گزارش پروژه یادگیری ماشین

دانیال ملک محمد ۹۴۱۰۰۰۹۲

سوال اول)

در ادامه تنظیم هایپر پارامترها برای روش های مختلف آمد است. در زیر هر تغییر هایپر پارامتری، accuracy آن آمده است. توجه کنید که این accuracy از طریق cross validation(3fold) و میانگین سه دقت آن بدست آمده است.

Logistic_Regression:

fit_intercept=True
0.9688440607286775

fit_intercept=False
0.9688443595679855

intercept_scaling
0.9688440607286775

intercept_scaling=0.1
0.9688440607286775

solver:
تنها solver یی که توانایی حل multinomial و l1,l2 loss را داشت saga بود هرچند بقیه نیز آزموده شدند:

#sag
0.9679428599512322
همگرایی دیر،کند

saga
0.9678134132428798
همگرایی دیر،کند

lbfgs
0.9657551056271295
سریع، اندکی دقت پاینتر

lgfgs one vs rest
0.9683295580159488

newton-cg
0.9688440607286775

در کل newton-cg سریعترین و بادقت ترین بود اما چون l1 norm ندارد، در ادامه saga استفاده می شود

:Random forest

criterion:

entropy
0.960477497609082

gini
0.9571294970780846

max_depth=10
0.9558427946303913

max_depth=7
0.9370476185420623

در ادامه همه با max_depth =7 انجام شد و با دقت بالایی مقایسه شد:

bootstrap = false
0.9362728812660697

oob=True
0.9343411819038625

warm start=True
0.9352426804833797

weight = balanced
0.9330535787892997

:SVM

تنظیم C:

C=100000
0.974511057402451

C=10000
0.974511057402451

C=1000
0.9746402063087315

C=500
0.9747687564991594

C=200
0.9752820669663645

C=100
0.9756666762676334

C=50
0.9736065725043254

C=10
0.96601071596039

C=1
0.9145040107339176

C=None
0.9145040107339176

C=0.1
0.8034027717563648

———————- —> C = 100

Kernel type:

Linear
0.9661428480735701

sigmoid
0.9691020597019303

sigmoid coef0=100 , coef0=1
0.18449865608209878

poly degree=3 , coef0=1

0.974380417411339

poly degree=5 , coef0=1
0.974381909533407

poly degree=10 , coef0=1
0.9764418612839701

poly degree=100 , coef0=1
0.9325362355472894

Poly degree=7 , coef0=1
0.9751556045021483

Poly degree=30 , coef0=1
0.9747683097960517

Poly degree=20 , coef0=1
0.9770856628412515

Poly degree=25 , coef0=1
0.9760556119968338

Poly degree=15 , coef0=1
0.9768276638679988

Poly degree=15 , coef0=10
0.973095055072969
ـــــــــــــــــــــــــــــــــ

و در این قسمت نتیجه گرفتم که کرنل rbf مطمئن تر و بهتر است

در ادامه کرنل rbf بود و c=100:

shrinking=False
0.9756666762676334

posibilty=True
0.9756666762676334

One vs rest
0.9756666762676334

One vs one
0.9756666762676334

:Adaboost

max_depth=10,n_estimators=50,learning_rate=1
[0.80871577 0.83127413 0.82894737]

max_depth=10,n_estimators=50,learning_rate=0.9,
[0.85653683 0.84671815 0.87345201]

max_depth=10,n_estimators=20,learning_rate=0.95
[0.91939838 0.94169884 0.9369195]

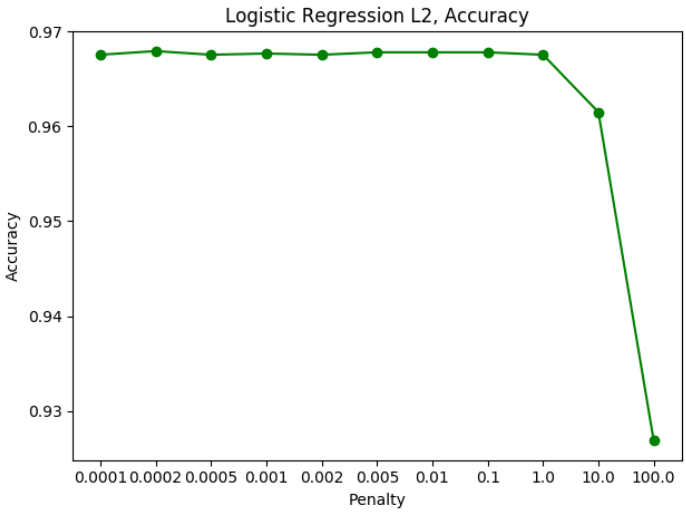
max_depth=7,n_estimators=50,learning_rate=1
0.9197878829678364

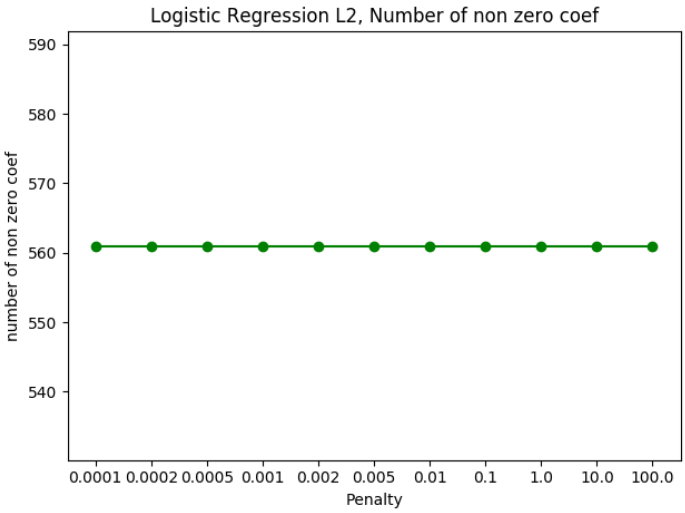
:Fully_connected

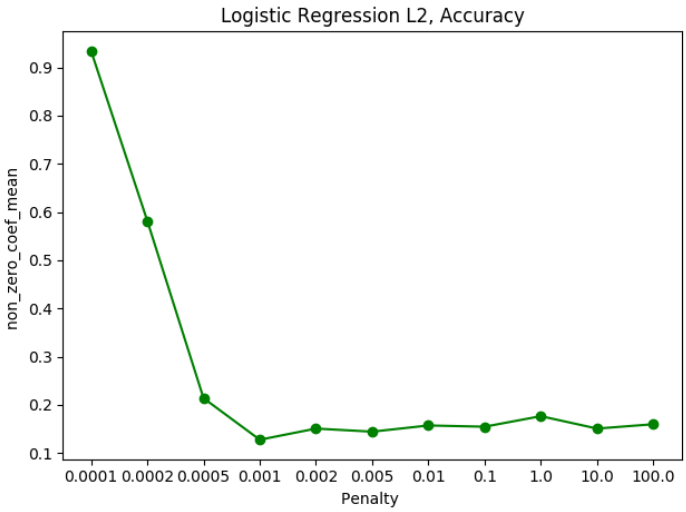
پس از تغییر پارامتر های زیاد، با dropout=0.2 ،batchsize=128،epoch=250و لایه اول ۱۵۰و لایه دوم ۵۰ تایی ، به دقت ۹۷ درصد روی داده ی ولیدیشن رسیدم. از اپتیمایز nadam استفاده کردم.

سوال دوم)

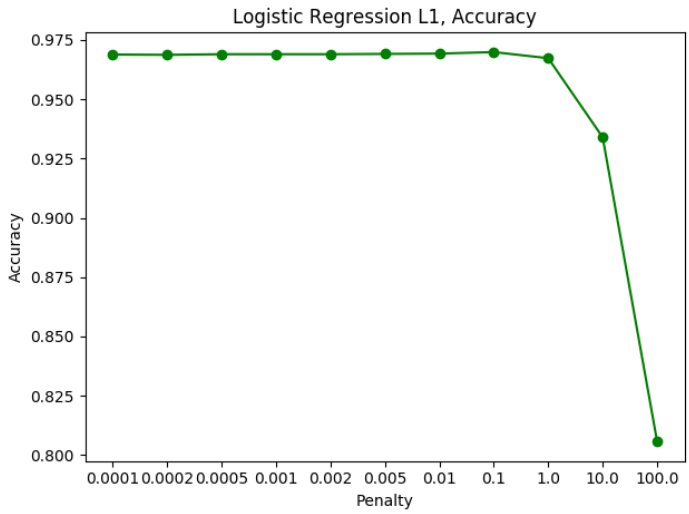
Logistic_Regression , L2 :

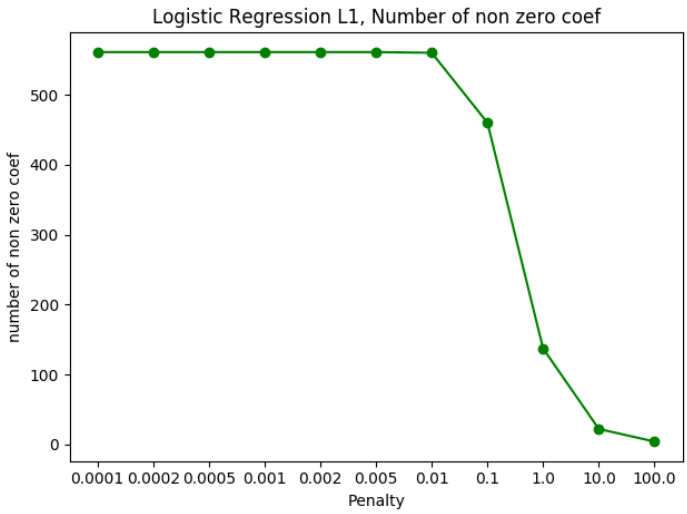




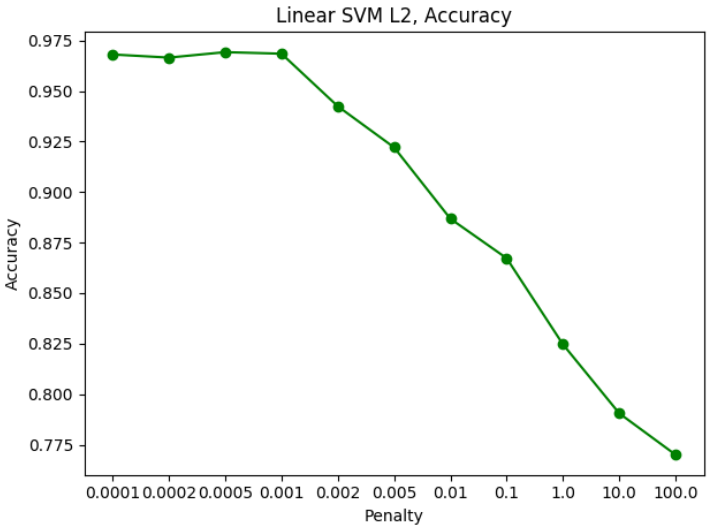


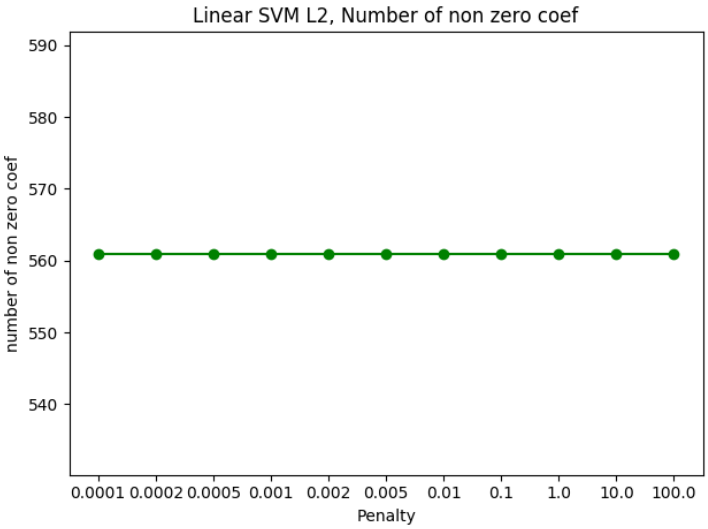
Logistic_Regression , L1 :

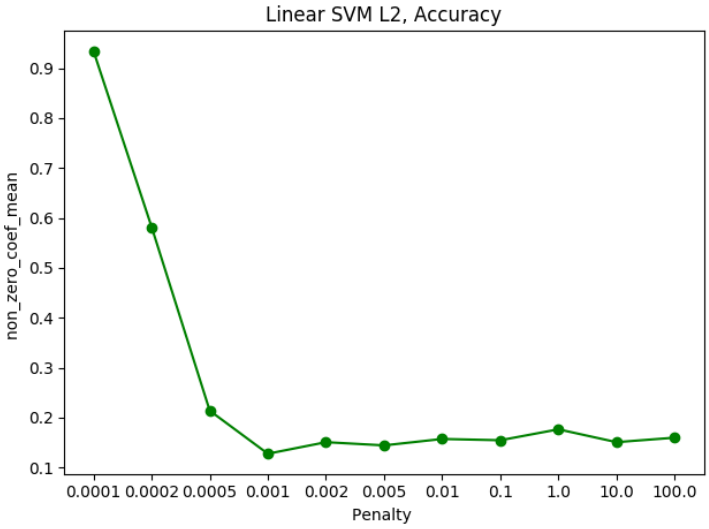




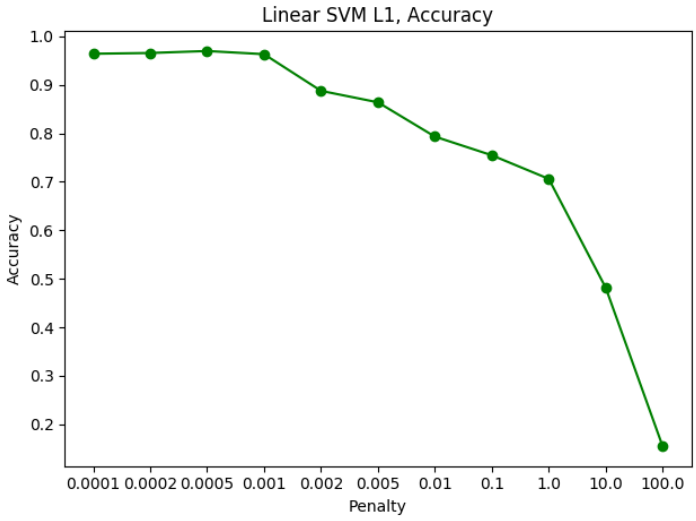
Linear SVM ， L2 :

