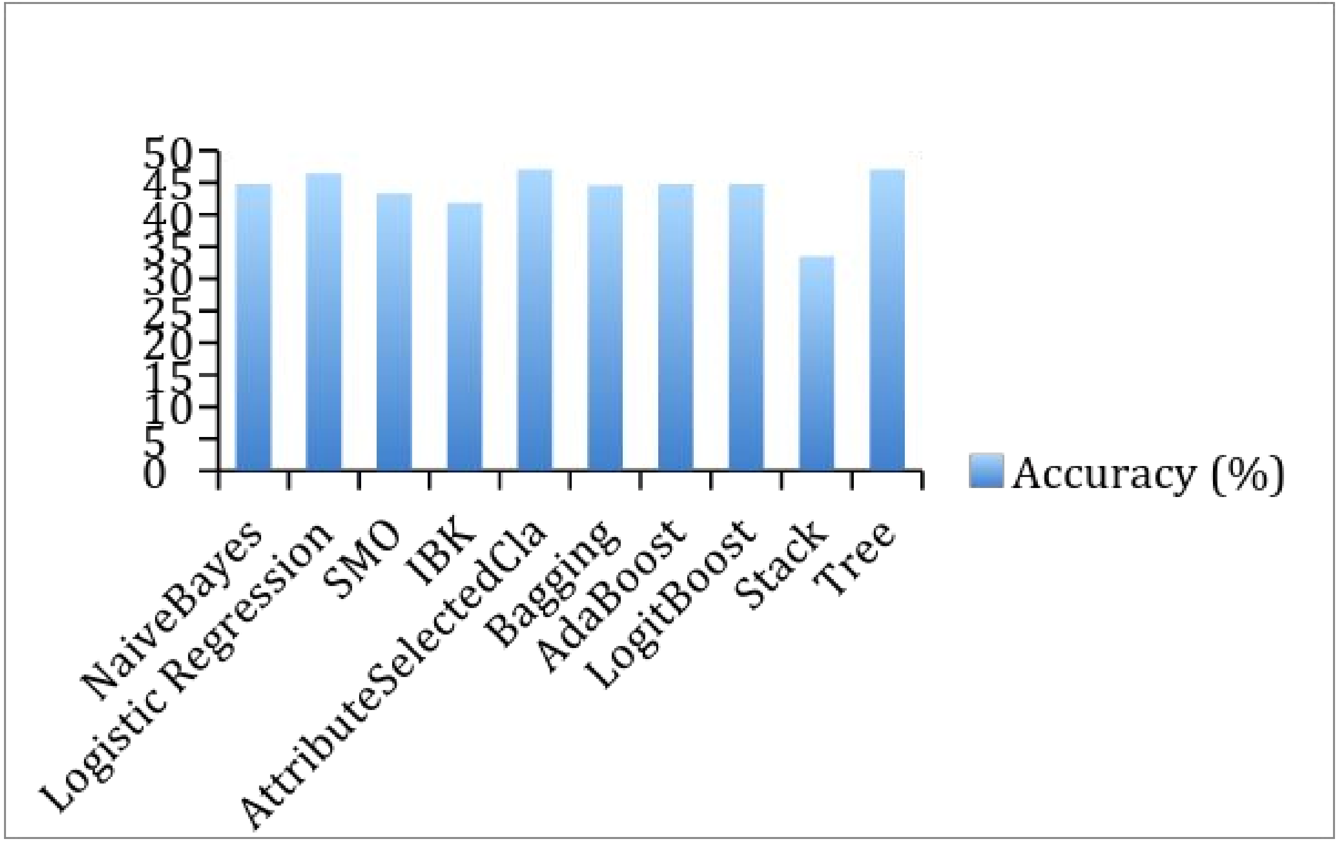
Predicting Students' First-year Success at UC Santa Cruz

Hairong Wu Kevin Doyle Connor McNeill

Table1 Best performance of different methods on 10-fold cross-validation

Methods on 10-fold cross-validation	Accuracy (%)
NaiveBayes	44.78
Logistic Regression	46.38
SMO (RBFkernel)	43.32
IBK (kNN=9, LinearNNSearch)	41.86
AttributeSelectedClassifier (LMT)	47.03
Bagging-RandomForest	44.61
AdaBoostM-HoeffdingTree	44.78
LogitBoost-DecisionStump	44.81
Stack-DecisionTable	33.51
Tree-LMT	47.03

