

# Reasons for using feature selection:

1. Improve data processing efficiency; reduce memory usage.
2. Reducing the number of features, to reduce overfitting and improve the generalization of models.
3. To gain a better understanding of the features and their relationship to the response variables.

## Feature selection overview

小鼠肾脏代谢组学.xlsx

WT-1 ~ 8 为control group, OIR-1 ~ 8为experiment group. Independent samples.

相关系数、info gain、etc

### 1. Load Data

```
In [2]: import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import cla
```

```
In [3]: # beware: EXCEL may cut off columns and save incomplete data
path = "小鼠肾脏代谢组学.txt"
data = pd.read_csv(path, delimiter = '\t') # ,header=None
X_names = data.iloc[:5022,4].values.tolist()
X = data.iloc[:5022,19:].T.values
y = np.array([0]*int(len(X)/2) + [1]*int(len(X)/2) )

X.shape, y.shape
```

Out[3]: ((16, 5022), (16,))

```
In [4]: import pylab, matplotlib
matplotlib.rcParams.update({'font.size': 18})

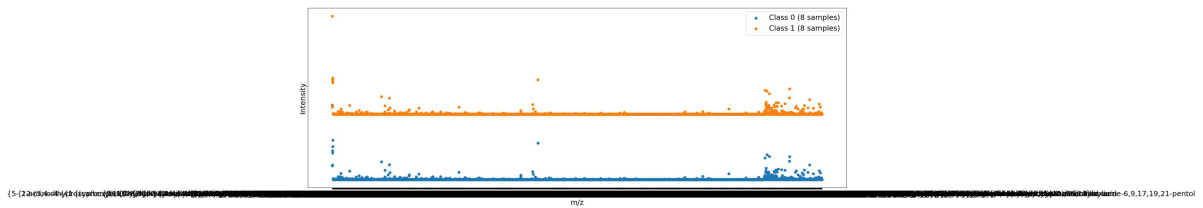
plt.figure(figsize = (24,8))

for c in set(y):
    Xc = X[y == c]
    yc = y[y == c]
    plt.scatter(X_names, np.mean(Xc,axis=0) + c*3*10e6, label= 'Class ' + str(c) +
    plt.legend()

# plt.title(u'Averaged Spectrums for Each Category\n')
plt.xlabel(r'm/z')
```

```
plt.ylabel('Intensity')
plt.yticks([])

matplotlib.rcParams.update({'font.size': 10})
```



## 2. Make a classifiability analysis with cla

```
cla.metrics.get_metrics(X,y)
```

```
{'classification.ACC': 1.0,
 'classification.Kappa': 1.0,
 'classification.F1_Score': 1.0,
 'classification.Jaccard': 1.0,
 'classification.Precision': 1.0,
 'classification.Recall': 1.0,
 'classification.CrossEntropy': 3.8652215771934146e-08,
 'classification.AP': 1.0,
 'classification.Brier': 1.1775589927875984e-14,
 'classification.ROC_AUC': 1.0,
 'classification.PR_AUC': 1.0,
 'classification.BER': 0.0,
 'correlation.IG.max': 0.7253718503718508,
 'correlation.r.max': 0.9473902863233524,
 'correlation.r.p.min': 2.601233179055872e-08,
 'correlation.rho.max': 0.8677218312746248,
 'correlation.rho.p.min': 1.329873727130143e-05,
 'correlation.tau.max': 0.7302967433402214,
 'correlation.tau.p.min': 0.0007775304469403846,
 'test.ES.max': 5.5373692024671515,
 'test.ANOVA.min': 2.601233179055904e-08,
 'test.ANOVA.min.log10': -7.5848207151481475,
 'test.ANOVA.F.max': 122.64983073772686,
 'test.MANOVA': 0.00224708,
 'test.MANOVA.log10': -2.6483814656582725,
 'test.MANOVA.F': 10.1095,
 'test.MWW.min': 0.00046955284955859495,
 'test.MWW.min.log10': -3.328315519521518,
 'test.MWW.U.min': 0.0,
 'test.KS.min': 0.00015540015540015537,
 'test.KS.min.log10': -3.8085485512404054,
 'test.KS.D.max': 1.0,
 'overlapping.F1.mean': 0.29540624419729367,
 'overlapping.F1.sd': 0.2013813320239405,
 'overlapping.F1v.mean': 0.7408278574273198,
 'overlapping.F1v.sd': 0.03313020408590226,
 'overlapping.F2.mean': 0.8125,
 'overlapping.F2.sd': 0.0,
 'overlapping.F3.mean': 0.0,
 'overlapping.F3.sd': 0.0,
 'overlapping.F4.mean': nan,
```

```
'overlapping.F4.sd': 0.008051990067659668,
'neighborhood.N1': 0.031185223435873426,
'neighborhood.N2.mean': 0.2,
'neighborhood.N2.sd': 0.4068381021724862,
'neighborhood.N3.mean': 0.28626741800257505,
'neighborhood.N3.sd': 0.06628412202991937,
'neighborhood.N4.mean': 0.1875,
'neighborhood.N4.sd': 0.4031128874149275,
'neighborhood.T1.mean': 0.016227474706987863,
'neighborhood.T1.sd': 0.06284873929116608,
'neighborhood.LSC': 313.875,
'linearity.L1.mean': 0.3125,
'linearity.L1.sd': 0.0009956192751891676}
```