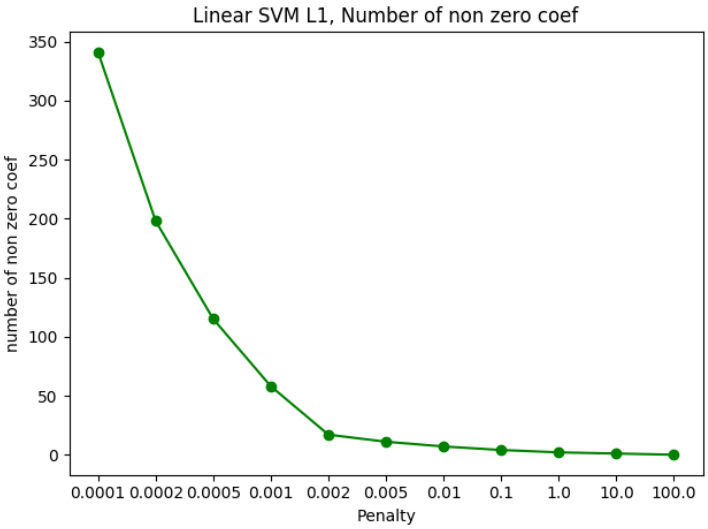




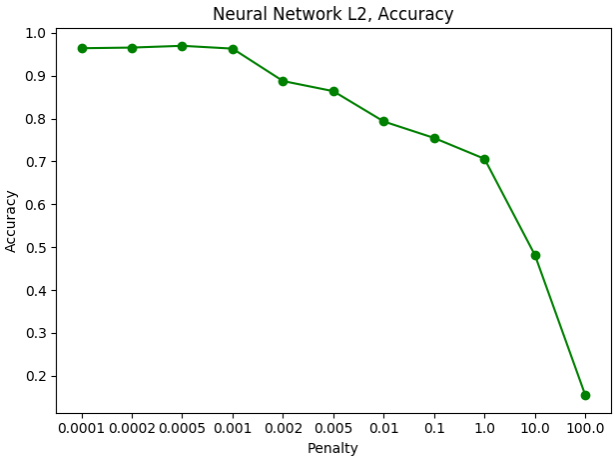


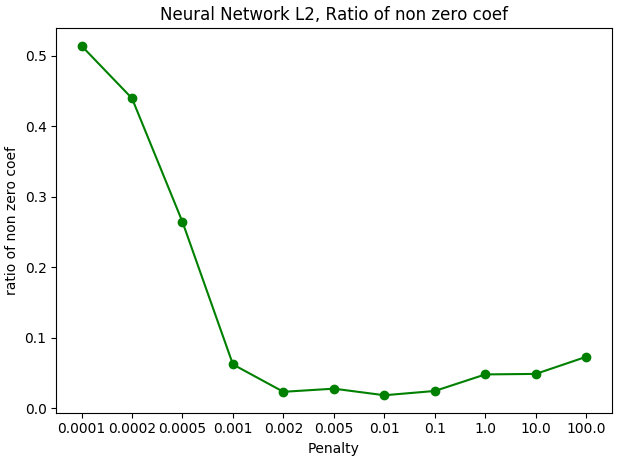
Linear SVM ， L1:

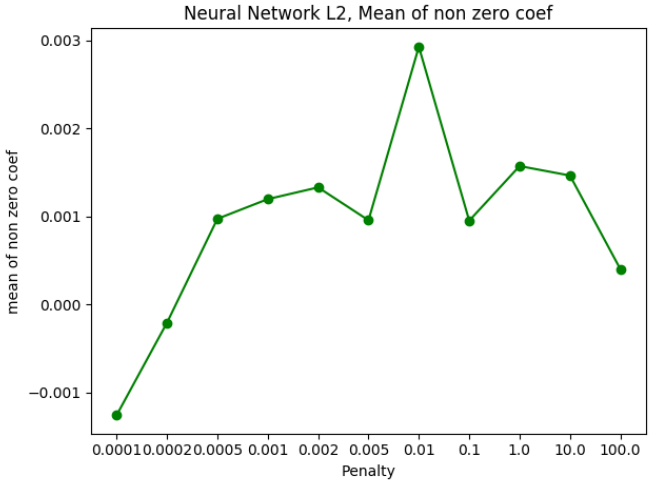




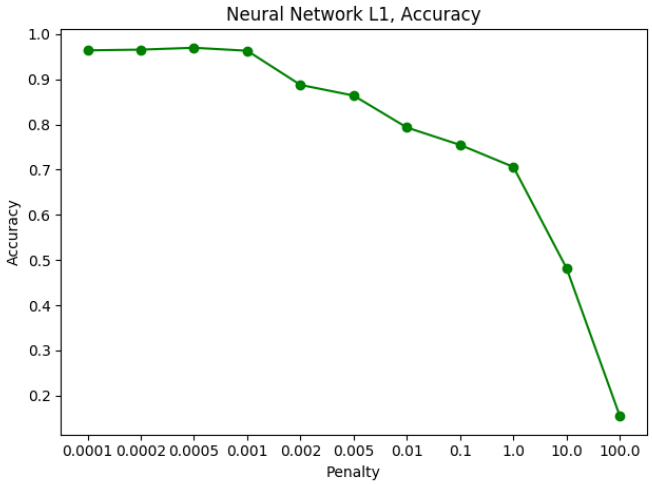
Neural Network , L2:

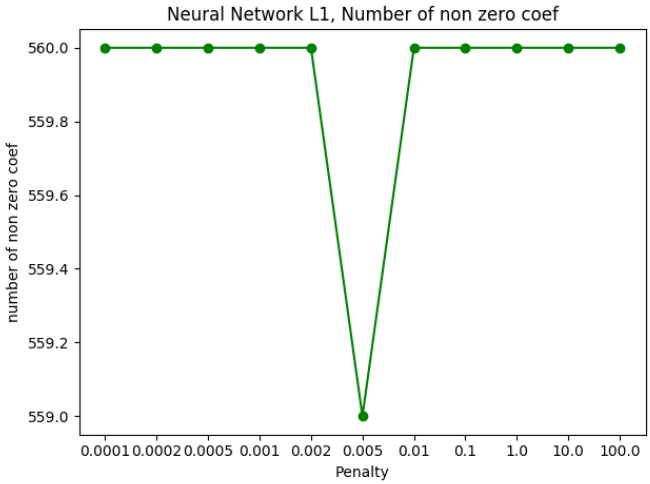


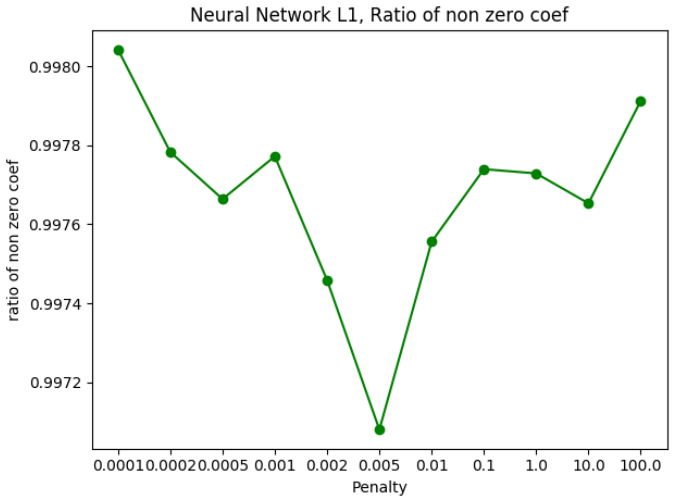




Neural Network, L1

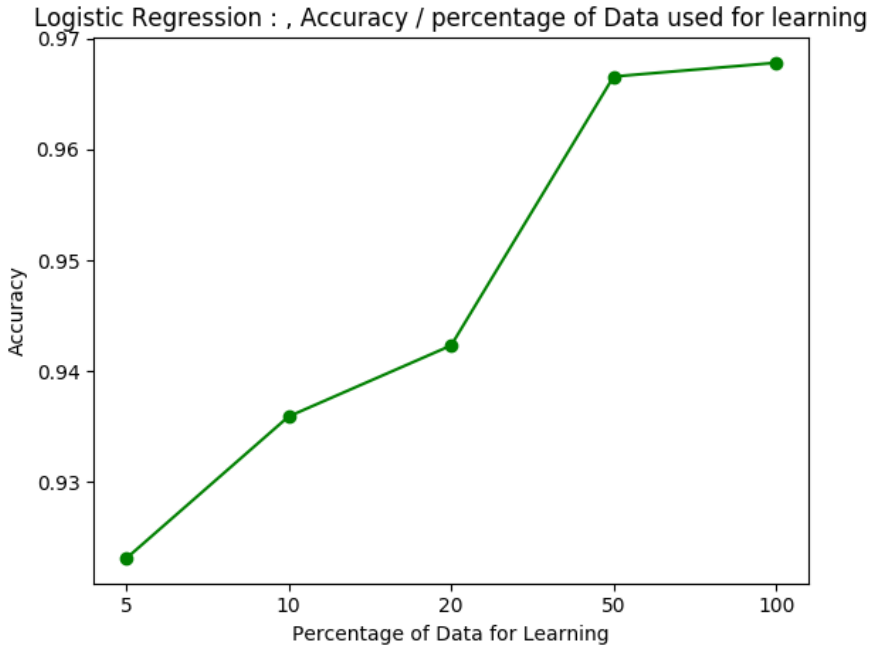


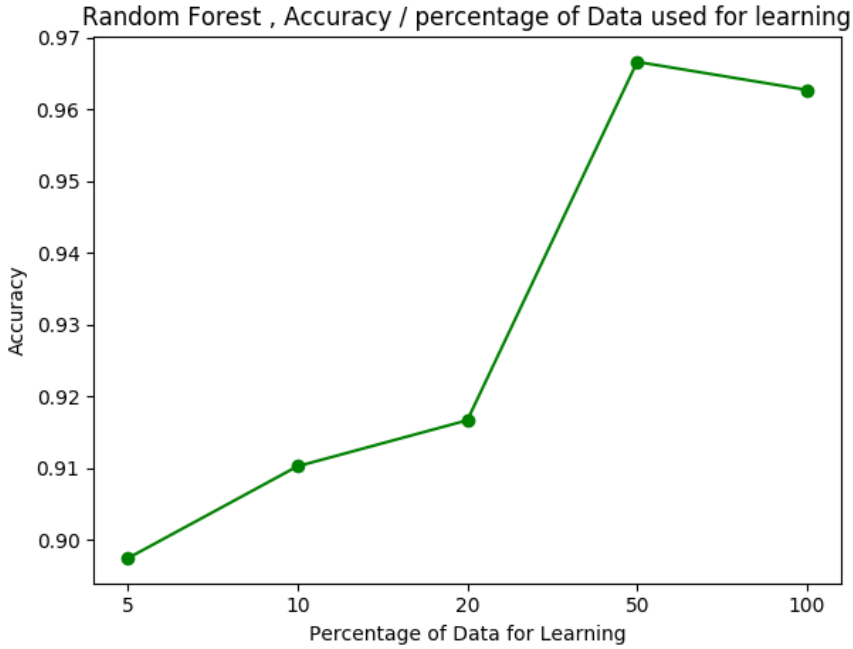


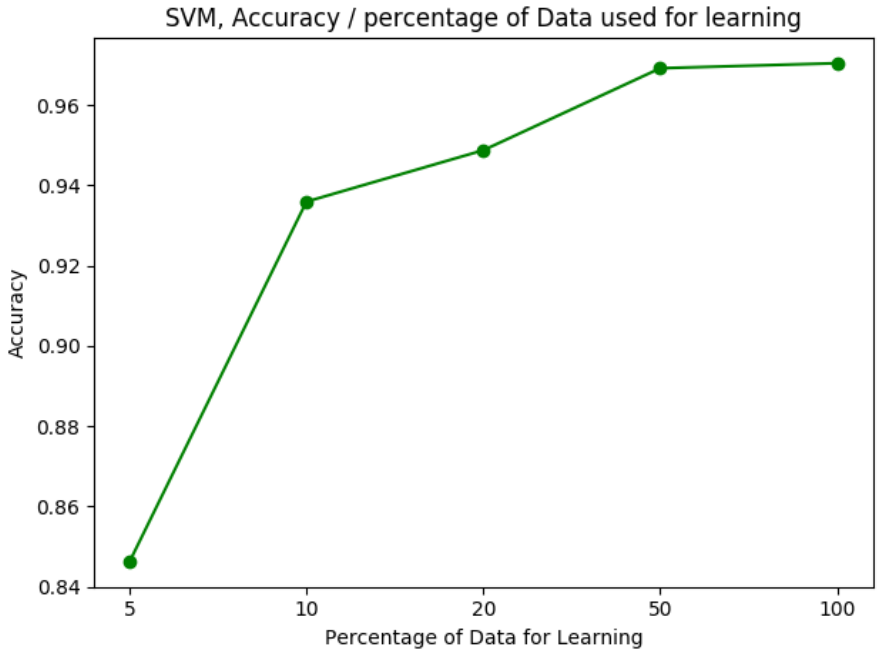


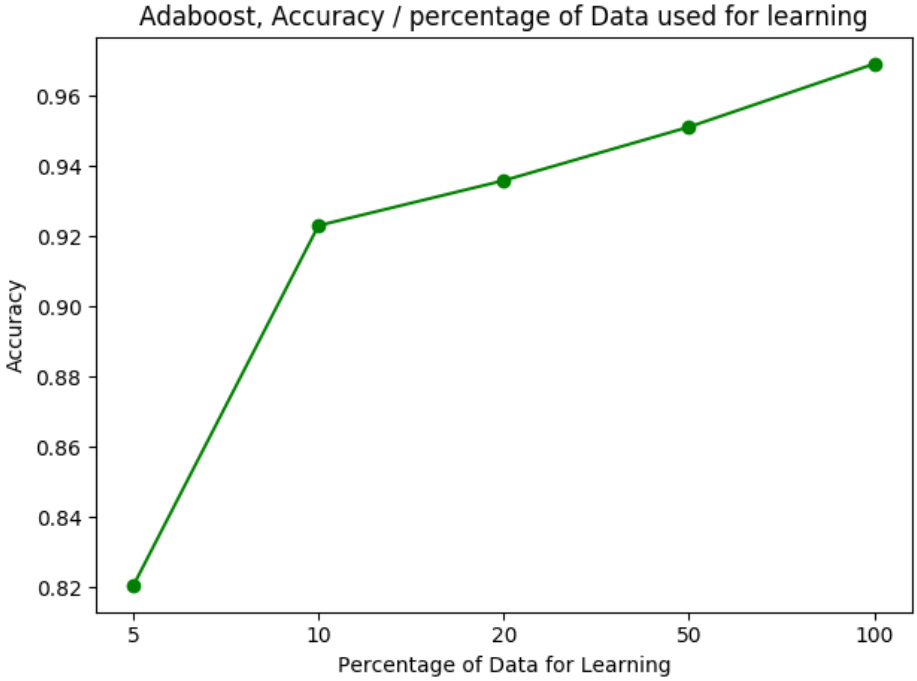
همانطور که دیدیم در همه ی مدل ها در L2 با افزایش ضریب جریمه، میانگین ضرایب کم می شود اما در L1 این امر موجب کاهش تعداد آن ها می شود.همچنین افزایش ضریب جریمه موجب کاهش دقت در مدل ها شد.

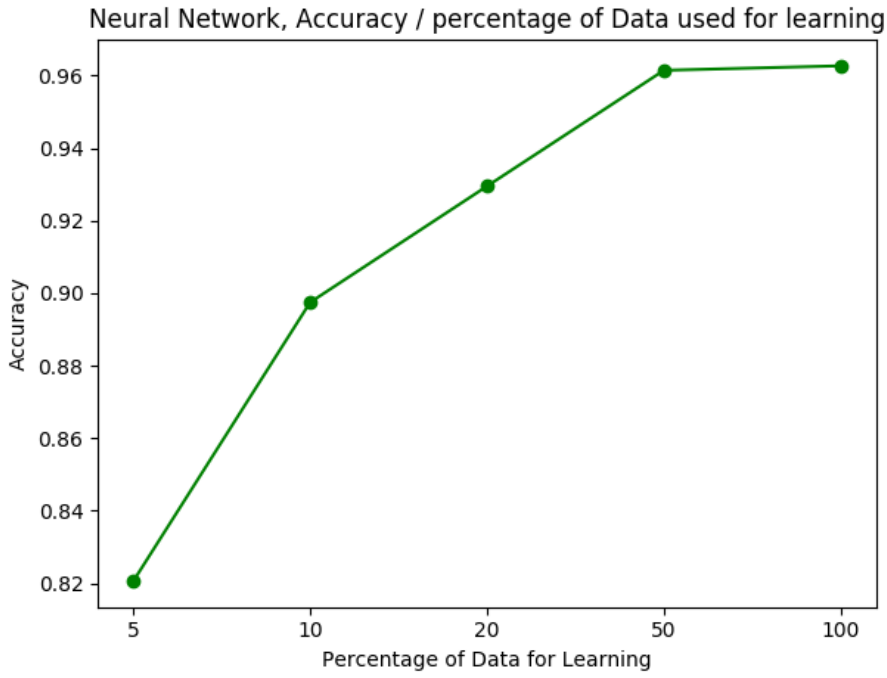
سوال سوم)


Logistic Regression : , Accuracy / percentage of Data used for learning
Accuracy
0.97
0.96
0.95
0.94
0.93
5
10
20
50
100
Percentage of Data for Learning


Random Forest , Accuracy / percentage of Data used for learning
Accuracy
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
5
10
20
50
100
Percentage of Data for Learning

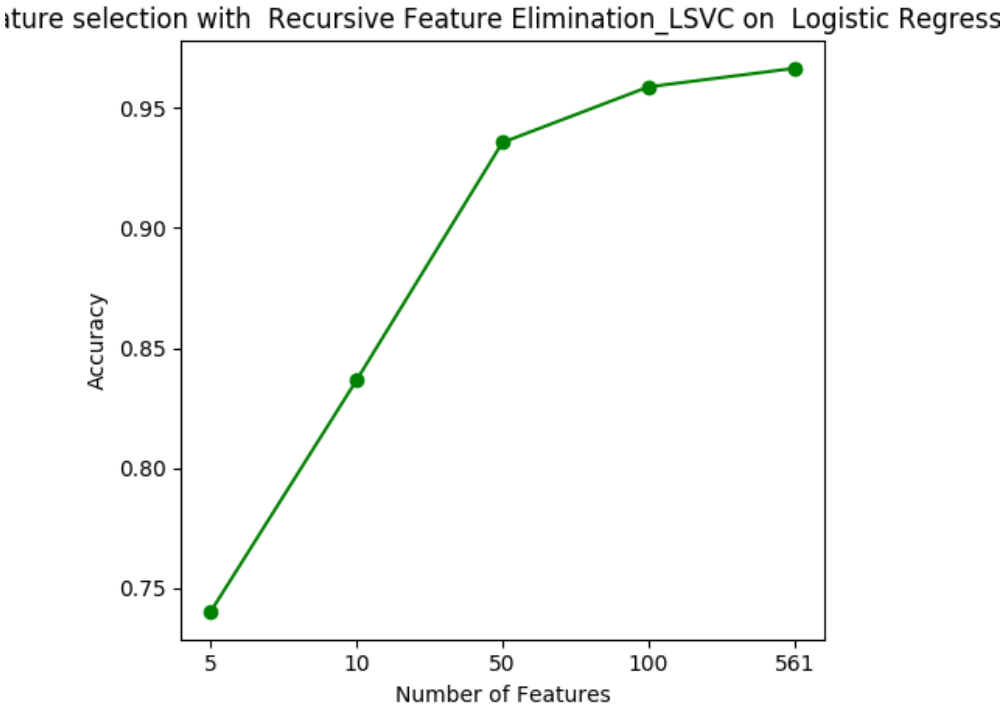

SVM, Accuracy / percentage of Data used for learning
Accuracy
0.96
0.94
0.92
0.90
0.88
0.86
0.84
5
10
20
50
100
Percentage of Data for Learning

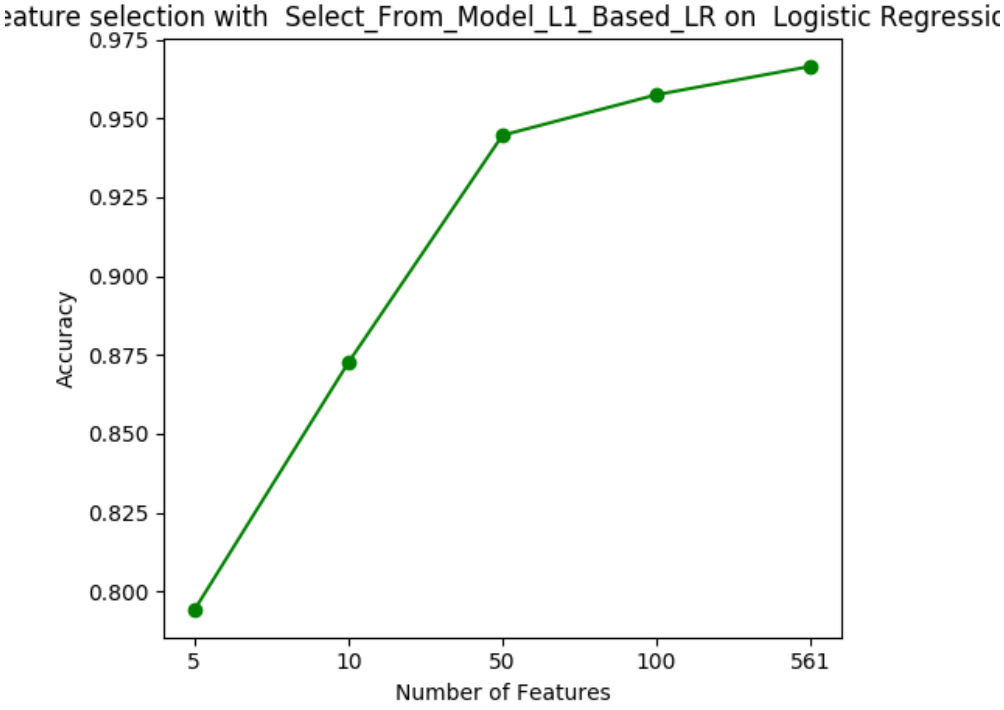

Adaboost, Accuracy / percentage of Data used for learning
Accuracy
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
5
10
20
50
100
Percentage of Data for Learning

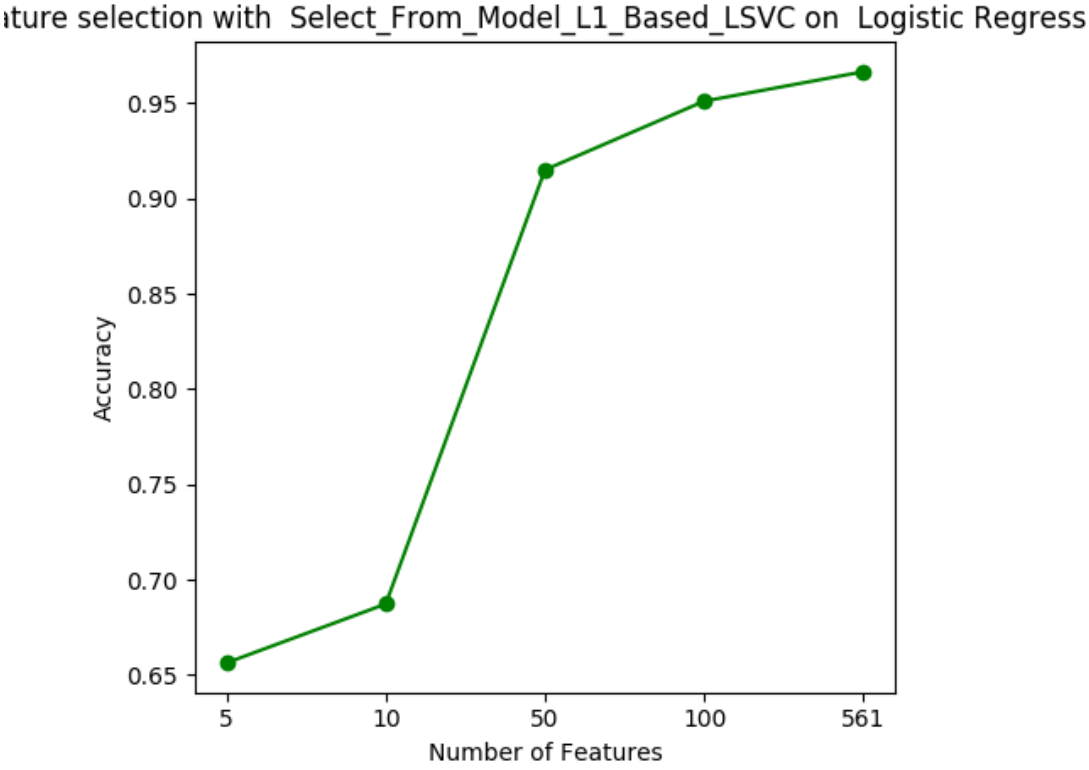

Neural Network, Accuracy / percentage of Data used for learning
Accuracy
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
5
10
20
50
100
Percentage of Data for Learning

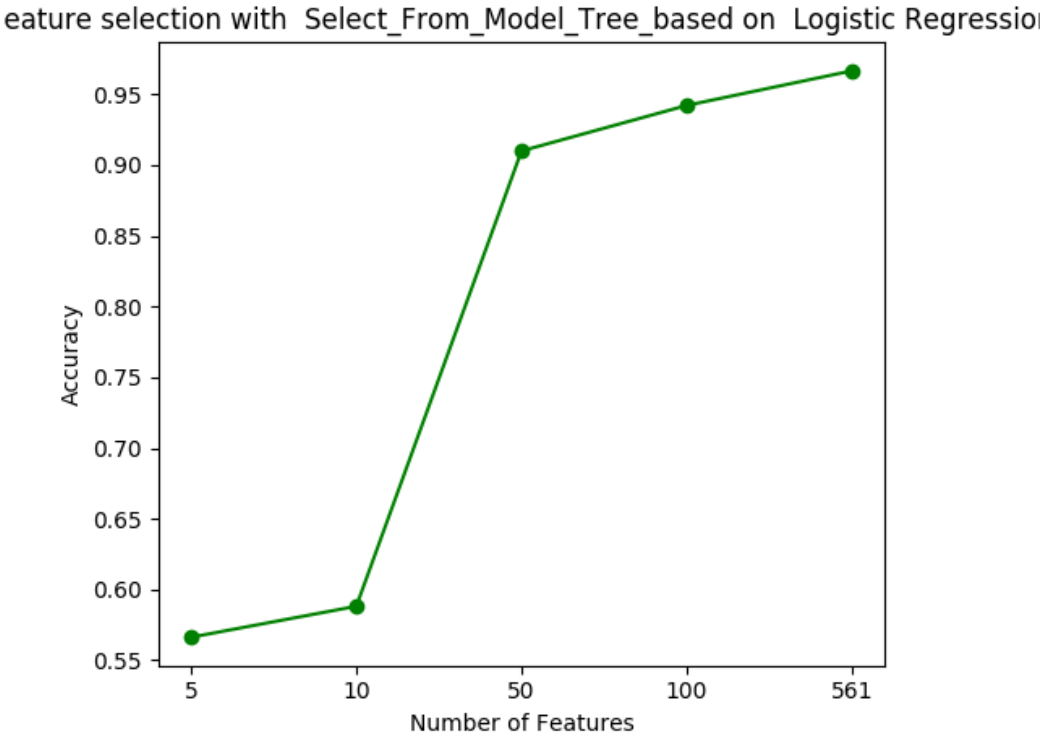
سوال چهارم)