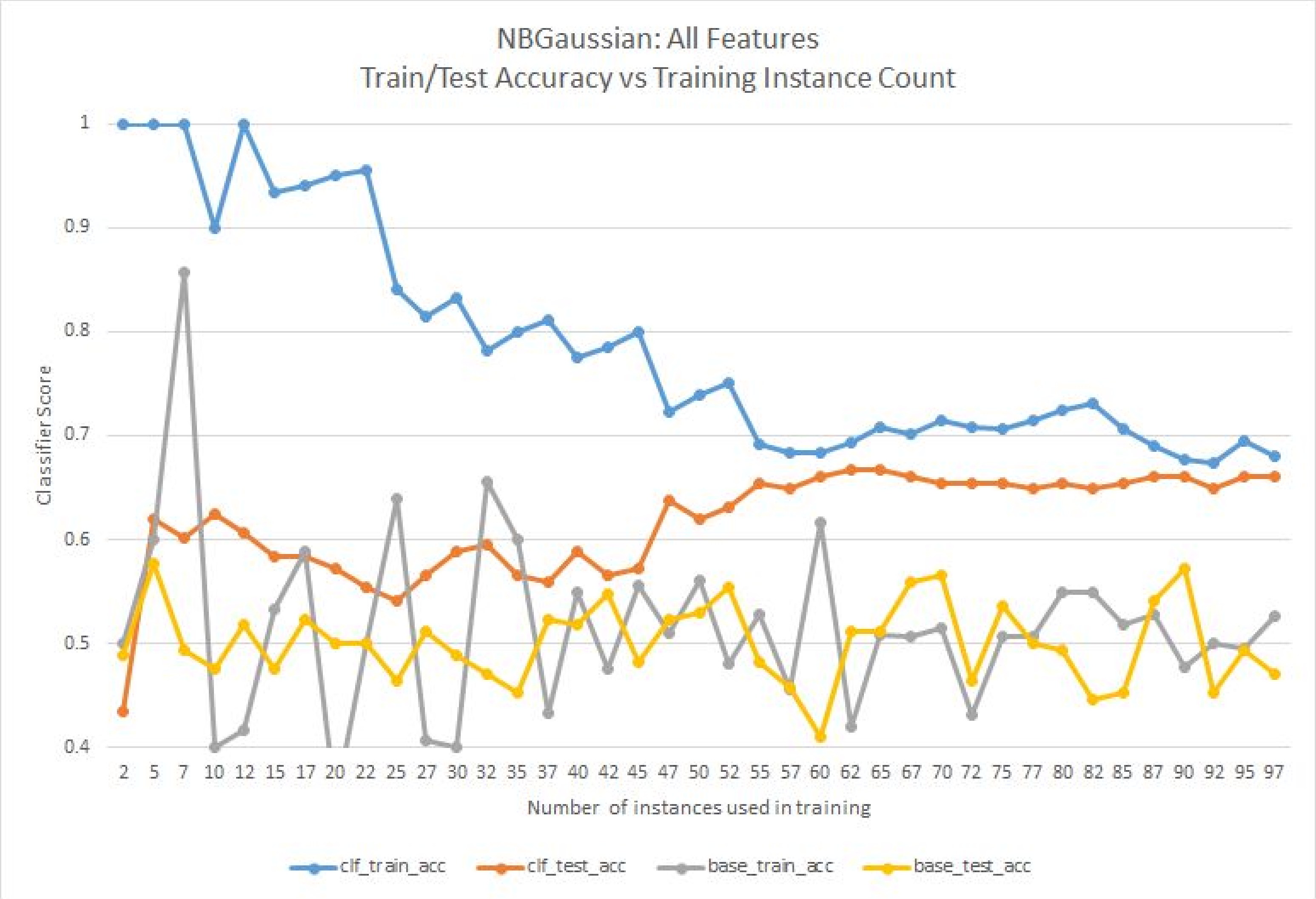

50
45
40
35
30
25
20
15
10
5
0
NaiveBayes
Logistic Regression
SMO
IBK
AttributeSelectedCla
Bagging
AdaBoost
LogitBoost
Stack
Tree
Accuracy (%)