根据 cla 分析结果，预计分类结果比较理想

### 3. Preview

#### PCA降维 (sklearn.decomposition.PCA)

```
In [5]: from cla import metrics
metrics.unsupervised_dimension_reductions(X, y)
```

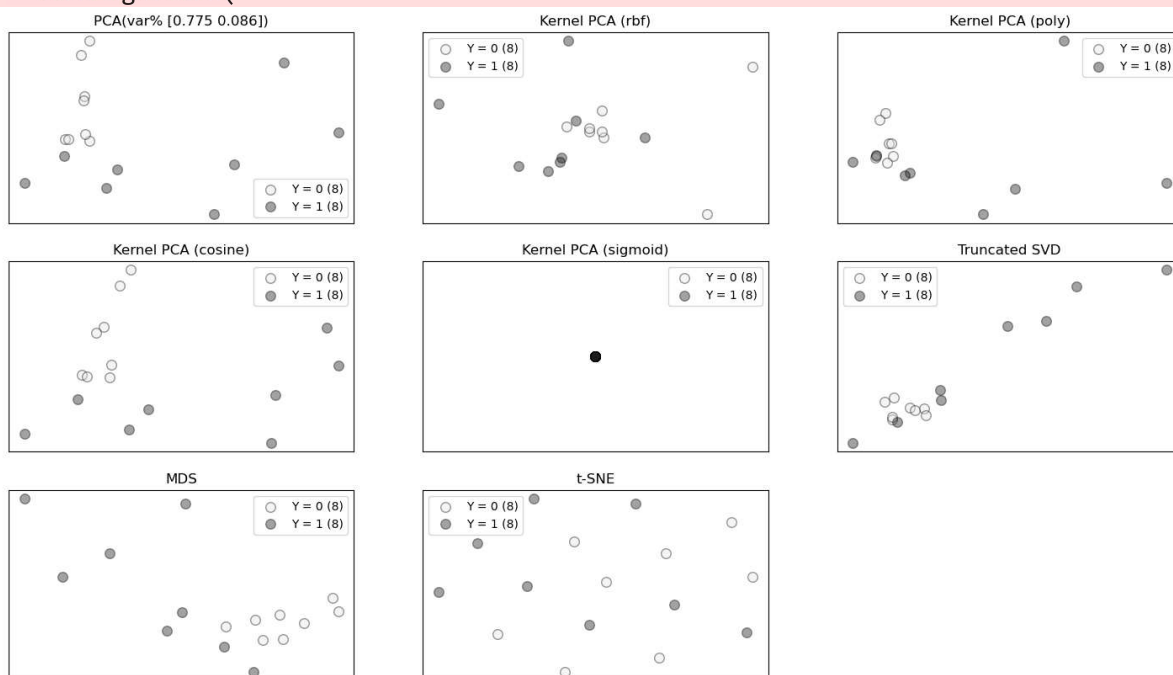
Unable to determine R home: [WinError 2] The system cannot find the file specified  
rpy2 3.X may not support Windows. ECoL metrics may not be available.

C:\Users\eleve\anaconda3\lib\site-packages\sklearn\manifold\\_t\_sne.py:780: FutureWarning: The default initialization in TSNE will change from 'random' to 'pca' in 1.2.

warnings.warn(

C:\Users\eleve\anaconda3\lib\site-packages\sklearn\manifold\\_t\_sne.py:790: FutureWarning: The default learning rate in TSNE will change from 200.0 to 'auto' in 1.2.

warnings.warn(



直接用非监督降维查看数据的可分性如上. 下面进行feature selection，选取有利于分类的特征。

# Feature Selection

```
In [6]: from sklearn.preprocessing import MinMaxScaler

mm_scaler = MinMaxScaler()
X_mm_scaled = mm_scaler.fit_transform(X)

from qsi import fs
_ = fs.RUN_ALL_FS(X_mm_scaled, y, X_names, N = 30)
```

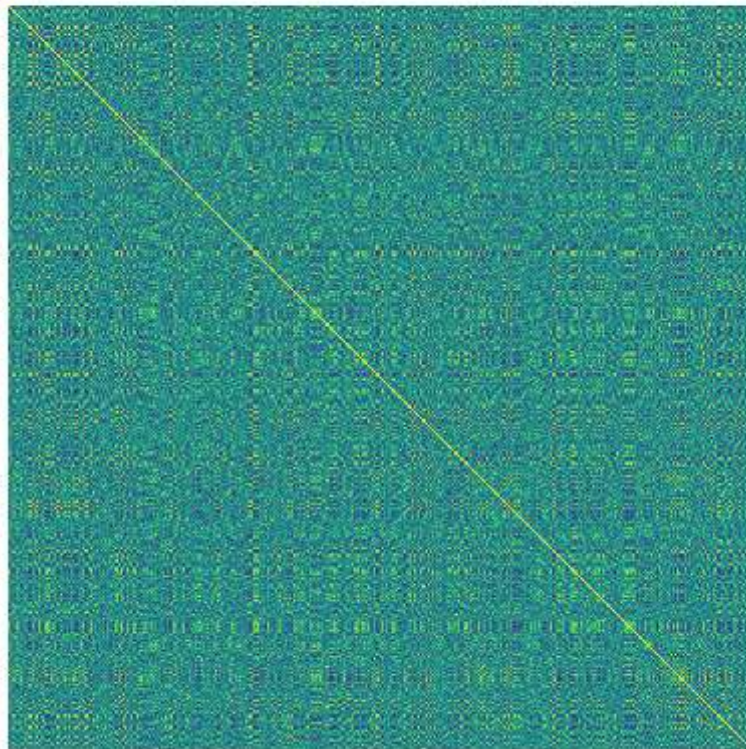
## Univariate feature selection

Univariate feature selection examines each feature individually to determine the strength of the relationship of the feature with the response variable.