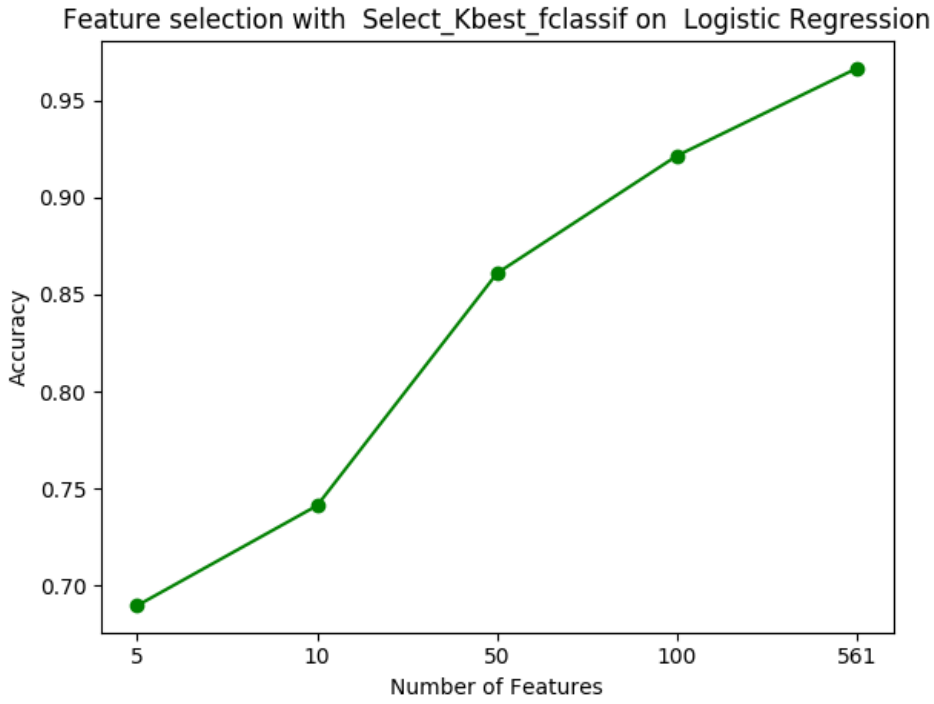
Logistic Regression :



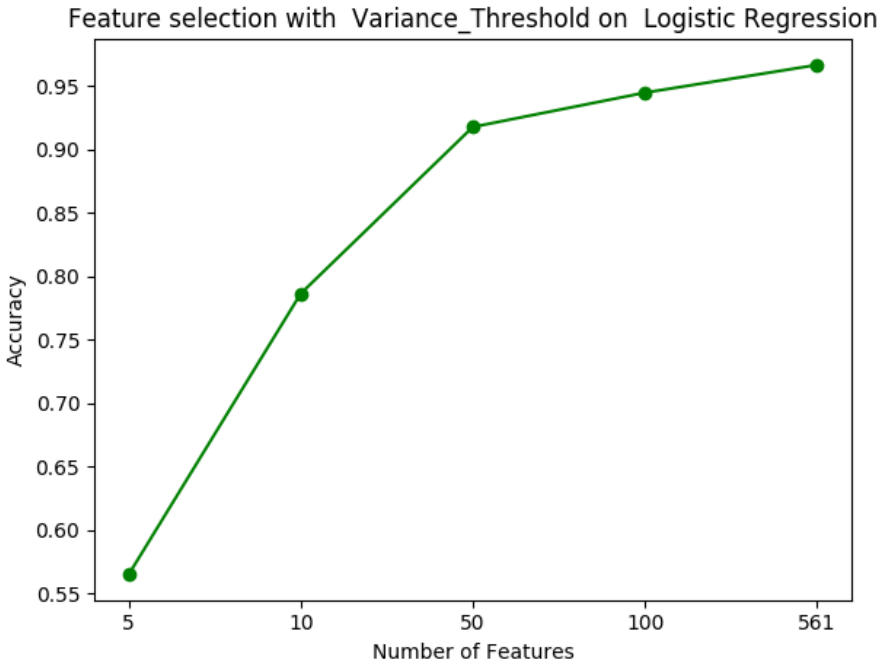



ıture selection with Select_From_Model_L1_Based_LSVC on Logistic Regress
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
5
10
50
100
561
Number of Features


eature selection with Select_From_Model_Tree_based on Logistic Regressio
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_fclassif on Logistic Regression
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
5
10
50
100
561
Number of Features


eature selection with Select_Kbest_mutual_info_classif on Logistic Regressio
0.9
0.8
0.7
0.6
Accuracy
5
10
50
100
561
Number of Features


Feature selection with Variance_Threshold on Logistic Regression
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
5
10
50
100
561
Number of Features

ForRecursive Feature Elimination_LSVC Features :
#Accuracies = [0.7400257400257401, 0.8365508365508365, 0.9356499356499357,
0.9588159588159588, 0.9665379665379665]

##################################
For Select_From_Model_L1_Based_LR Features :
#Accuracies = [0.7940797940797941, 0.8725868725868726, 0.9446589446589446,
0.9575289575289575, 0.9665379665379665]

##################################
For Select_From_Model_L1_Based_LSVC Features :
#Accuracies = [0.6563706563706564, 0.6872586872586872, 0.915057915057915,
0.9510939510939511, 0.9665379665379665]

##################################
For Select_From_Model_Tree_based Features :
#Accuracies = [0.5662805662805663, 0.5881595881595881, 0.9099099099099099,
0.9420849420849421, 0.9665379665379665]

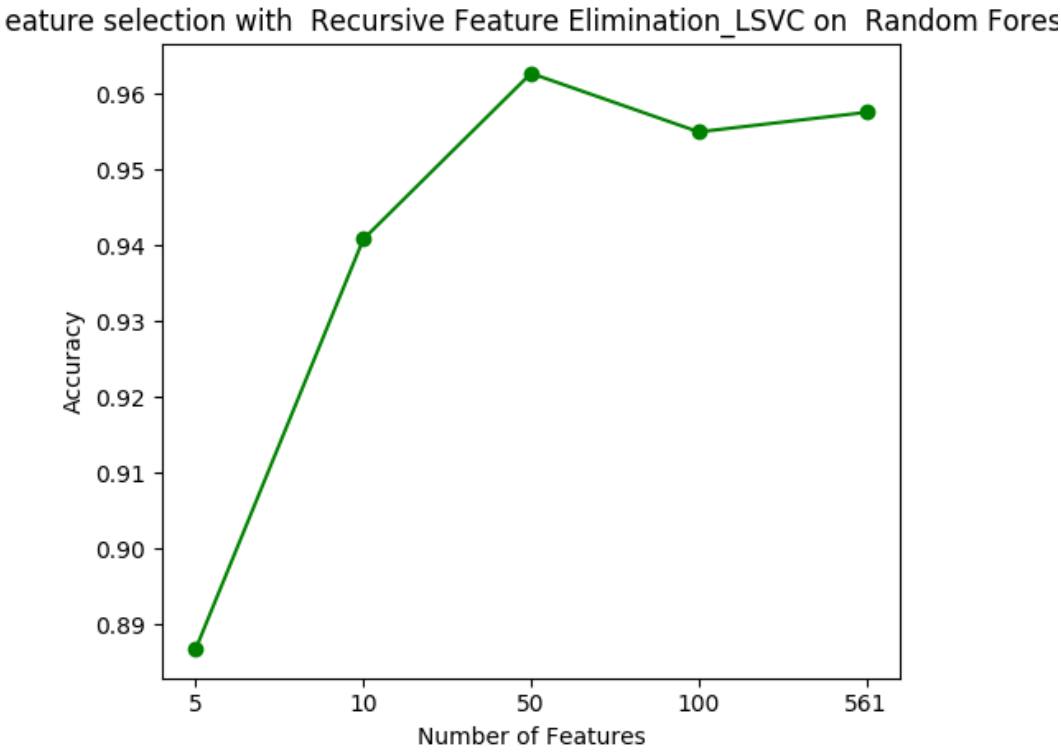
##################################
For Select_Kbest_fclassif Features :
#Accuracies = [0.6898326898326899, 0.7413127413127413, 0.861003861003861,
0.9214929214929215, 0.9665379665379665]

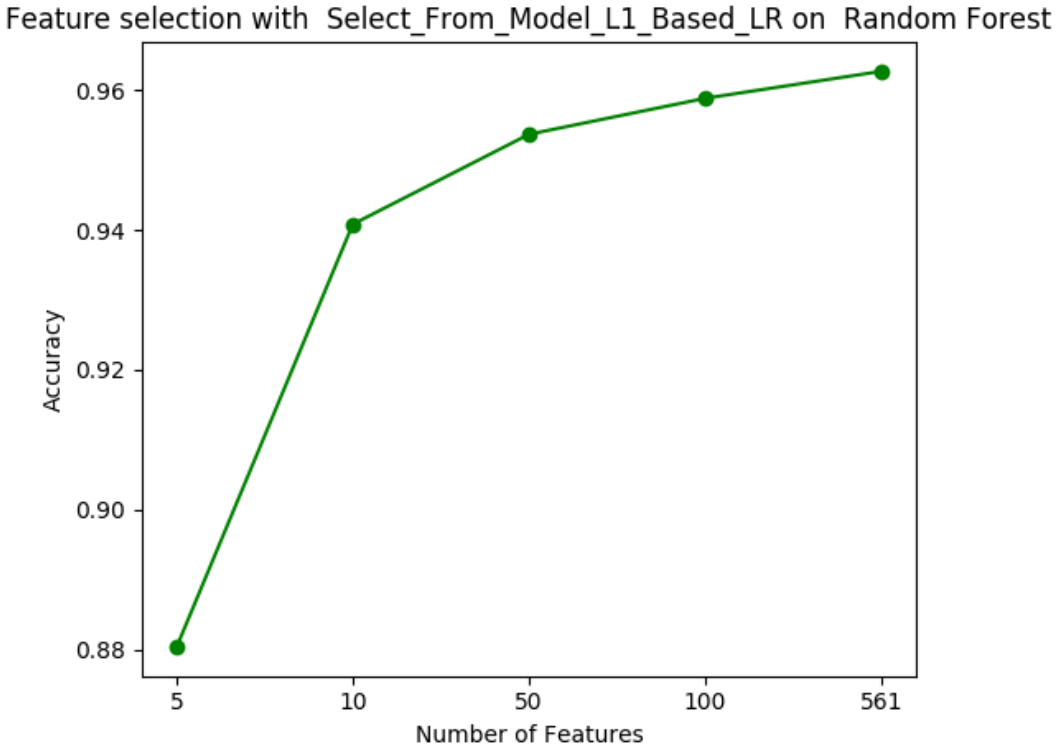
##################################
For Select_Kbest_mutual_info_classif Features :
#Accuracies = [0.5521235521235521, 0.574002574002574, 0.8532818532818532,
0.9073359073359073, 0.9665379665379665]

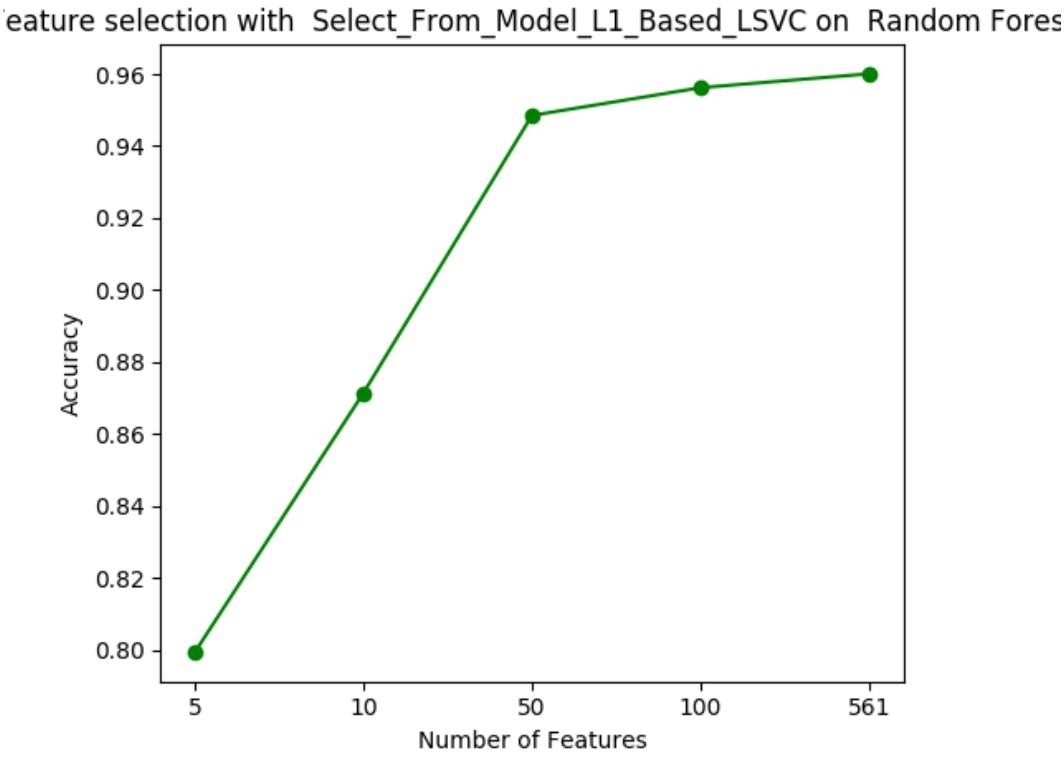
##################################
For Variance_Threshold Features :
#Accuracies = [0.564993564993565, 0.7863577863577863, 0.9176319176319176,
0.9446589446589446, 0.9665379665379665]