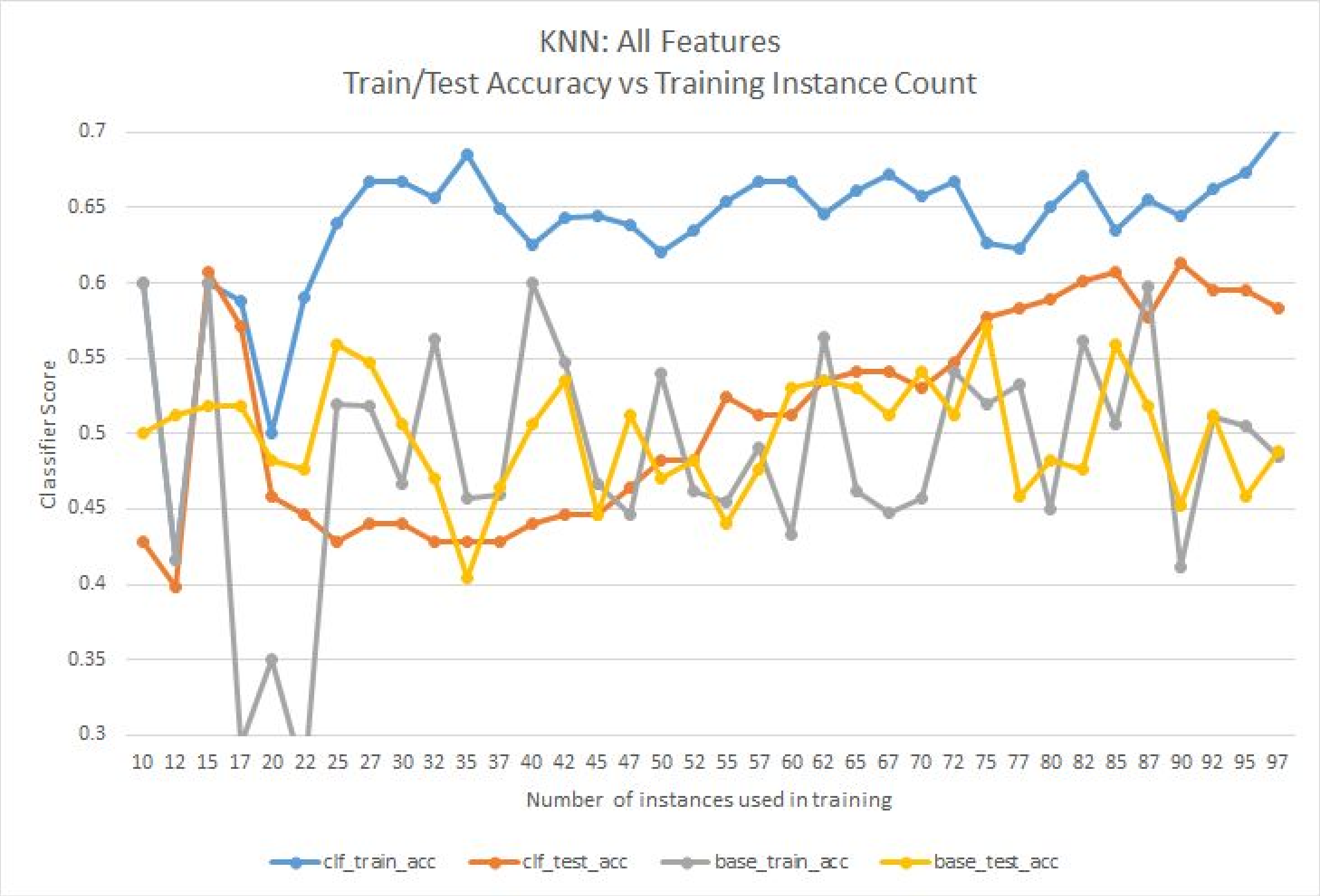
linear regression

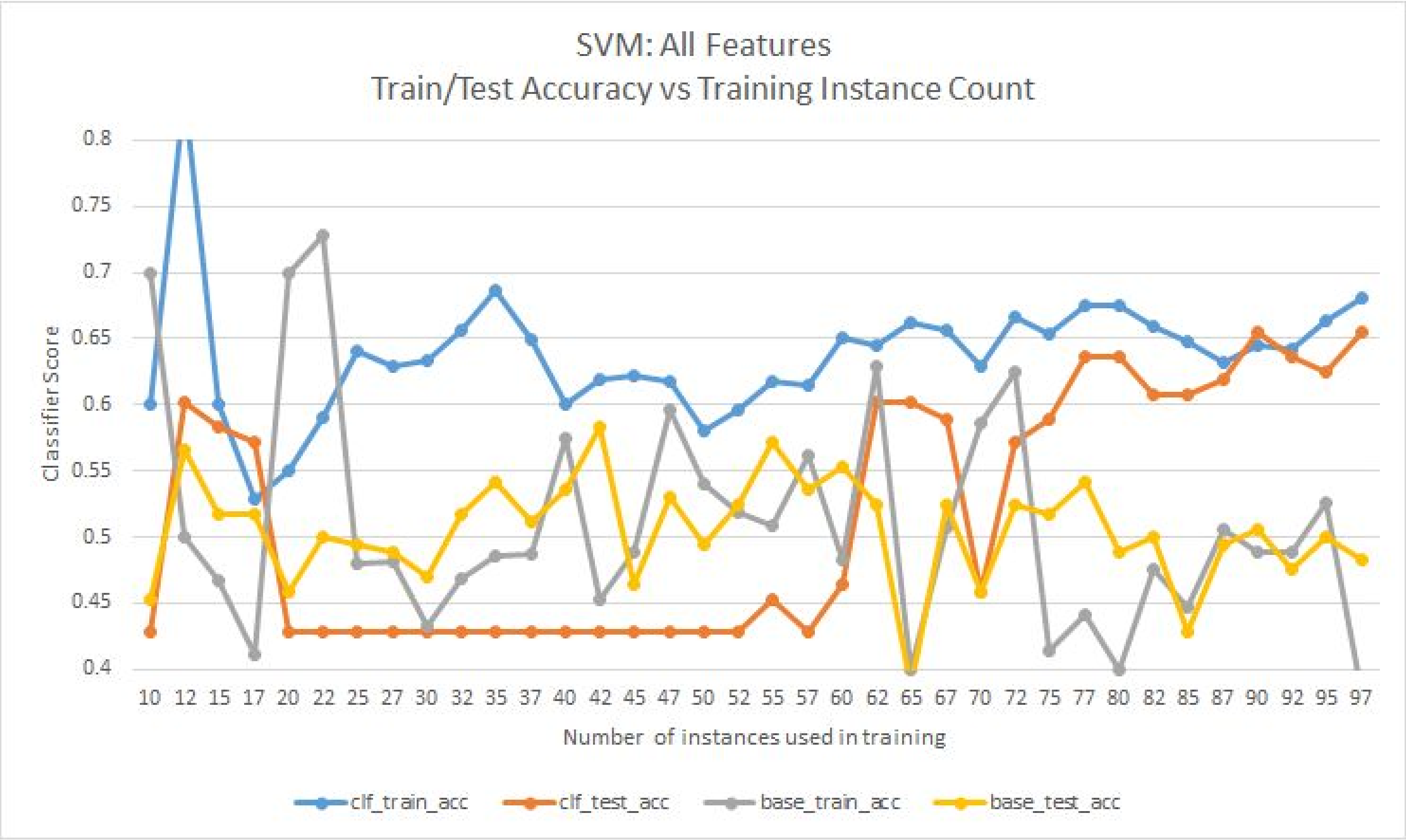
$$\begin{aligned} Y= {} & 0.2045 * \text{gender} \\ & - 0.0003 * \text{Firgen} \\ & - 0.0004 * \text{SATCRDG} \\ & + 0.0004 * \text{SATMATH} \\ & + 0.0011 * \text{SATWRTG} \\ & + 0.0001 * \text{SATTotal} \\ & + 0.4487 * \text{HSGPA} \\ & + 0.0175 * \text{ACTEngWrit} \\ & + 0.0198 * \text{APIScore} \\ & + 0.0591 * \text{FirstLang} \\ & - 0.302 \end{aligned}$$