### pearson-r

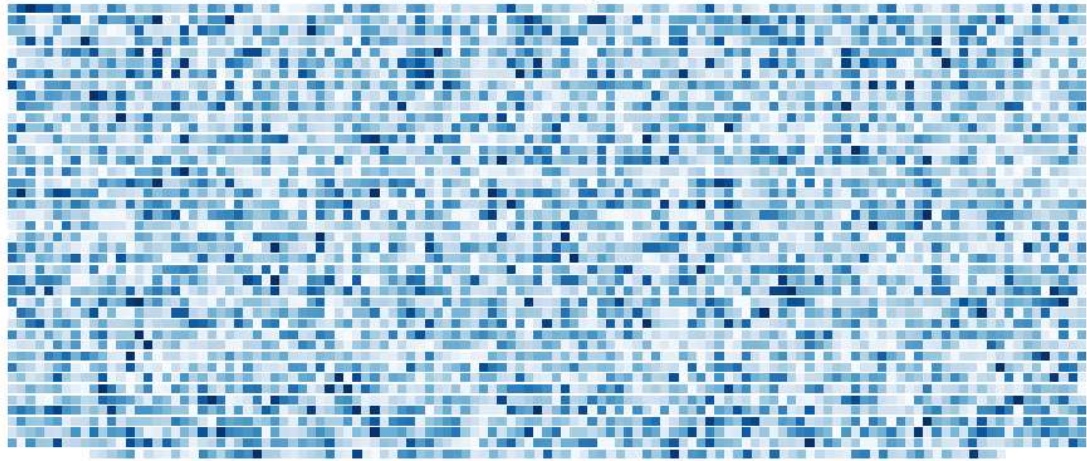
<Figure size 2000x2000 with 0 Axes>

Correlation Coef Matrix between all the X and y.  
y is at the last row/col.