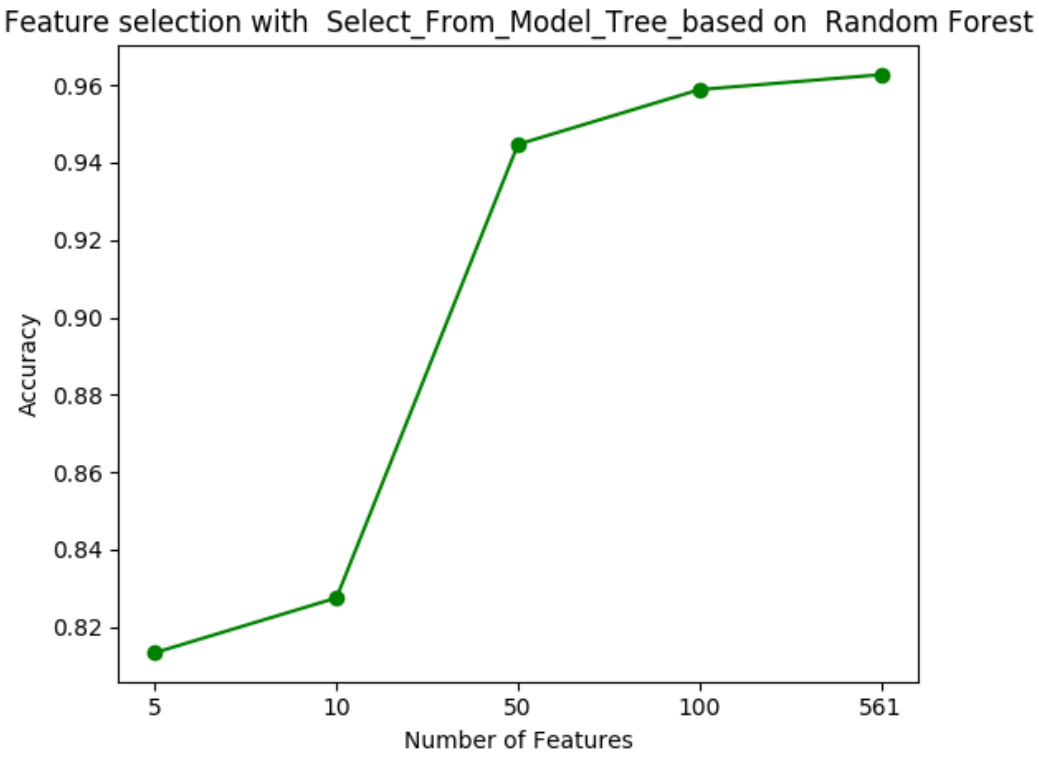
##################################

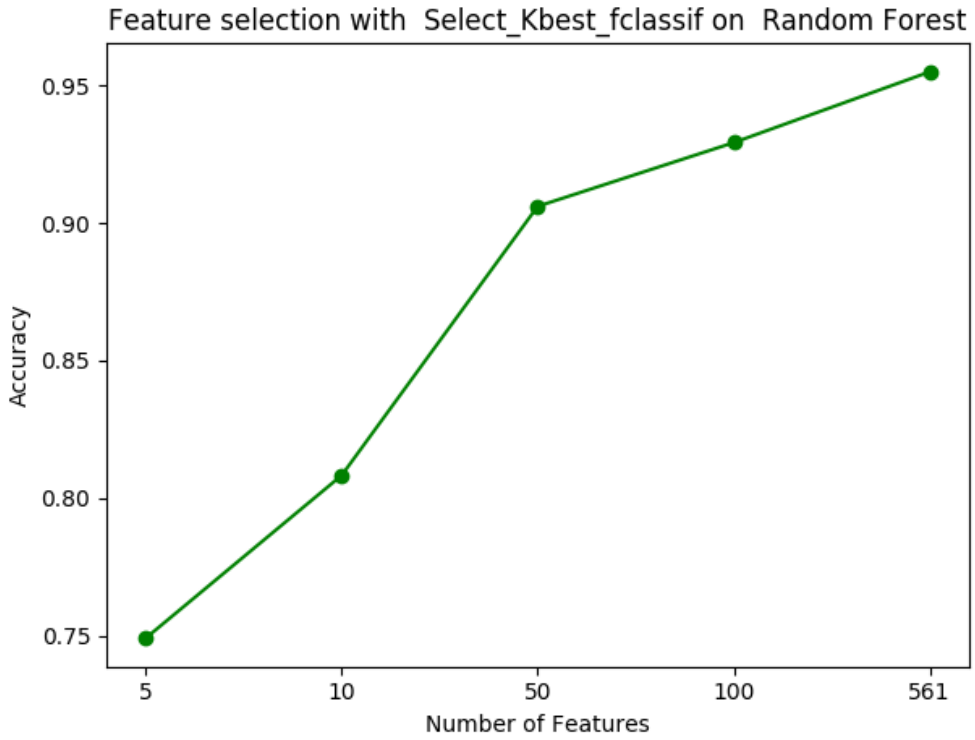
Random Forest:

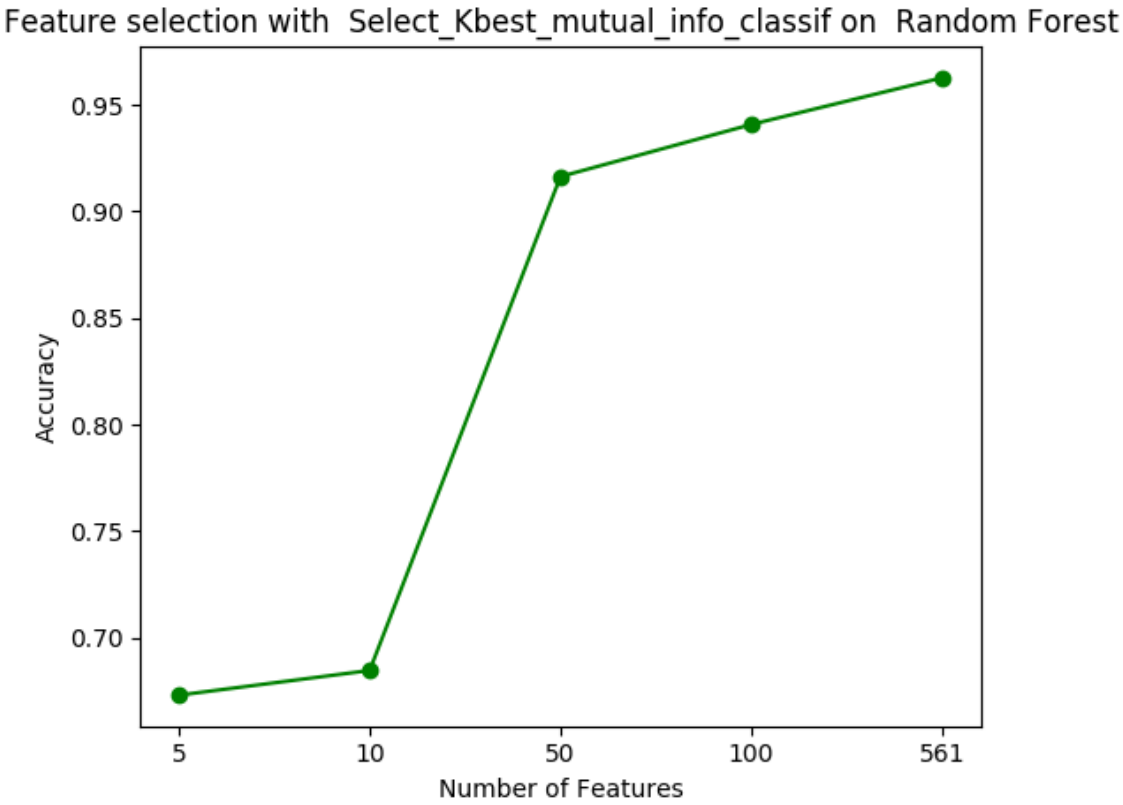


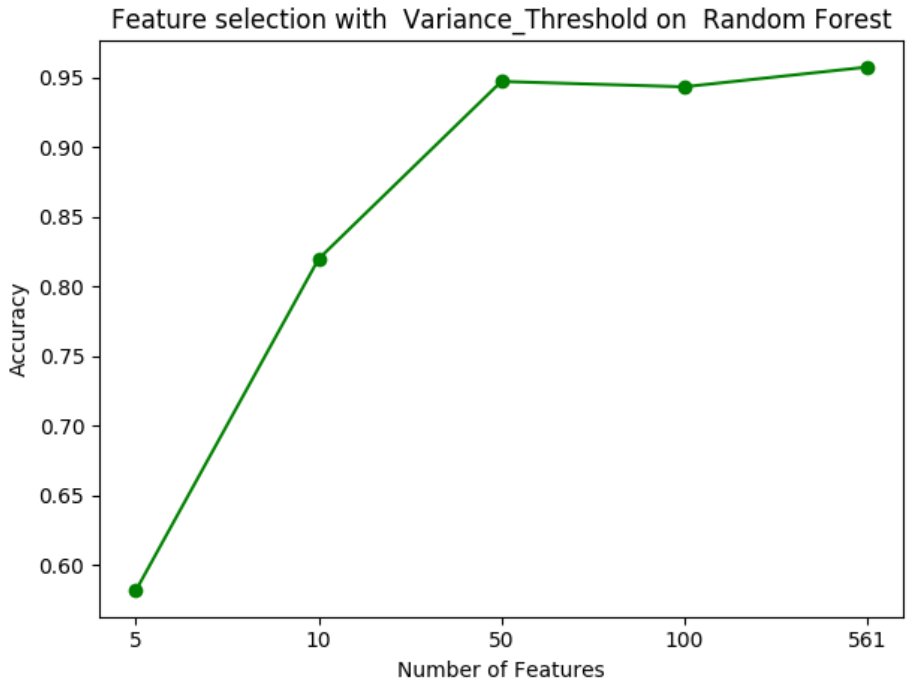



eature selection with Select_From_Model_L1_Based_LSVC on Random Fores
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
Accuracy
5
10
50
100
561
Number of Features


Feature selection with Select_From_Model_Tree_based on Random Forest
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
Accuracy
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_fclassif on Random Forest
Accuracy
0.95
0.90
0.85
0.80
0.75
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_mutual_info_classif on Random Forest
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
5
10
50
100
561
Number of Features


Feature selection with Variance_Threshold on Random Forest
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
50
100
561
Number of Features

For Recursive Feature Elimination_LSVC Features :
#Accuracies = [0.8867438867438867, 0.9407979407979408, 0.9626769626769627,
0.954954954954955, 0.9575289575289575]

##################################
For Select_From_Model_L1_Based_LR Features :
#Accuracies = [0.8803088803088803, 0.9407979407979408, 0.9536679536679536,
0.9588159588159588, 0.9626769626769627]

##################################
For Select_From_Model_L1_Based_LSVC Features :
#Accuracies = [0.7992277992277992, 0.8712998712998713, 0.9485199485199485,
0.9562419562419563, 0.9601029601029601]

##################################
For Select_From_Model_Tree_based Features :
#Accuracies = [0.8133848133848134, 0.8275418275418276, 0.9446589446589446,
0.9588159588159588, 0.9626769626769627]

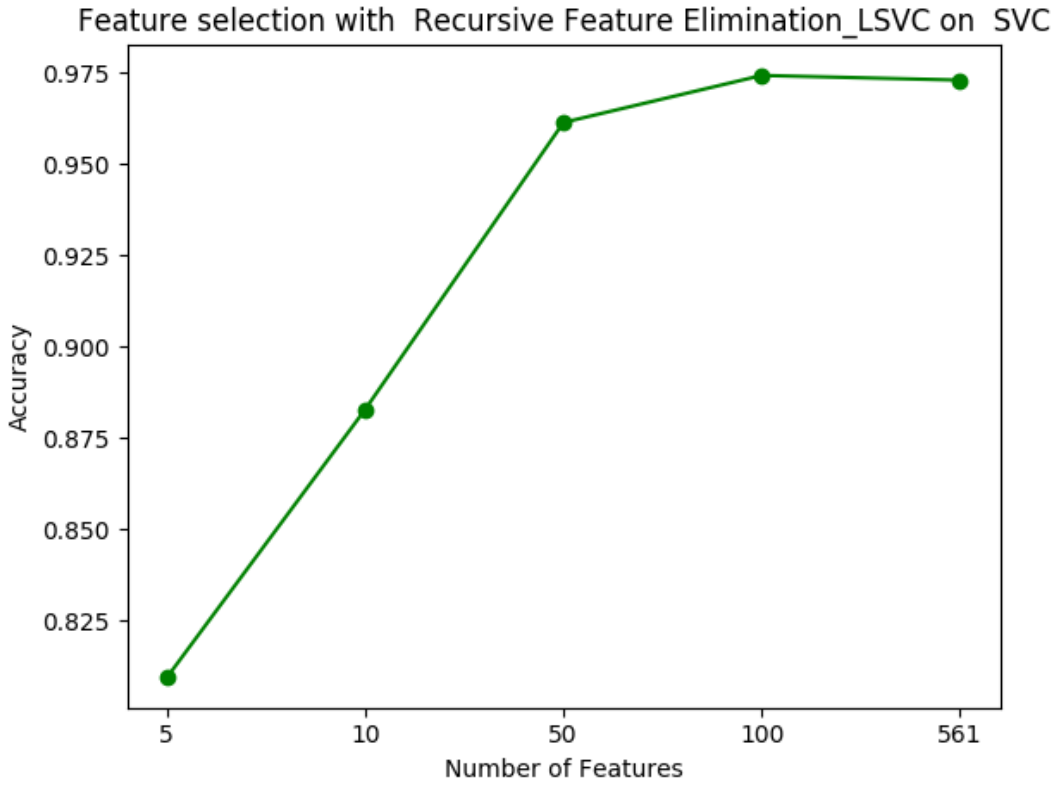
##################################
For Select_Kbest_fclassif Features :
#Accuracies = [0.749034749034749, 0.8082368082368082, 0.9060489060489061,
0.9292149292149292, 0.954954954954955]