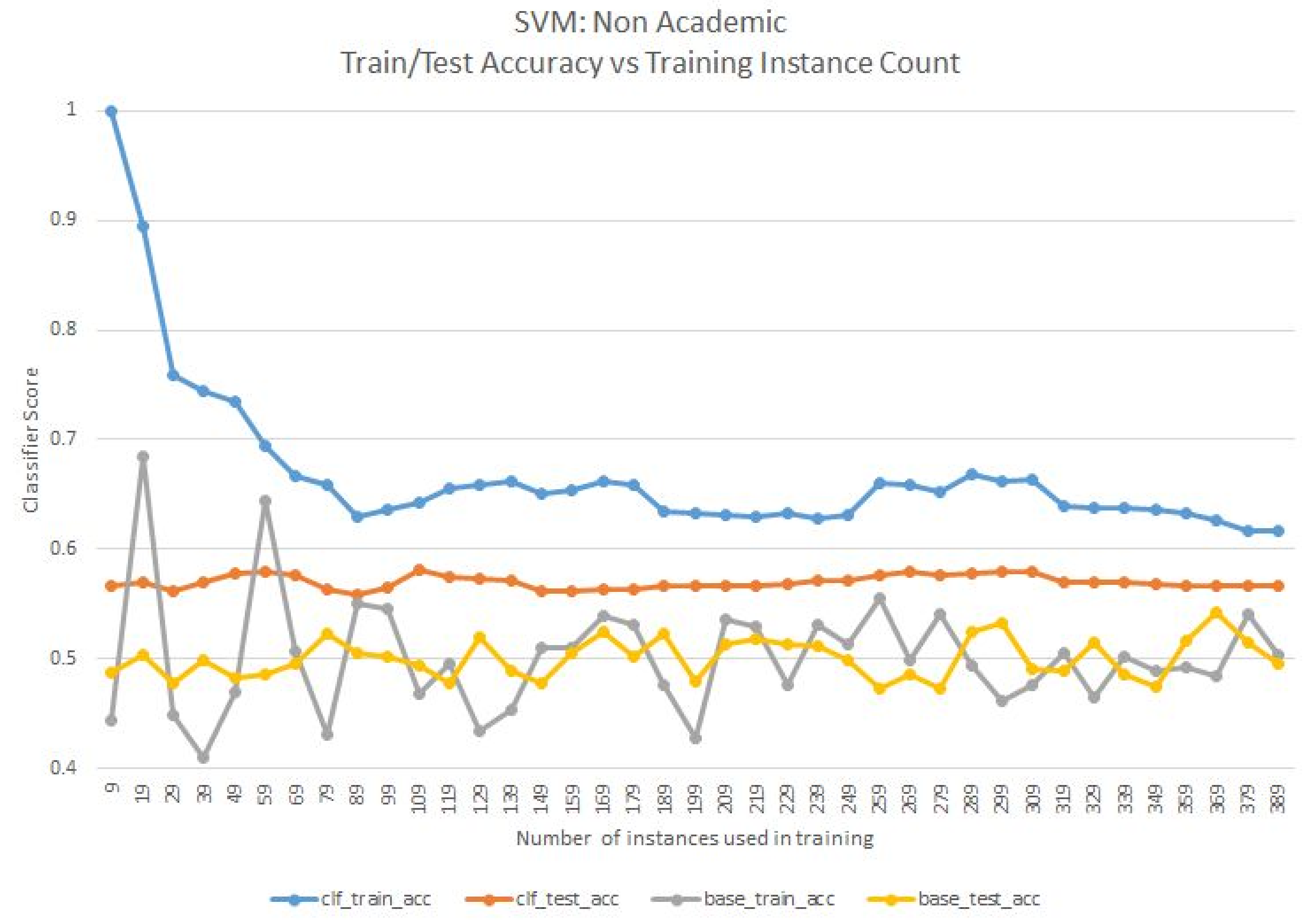
Properties	Correlation coefficient
HSGPA	0.2074
gender	0.088
FirstLang	0.1355
APIScore	0.1687
ACTEngWrit	0.1612
SATWRTG	0.2447
SATCRDG	0.1915
SATMATH	0.1764
Firgen	0.2447
SATTotal	0.1915
famincome	0.1764
ACTRead	0.1590