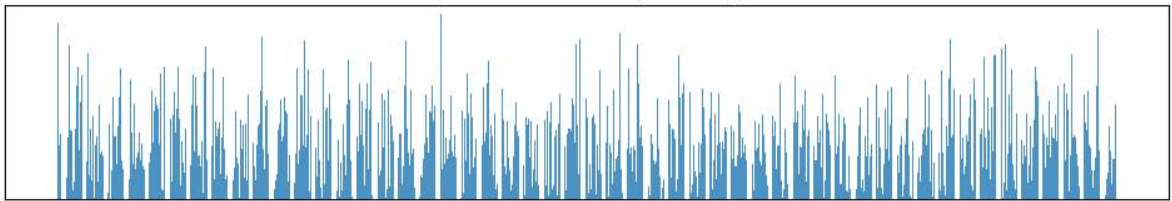




feature-wise coefs/values  
importance marked by color depth



feature-wise coefs/values  
importance marked by bar height



Important feature Number: 30

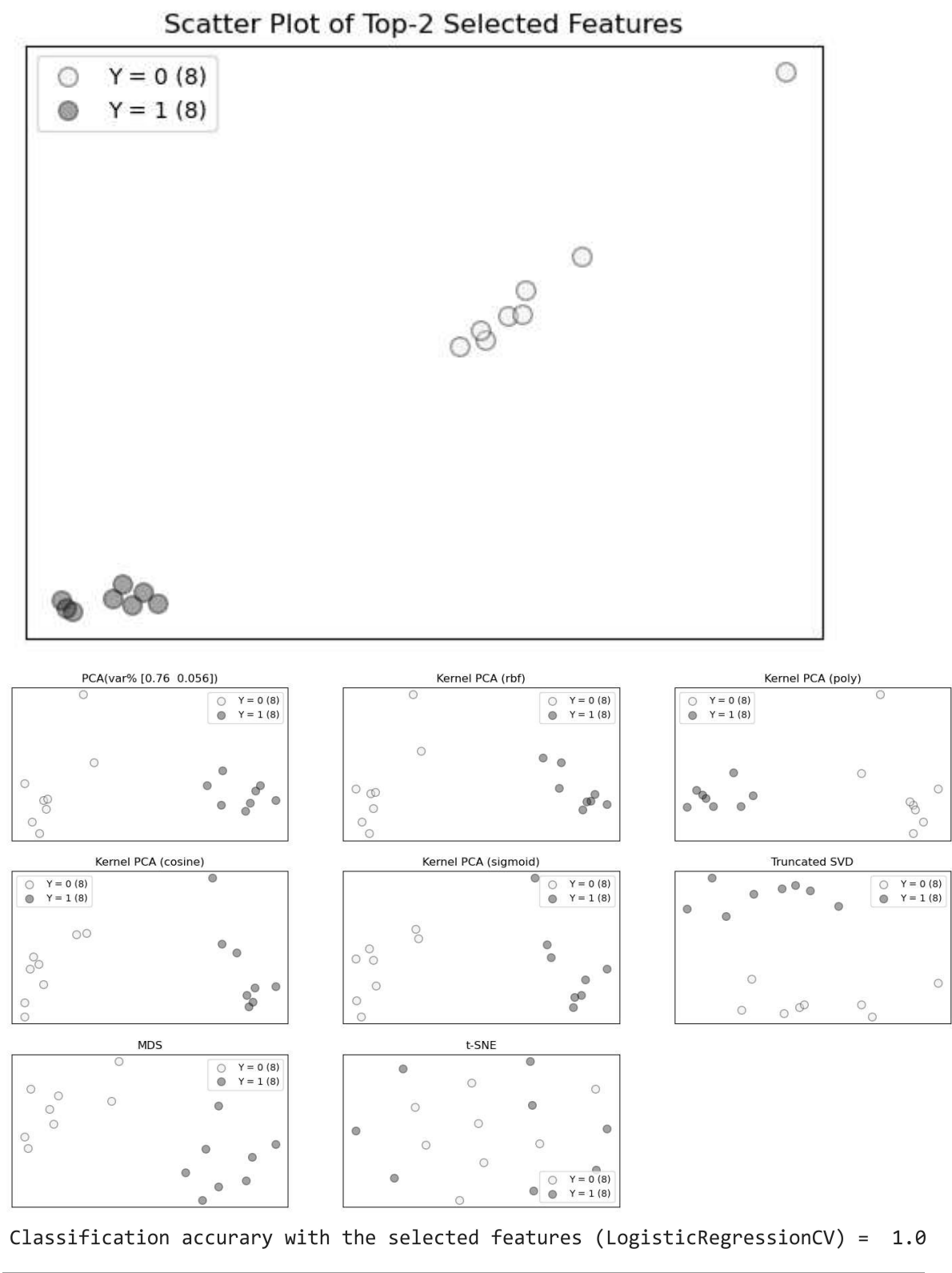
Important features indices: [1819 969 2495 2 1212 4938 3954 3853 4431 2669 37  
35 4888 4984 4342  
4720 5017 970 4069 1510 2479 4703 4237 2956 4780 4210 1653 1172 1480  
1399 4349]

Important features names: ['Shikimic acid' 'Pimelic acid' 'N-phenethyltridecanami  
de'

'PC(14:1(9Z)/20:2(11Z,14Z))' 'Alanine' '2,4,6-octatrienal'  
'28-Homobrassinolide' 'Ethyl isopropyl disulfide' 'L-Histidine'  
'3-Indoxyl phosphate' 'bismuth subgallate' '3-Methylbenzaldehyde'  
'5-Acetoxy-tetradeca-2E,4E,6E-trienoic acid' 'SM(d18:1/18:1(9Z))'  
'2,3-dihydrobenzofuran' '6-Methylquinoline' 'Methylsuccinic acid'  
'N-Acetyl-9-O-acetylneuraminic acid' 'G? 6983'  
'3,4-Dihydroxyphenyl ethanol' 'Tetralin'  
'MG(0:0/22:4(7Z,10Z,13Z,16Z)/0:0)' 'Dihydronepetalactone' 'Guanine'  
'Daimuron' '1-O-(2R-methoxy-hexadecyl)-sn-glycerol' '2-Aminopyrazine'  
'Quinoxaline-2-carboxylic acid' "N3'-Acetyltobramycin"  
'L-3-Hydroxykynurenine']

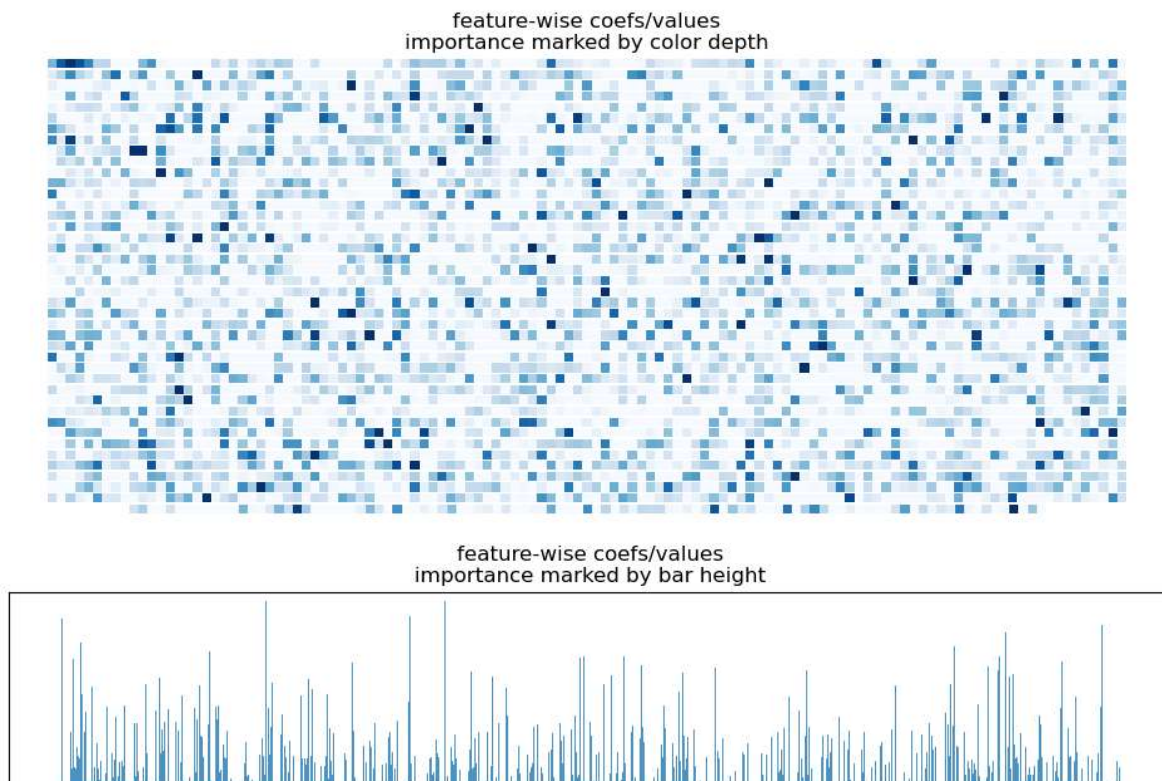
Top-30 feature Importance: [0.94739029 0.93587422 0.90749876 0.90385268 0.8931008  
3 0.8759588