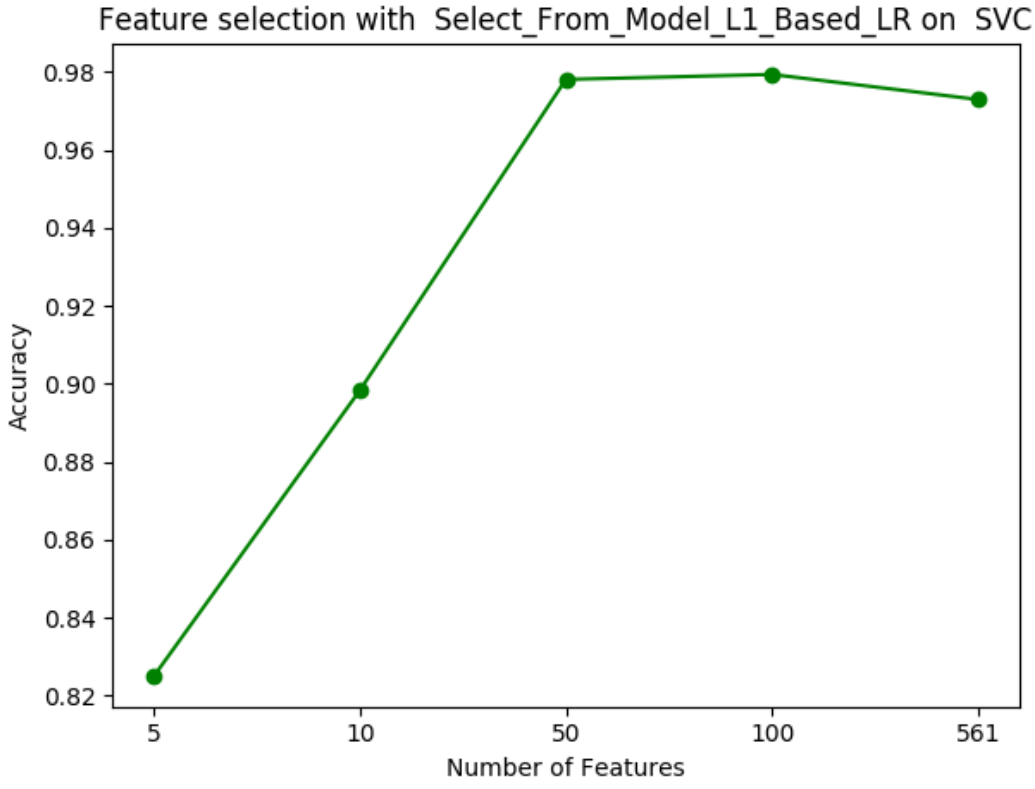
##################################
For Select_Kbest_mutual_info_classif Features :
#Accuracies = [0.6731016731016731, 0.6846846846846847, 0.9163449163449163,
0.9407979407979408, 0.9626769626769627]

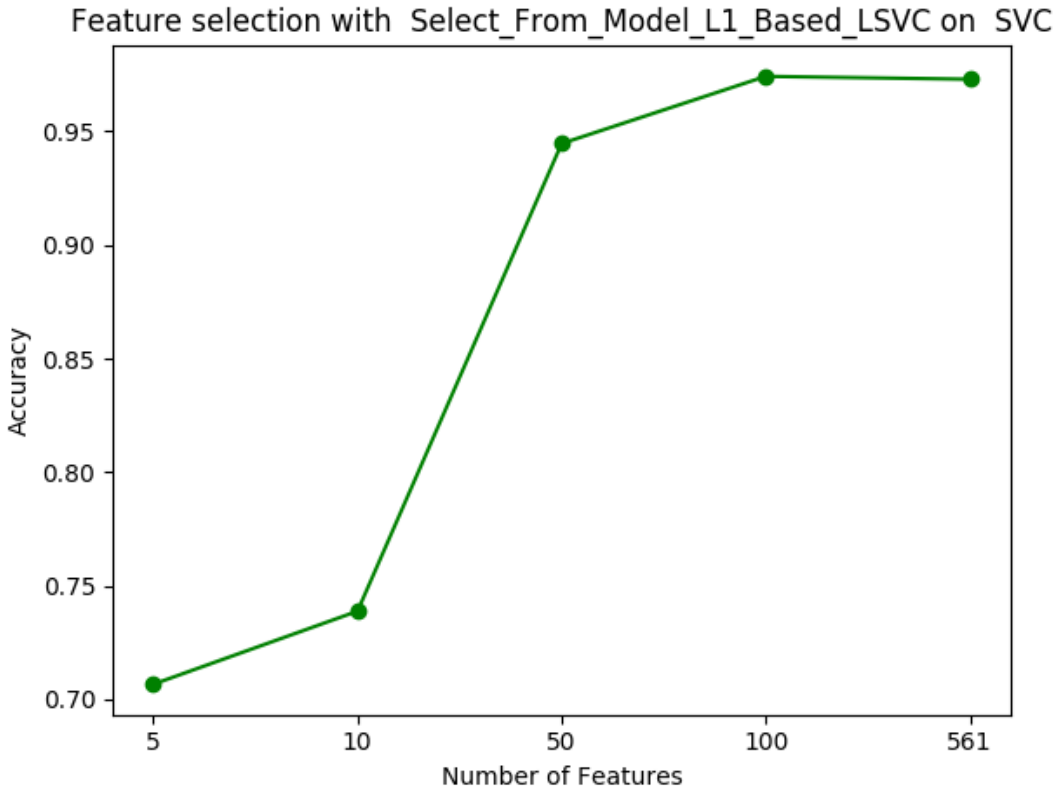
##################################
For Variance_Threshold Features :
#Accuracies = [0.5817245817245817, 0.8198198198198198, 0.9472329472329473,
0.9433719433719434, 0.9575289575289575]

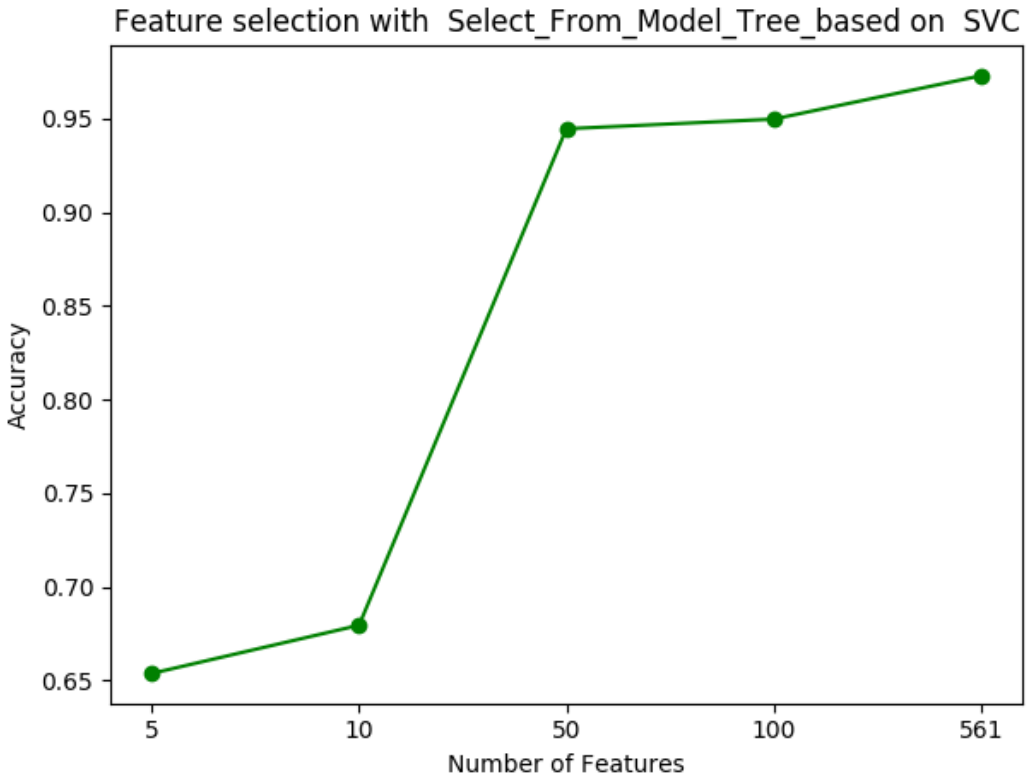
##################################

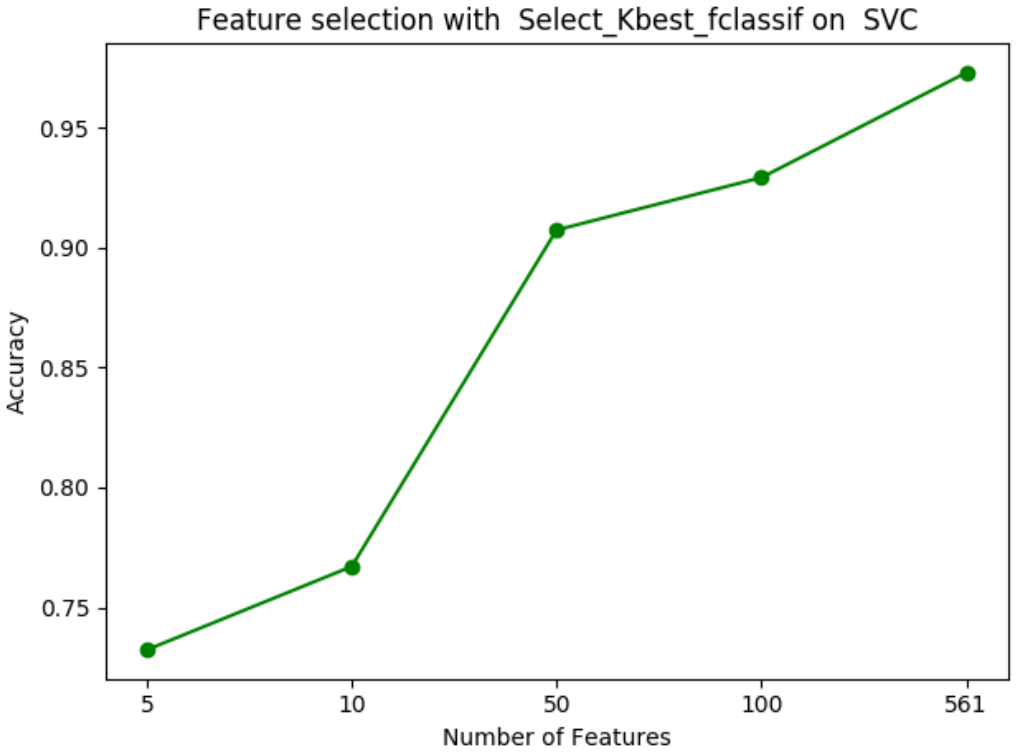
SVC:

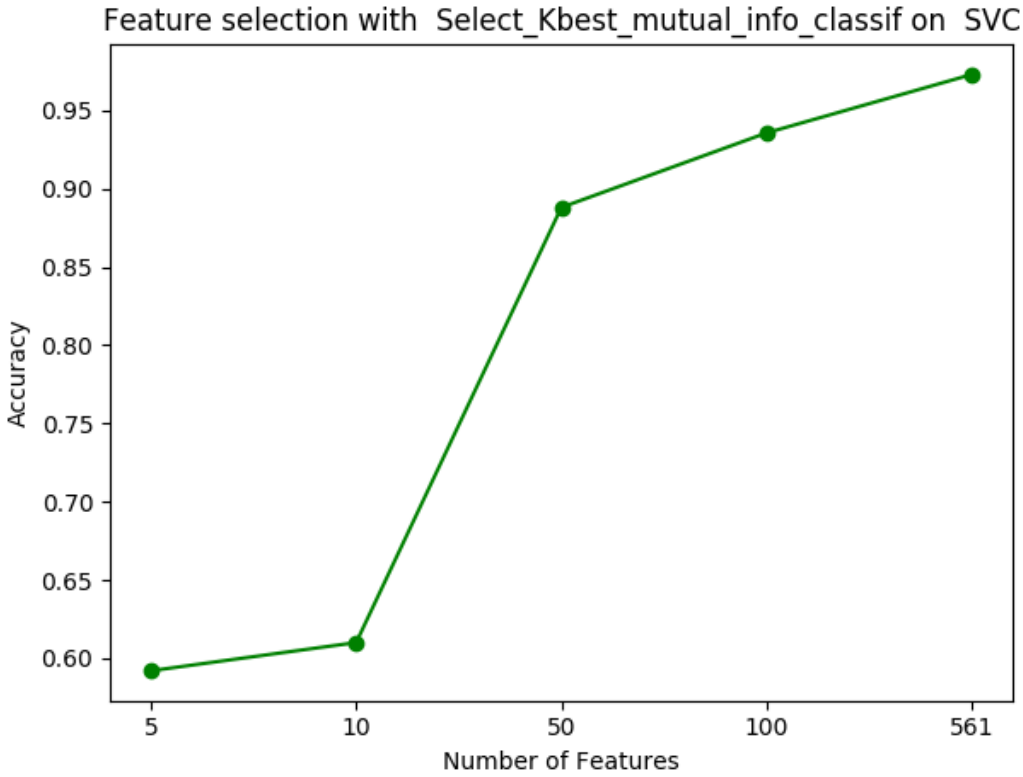

Feature selection with Recursive Feature Elimination_LSVC on SVC
Accuracy
0.975
0.950
0.925
0.900
0.875
0.850
0.825
5
10
50
100
561
Number of Features