HSGPA	0.2074
HSGPA, gender	0.2199
HSGPA, gender, FirstLang	0.2482
HSGPA, gender, FirstLang, APIScore	0.2914
HSGPA, gender, FirstLang, APIScore, ACTEngWrit	0.3009
HSGPA, gender, FirstLang, APIScore, ACTEngWrit, SATWRTG	0.3311
HSGPA, gender, FirstLang, APIScore, ACTEngWrit, SATWRTG, SATMATH	0.3346
HSGPA, gender, FirstLang, APIScore, ACTEngWrit, SATWRTG, SATMATH, SATCRDG	0.3346
HSGPA, gender, FirstLang, APIScore, ACTEngWrit, SATWRTG, SATMATH, SATCRDG, Firgen	0.3356
HSGPA, gender, FirstLang, APIScore, ACTEngWrit, SATWRTG, SATMATH, SATCRDG, Firgen, SATTotal	0.3352

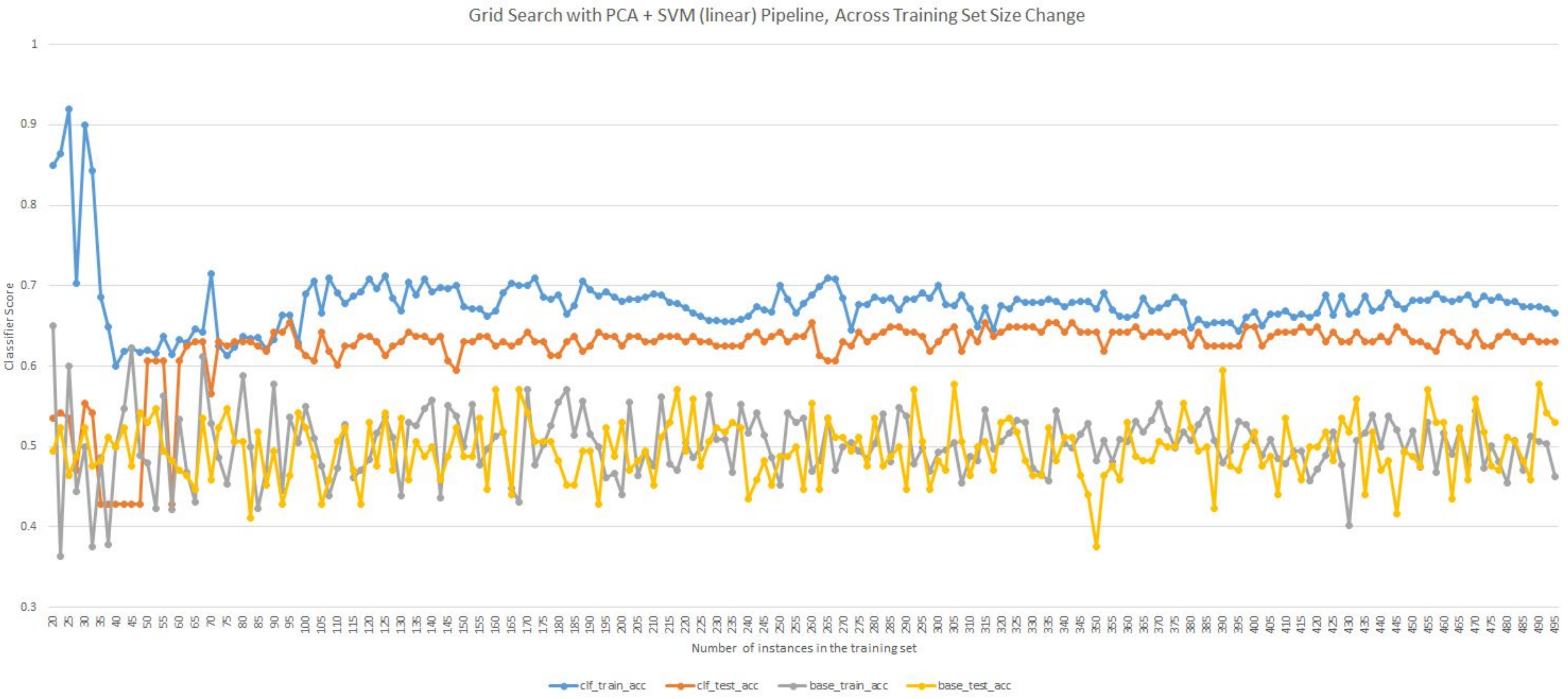

NBGaussian: All Features
Train/Test Accuracy vs Training Instance Count
Classifier Score
1
0.9
0.8
0.7
0.6
0.5
0.4
2 5 7 10 12 15 17 20 22 25 27 30 32 35 37 40 42 45 47 50 52 55 57 60 62 65 67 70 72 75 77 80 82 85 87 90 92 95 97
Number of instances used in training
clf_train_acc
clf_test_acc
base_train_acc
base_test_acc