0.85746906 0.85701622 0.85533418 0.85349397 0.85151067 0.85056171  
0.84639826 0.84611839 0.83928012 0.83730333 0.83238263 0.82996959  
0.82934025 0.82363768 0.82048911 0.81979575 0.81848152 0.8171575  
0.81621014 0.81501828 0.81201075 0.81134593 0.81001194 0.80670644]



info-gain / mutual information

Information gain has been used in decision tree. For a specific feature, Information gain (IG) measures how much "information" a feature gives us about the class.  $IG(Y|X)=H(Y)-H(Y|X)$  IG/MI returns zero for independent variables and higher values the more dependence there is between the variables (can be used to rank features by their independence). In information theory, IG answers "if we transmit Y, how many bits can be saved if both sender and receiver know X?" Or "how much information of Y is implied in X?" Attribute/feature X with a high IG is a good split on Y. Pearson r only captures linear correlations, while information gain also captures non-linear correlations.

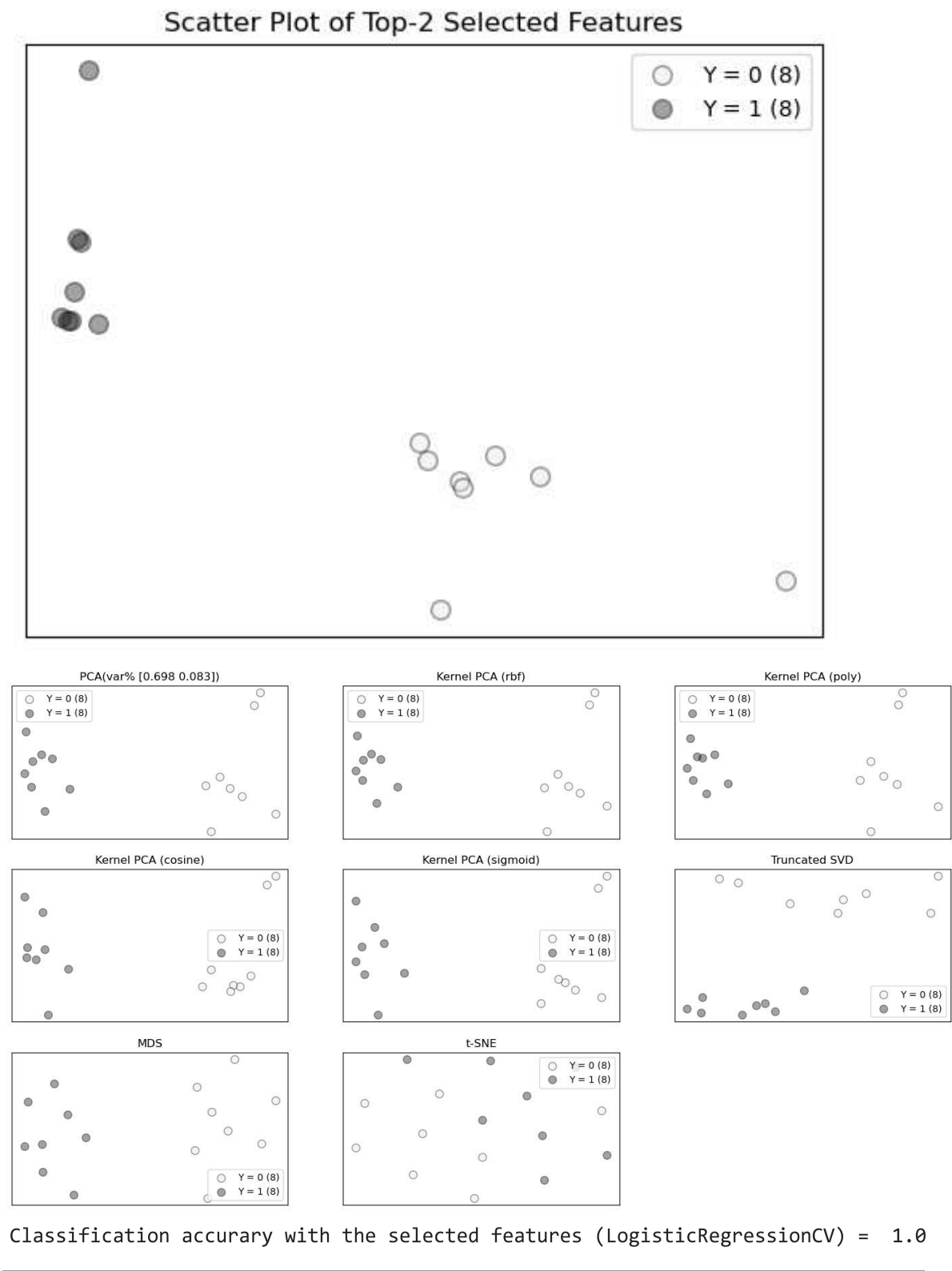


Important feature Number: 30

Important features indices: [ 969 3954 1819 970 2495 4817 1212 2093 5017 1653  
2 527 3254 4888  
4349 4938 3614 4703 4906 4480 1510 1254 3550 4984 2500 3 1123 91  
136 935]

Important features names: ['Pimelic acid' '28-Homobrassinolide' 'Shikimic acid'  
'Methylsuccinic acid' 'N-phenethyltridecanamide'  
'PE(18:1(11Z)/20:4(5Z,8Z,11Z,14Z))' 'Alanine'  
'S1P Lyase Fluorogenic Substrate' '6-Methylquinoline'  