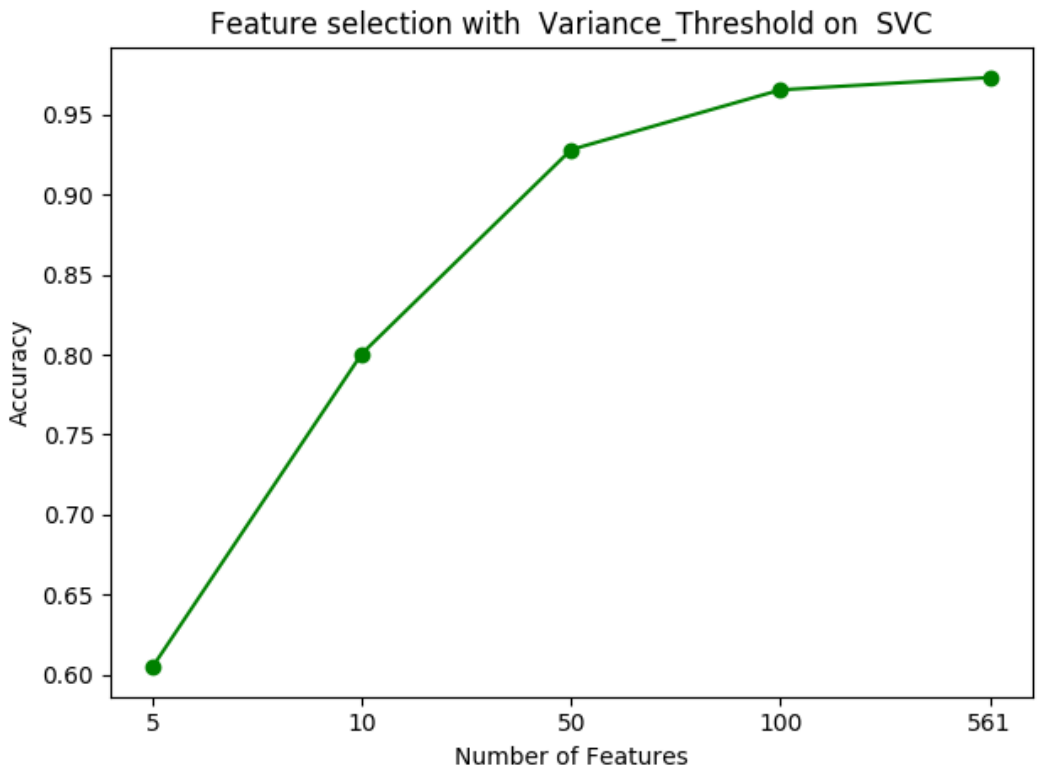

Feature selection with Select_From_Model_L1_Based_LR on SVC
Accuracy
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
5
10
50
100
561
Number of Features


Feature selection with Select_From_Model_L1_Based_LSVC on SVC
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
5
10
50
100
561
Number of Features


Feature selection with Select_From_Model_Tree_based on SVC
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_fclassif on SVC
Accuracy
0.95
0.90
0.85
0.80
0.75
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_mutual_info_classif on SVC
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
50
100
561
Number of Features


Feature selection with Variance_Threshold on SVC
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
50
100
561
Number of Features

For Recursive Feature Elimination_LSVC Features :
#Accuracies = [0.8095238095238095, 0.8828828828828829, 0.9613899613899614,
0.9742599742599742, 0.972972972972973]

##################################
For Select_From_Model_L1_Based_LR Features :
#Accuracies = [0.824967824967825, 0.8983268983268984, 0.9781209781209781,
0.9794079794079794, 0.972972972972973]

##################################
For Select_From_Model_L1_Based_LSVC Features :
#Accuracies = [0.7065637065637066, 0.7387387387387387, 0.9446589446589446,
0.9742599742599742, 0.972972972972973]

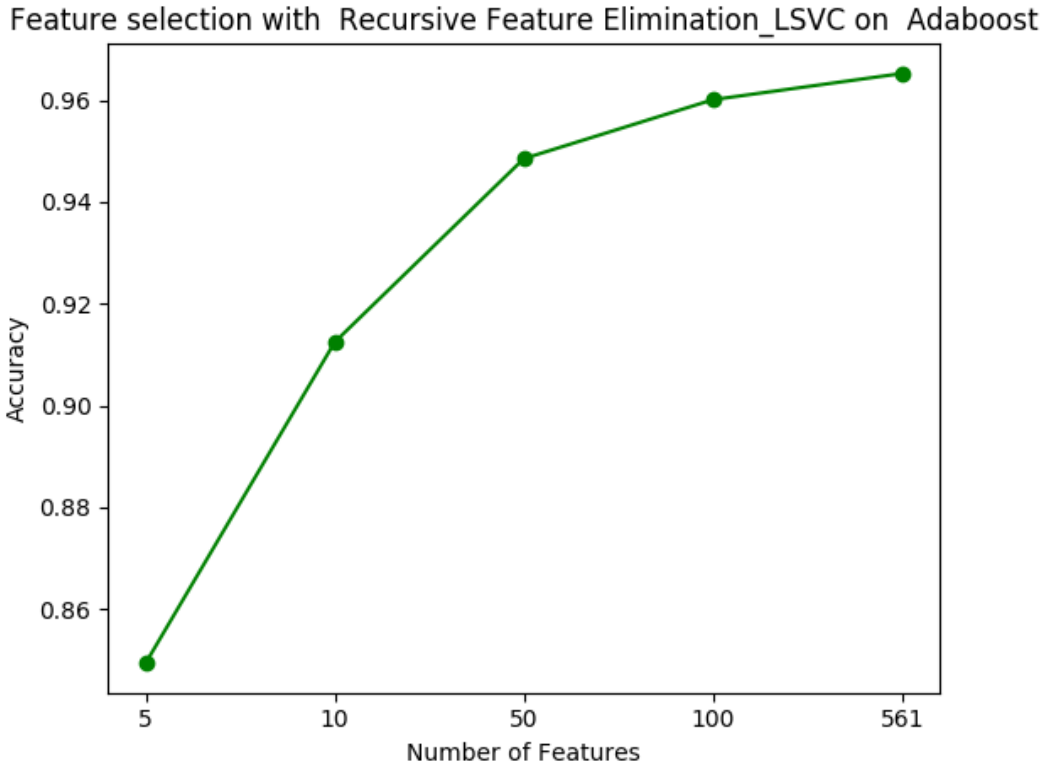
##################################
For Select_From_Model_Tree_based Features :
#Accuracies = [0.6537966537966537, 0.6795366795366795, 0.9446589446589446,
0.9498069498069498, 0.972972972972973]

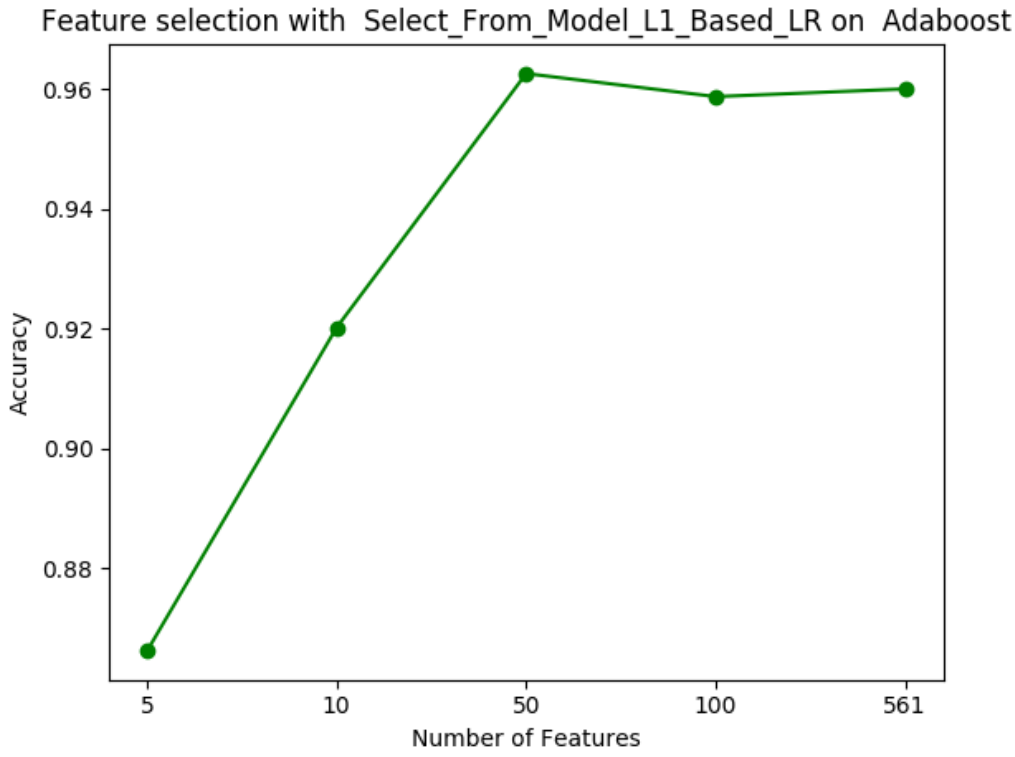
##################################
For Select_Kbest_fclassif Features :
#Accuracies = [0.7323037323037324, 0.767052767052767, 0.9073359073359073,
0.9292149292149292, 0.972972972972973]

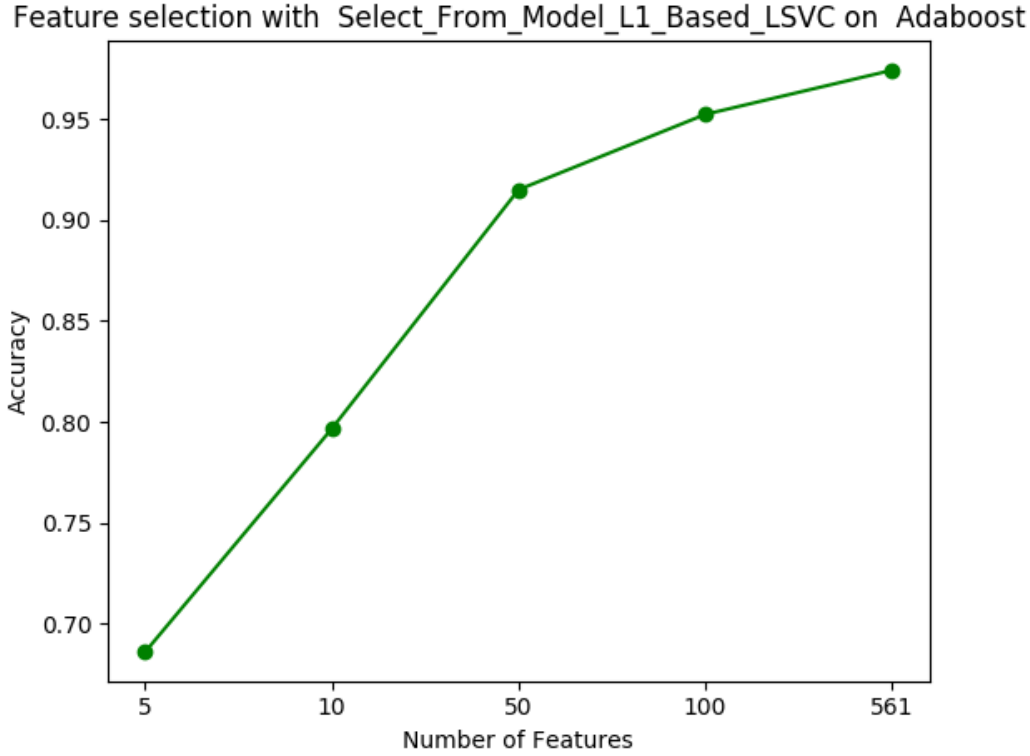
##################################
For Select_Kbest_mutual_info_classif Features :
#Accuracies = [0.592020592020592, 0.61003861003861, 0.888030888030888,
0.9356499356499357, 0.972972972972973]