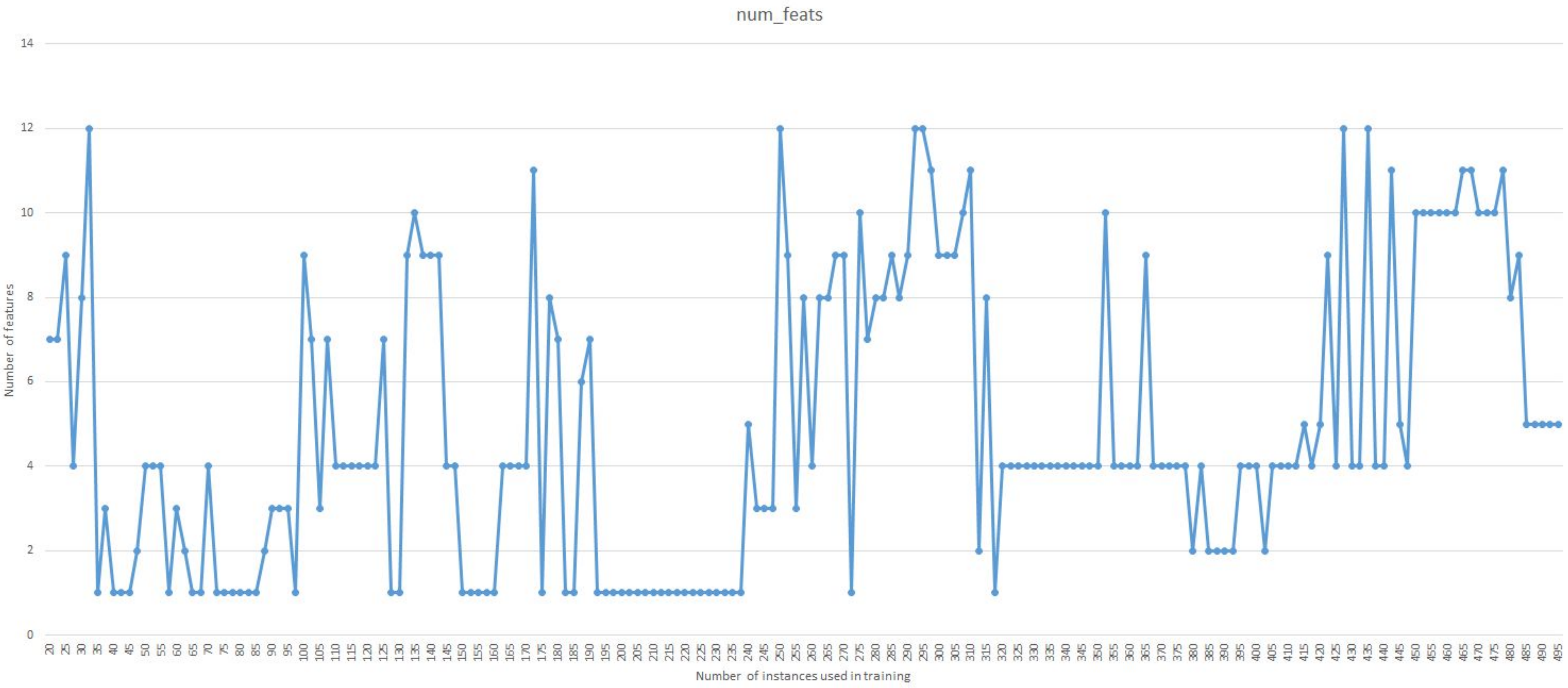

KNN: All Features
Train/Test Accuracy vs Training Instance Count
Classifier Score
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
10
12
15
17
20
22
25
27
30
32
35
37
40
42
45
47
50
52
55
57
60
62
65
67
70
72
75
77
80
82
85
87
90
92
95
97
Number of instances used in training
clf_train_acc
clf_test_acc
base_train_acc
base_test_acc