'1-O-(2R-methoxy-hexadecyl)-sn-glycerol' 'PC(14:1(9Z)/20:2(11Z,14Z))'  
'7-Chloroemodin' 'p-Mentha-1,3,5,8-tetraene' '3-Methylbenzaldehyde'  
'L-3-Hydroxykynurenine' '2,4,6-octatrienal' '2-Heptanethiol' 'Tetralin'  
'1-(3-Methyl-2-butenoyl)-6-apiosylglucose' 'L-Histidinol' 'G? 6983'  
'Absciscic alcohol'  
'(3beta,5alpha,6beta,7alpha,22E,24R)-Ergosta-8,22-diene-3,5,6,7-tetrol'  
'5-Acetoxy-tetradeca-2E,4E,6E-trienoic acid' '(Z)-13-Hexadecenoic acid'  
'L-Isoleucine' 'Debrisoquine' 'Isovalerylcarnitine' 'Nicotyrine'  
'L-Agaritine']

Top-30 feature Importance: [0.72537185 0.72537185 0.72537185 0.72537185 0.7045385  
2 0.70453852  
0.70453852 0.68891352 0.67641352 0.66808018 0.65706828 0.64925578  
0.64516352 0.63623495 0.63623495 0.63606133 0.62060995 0.62060995  
0.615228 0.60349685 0.59456828 0.59439467 0.58675578 0.57946411  
0.57356133 0.56644328 0.56644328 0.56592245 0.56592245 0.56592245]

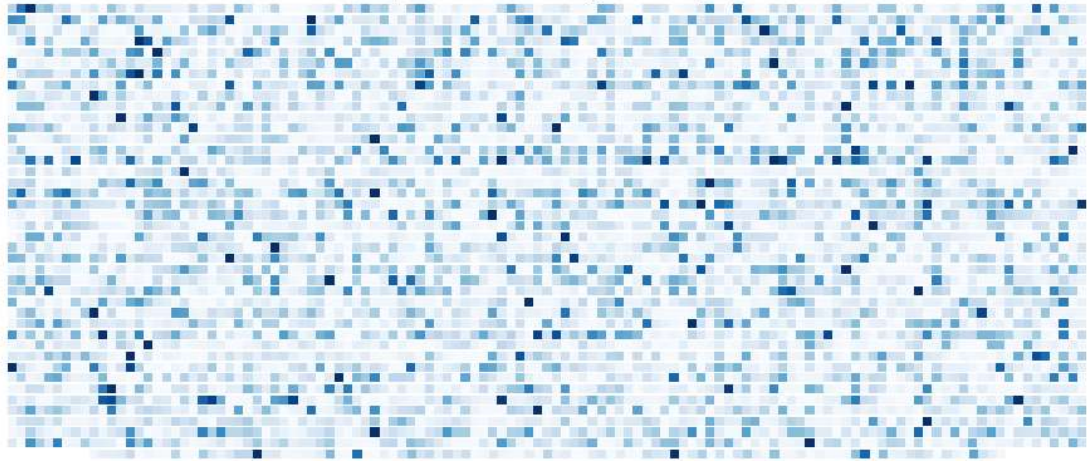


## chi-squared statistic

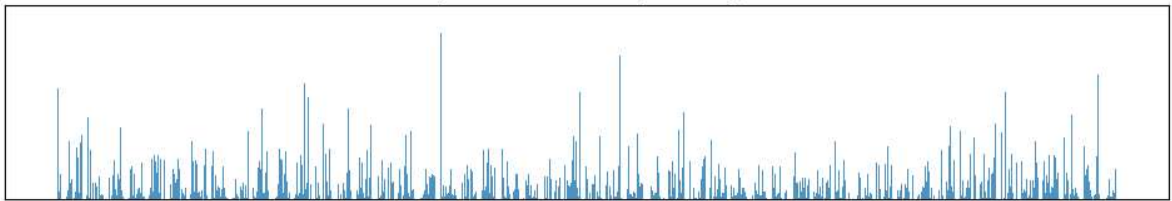
This score can be used to select the `n_features` features with the highest values for the test chi-squared statistic from `X`, which must contain only non-negative features such as booleans or frequencies (e.g., term counts in document classification), relative to the classes. Recall that the chi-square test measures dependence between stochastic variables, so using this function “weeds out” the features that are the most likely to be independent of class and therefore irrelevant for classification.