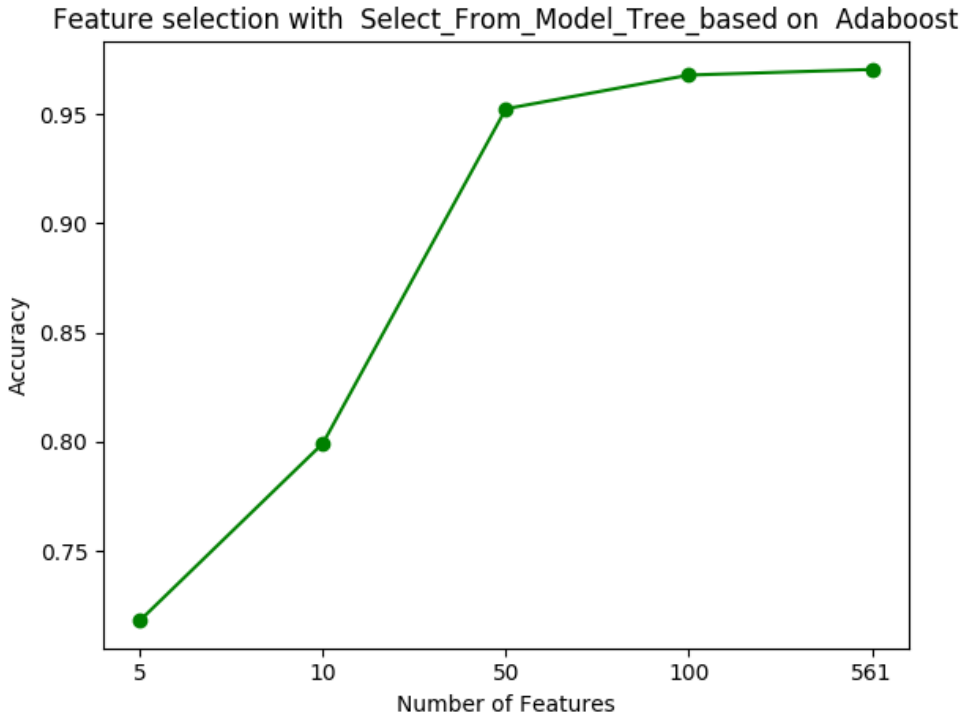
##################################
For Variance_Threshold Features :
#Accuracies = [0.6048906048906049, 0.8005148005148005, 0.9279279279279279,
0.9652509652509652, 0.972972972972973]

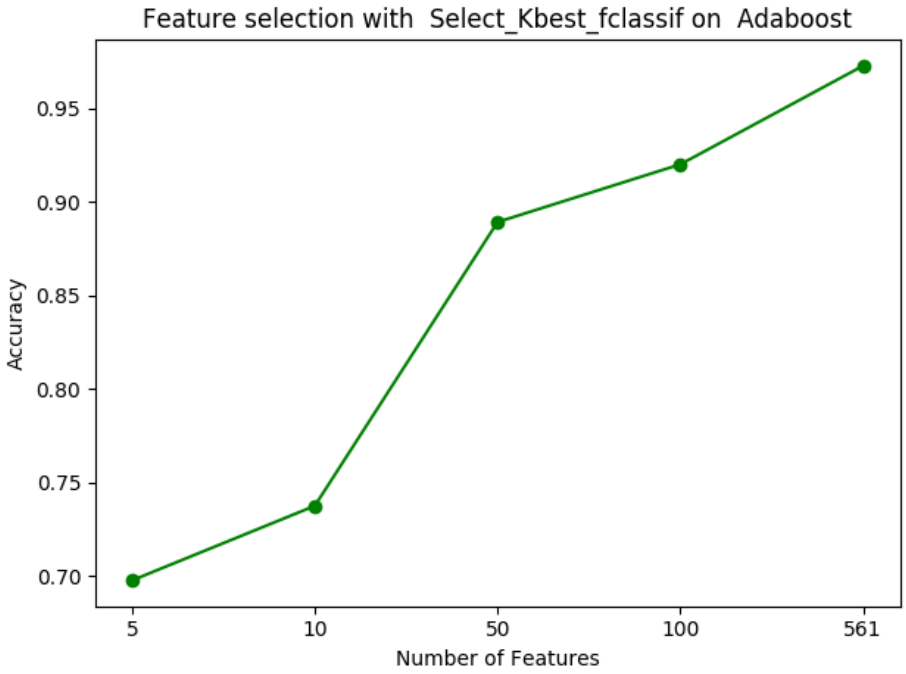
Adaboost:

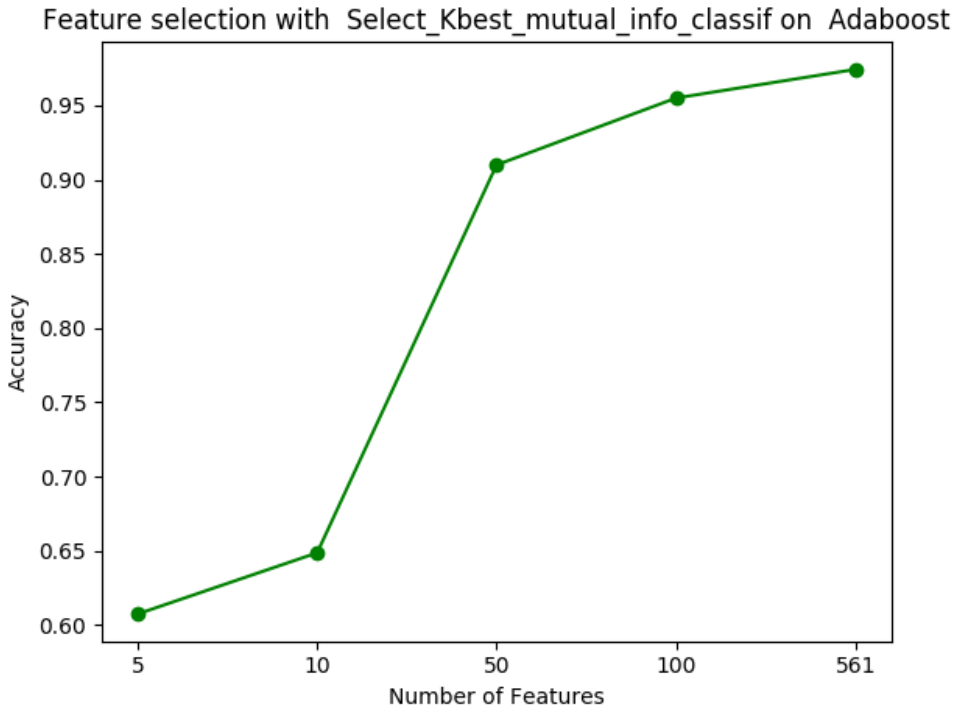

Feature selection with Recursive Feature Elimination_LSVC on Adaboost
Accuracy
0.96
0.94
0.92
0.90
0.88
0.86
5
10
50
100
561
Number of Features

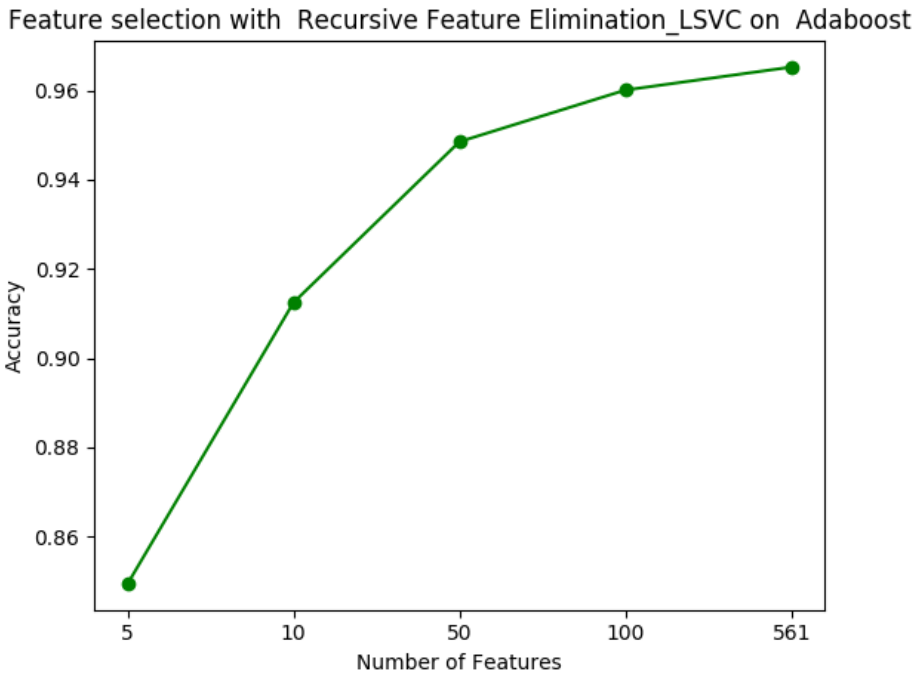

Feature selection with Select_From_Model_L1_Based_LR on Adaboost
Accuracy
0.96
0.94
0.92
0.90
0.88
5
10
50
100
561
Number of Features


Feature selection with Select_From_Model_L1_Based_LSVC on Adaboost
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
5
10
50
100
561
Number of Features


Feature selection with Select_From_Model_Tree_based on Adaboost
Accuracy
0.95
0.90
0.85
0.80
0.75
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_fclassif on Adaboost
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
5
10
50
100
561
Number of Features


Feature selection with Select_Kbest_mutual_info_classif on Adaboost
Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5
10
50
100
561
Number of Features


Feature selection with Recursive Feature Elimination_LSVC on Adaboost
Accuracy
0.96
0.94
0.92
0.90
0.88
0.86
5
10
50
100
561
Number of Features

For Recursive Feature Elimination_LSVC Features :
#Accuracies = [0.8494208494208494, 0.9124839124839125, 0.9485199485199485,
0.9601029601029601, 0.9652509652509652]

##################################
For Select_From_Model_L1_Based_LR Features :
#Accuracies = [0.8661518661518661, 0.9202059202059202, 0.9626769626769627,
0.9588159588159588, 0.9601029601029601]

##################################
For Select_From_Model_L1_Based_LSVC Features :
#Accuracies = [0.685971685971686, 0.7966537966537967, 0.915057915057915,
0.9523809523809523, 0.974259742599742]

##################################
For Select_From_Model_Tree_based Features :
#Accuracies = [0.7181467181467182, 0.7992277992277992, 0.9523809523809523,
0.9678249678249679, 0.9703989703989704]

##################################
For Select_Kbest_fclassif Features :
#Accuracies = [0.6975546975546976, 0.7374517374517374, 0.8893178893178894,
0.9202059202059202, 0.972972972972973]

##################################
For Select_Kbest_mutual_info_classif Features :
#Accuracies = [0.6074646074646075, 0.6486486486486487, 0.9099099099099099,
0.954954954954955, 0.974259742599742]

##################################
For Variance_Threshold Features :
#Accuracies = [0.546975546975547, 0.7554697554697555, 0.9420849420849421,
0.9575289575289575, 0.9781209781209781]

سوال پنجم)

علاوه بر روش های قبل، روش
PCA و TruncatedSVD
استفاده شد اما همان دقت ۸۵ تا ۹۰ درصد گرفته شد.
لذا بهترین دقت را با

Feature selection with Select_From_Model_L1_Based_LR on Adaboost

انجام شد که دقت ۹۲ درصد داد.

خوشه بندی سوال یک)

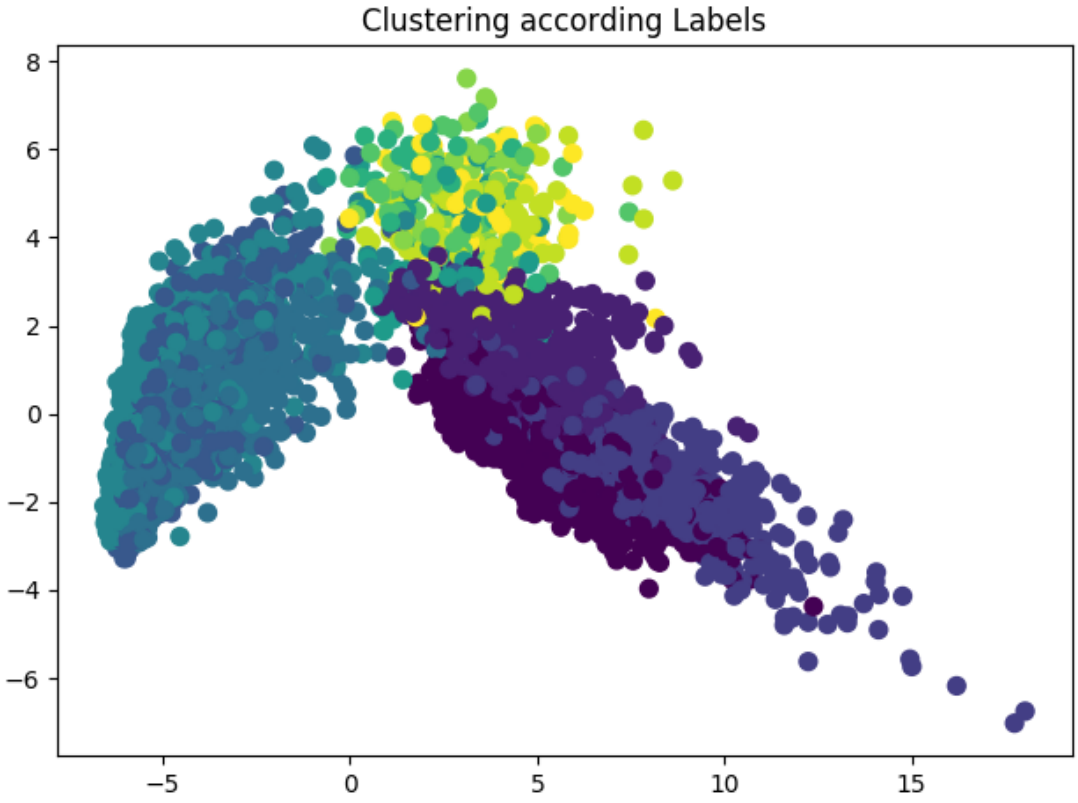
برای هر دو روش با پارامترهای مختلف چندین بار اجرا شد و در پوشه های clustering/kmm و clustering/gmm ذخیره شدند.

بهترین randindex برای kmeans ، 0.39 شد. برای جی ام ام 0.53 شد.

خوشه بندی سوال دو)

توجه کنید در دو نمودار بالا، یک رنگ مشخص برای هردویشان به معنای یک دسته ی یکسان نیست بلکه برای هر عکس، بیانگر یک کلاستر مجزا است. همانگونه که می بینید در
PCA ها به نحو احسن توانسته اند تفاوت ها برای کلاسترینگ را نشان دهند.


Clustering according to 20 PCA feature
8
6
4
2
0
−2
−4
−6
−5
0
5
10
15


Clustering according Labels
8
6
4
2
0
−2
−4
−6
−5
0
5
10
15