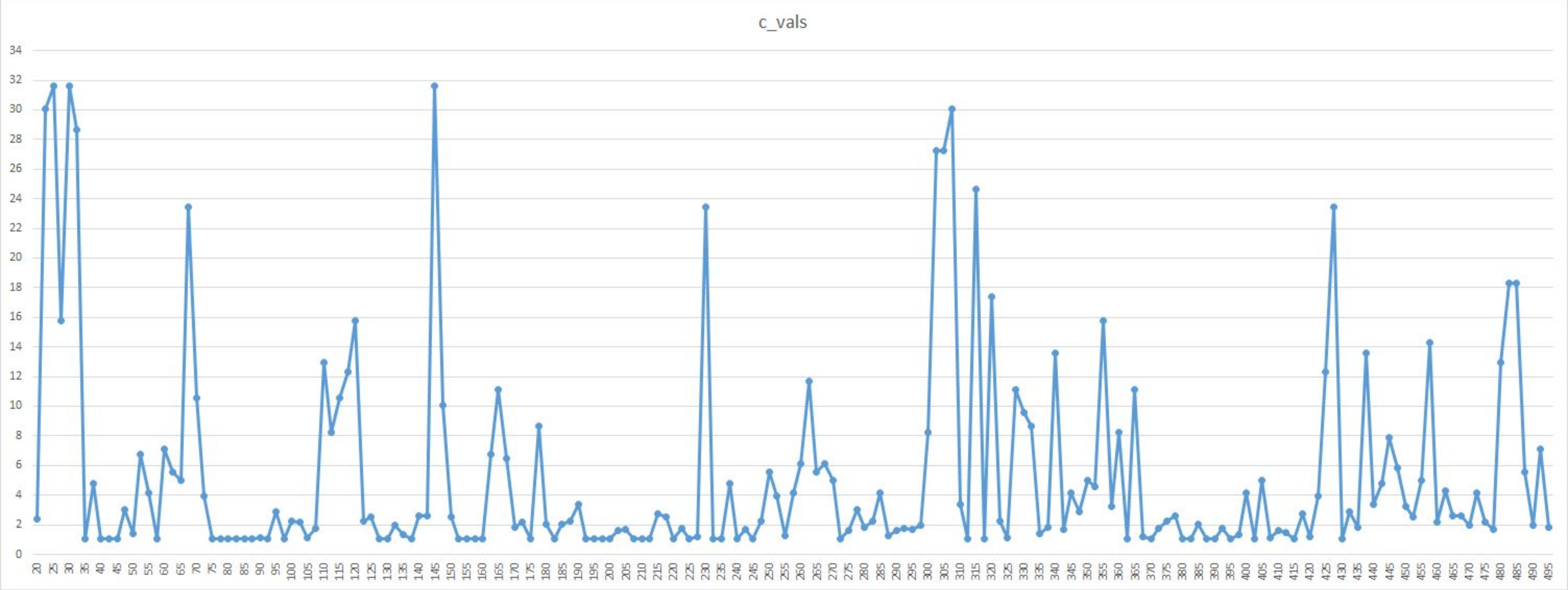

SVM: All Features
Train/Test Accuracy vs Training Instance Count
Classifier Score
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
10
12
15
17
20
22
25
27
30
32
35
37
40
42
45
47
50
52
55
57
60
62
65
67
70
72
75
77
80
82
85
87
90
92
95
97
Number of instances used in training
clf_train_acc
clf_test_acc
base_train_acc
base_test_acc


SVM: Non Academic
Train/Test Accuracy vs Training Instance Count
Classifier Score
1
0.9
0.8
0.7
0.6
0.5
0.4
9
19
29
39
49
59
69
79
89
99
109
119
129
139
149
159
169
179
189
199
209
219
229
239
249
259
269
279
289
299
309
319
329
339
349
359
369
379
389
Number of instances used in training
clf_train_acc
clf_test_acc
base_train_acc
base_test_acc

Break the bias..? Nope.




num_feats
Number of features
Number of instances used in training
0
2
4
6
8
10
12
14
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495



c_vals

34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495