feature-wise coefs/values  
importance marked by color depth



feature-wise coefs/values  
importance marked by bar height



Important feature Number: 30

Important features indices: [ 969 1819 1480 1212 2669 4661 4888 1254 3735 4720 4984 4938 2495 3297

1172 2 4498 2479 4211 3555 1190 4431 684 4279 5005 4780 4374 323 3853 1098]

Important features names: ['Pimelic acid' 'Shikimic acid' 'Quinoxaline-2-carboxylic acid' 'Alanine'

'3-Indoxyl phosphate' '2-Phenylethanol' '3-Methylbenzaldehyde'

'Absciscic alcohol' 'bismuth subgallate' '2,3-dihydrobenzofuran'

'5-Acetoxy-tetradeca-2E,4E,6E-trienoic acid' '2,4,6-octatrienal'

'N-phenethyltridecanamide' '4-phenyl-5-methyl-1,2,3-Thiadiazole'

'2-Aminopyrazine' 'PC(14:1(9Z)/20:2(11Z,14Z))' 'Guanosine'

'3,4-Dihydroxyphenyl ethanol'

'N-Acetyl-5'-hydroxysulfapyridine glucuronide"

'1-Propanamine, N,N-dimethyl-3-(5-oxidodibenzo[b,e]thiepin-11(6H)-ylidene)-, (E)-(9CI)'

'PC(2:0/0:0)[U]' 'L-Histidine' 'Osmundalactone' 'PS(0-16:0/19:1(9Z))'

'Lotaustralin' 'Guanine' 'Piretanide glucuronide' '7-Methylinosine'

'Ethyl isopropyl disulfide' 'Azaserine']

Top-30 feature Importance: [4.38254112 3.95816283 3.48227071 3.42887804 3.40606998 3.26942337

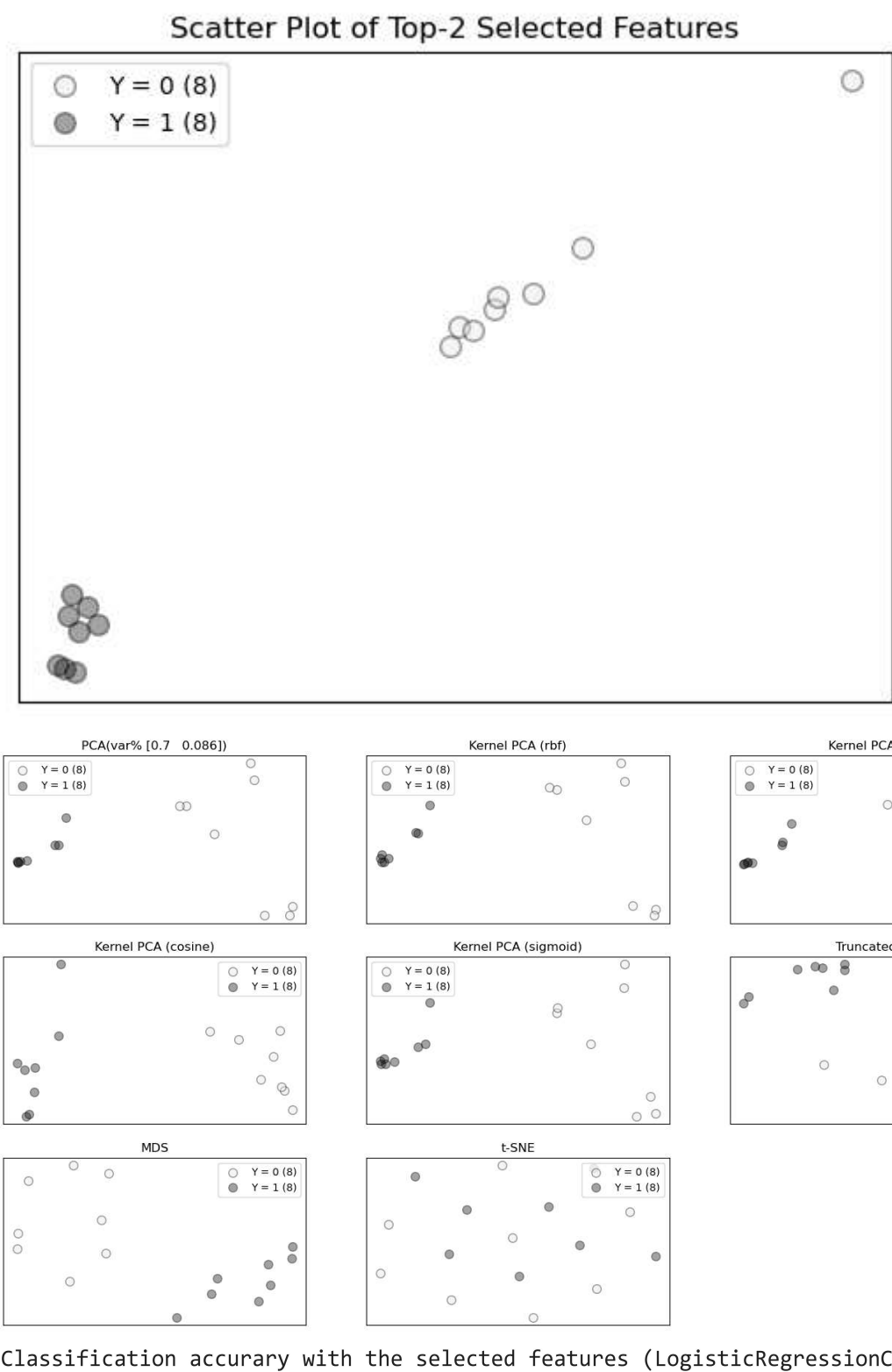
3.11345615 3.07822911 3.05681359 3.02590915 2.96064513 2.9498349

2.8041542 2.79085743 2.73969629 2.63293886 2.55840048 2.53883228

2.5015516 2.47126538 2.43034889 2.42686171 2.4217823 2.41729243

2.34398231 2.33677065 2.33180459 2.33088985 2.3209658 2.28212166]

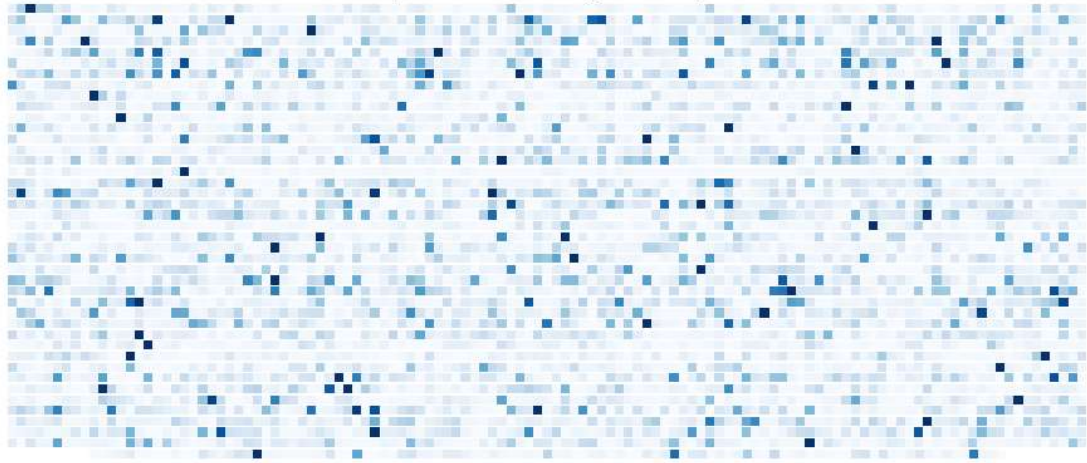




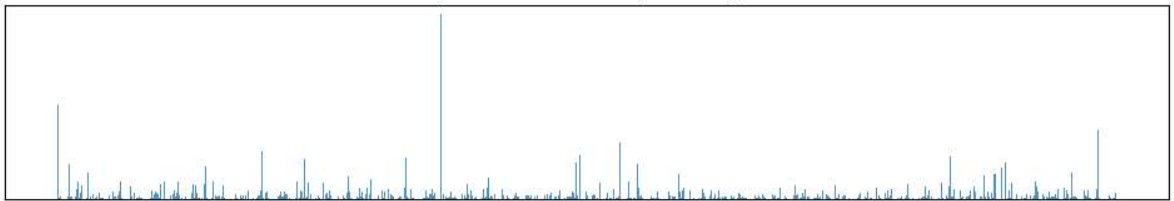
**anova statistic**

ANOVA F-value

feature-wise coefs/values  
importance marked by color depth



feature-wise coefs/values  
importance marked by bar height



Important feature Number: 30

Important features indices: [1819 969 2495 2 1212 4938 3954 3853 4431 2669 37 35 4888 4984 4342

4720 5017 970 4069 1510 2479 4703 4237 2956 4780 4210 1653 1172 1480

1399 4349]

Important features names: ['Shikimic acid' 'Pimelic acid' 'N-phenethyltridecanamide'

'PC(14:1(9Z)/20:2(11Z,14Z))' 'Alanine' '2,4,6-octatrienal'

'28-Homobrassinolide' 'Ethyl isopropyl disulfide' 'L-Histidine'

'3-Indoxyl phosphate' 'bismuth subgallate' '3-Methylbenzaldehyde'

'5-Acetoxy-tetradeca-2E,4E,6E-trienoic acid' 'SM(d18:1/18:1(9Z))'

'2,3-dihydrobenzofuran' '6-Methylquinoline' 'Methylsuccinic acid'

'N-Acetyl-9-O-acetylneuraminic acid' 'G? 6983'

'3,4-Dihydroxyphenyl ethanol' 'Tetralin'

'MG(0:0/22:4(7Z,10Z,13Z,16Z)/0:0)' 'Dihydronepetalactone' 'Guanine'

'Daimuron' '1-O-(2R-methoxy-hexadecyl)-sn-glycerol' '2-Aminopyrazine'

'Quinoxaline-2-carboxylic acid' 'N3'-Acetyltobramycin"

'L-3-Hydroxykynurenine']

Top-30 feature Importance: [122.64983074 98.77641062 65.34439001 62.48169429 55.17990667

46.16428701 38.88071223 38.72609131 38.16028534 37.55625285

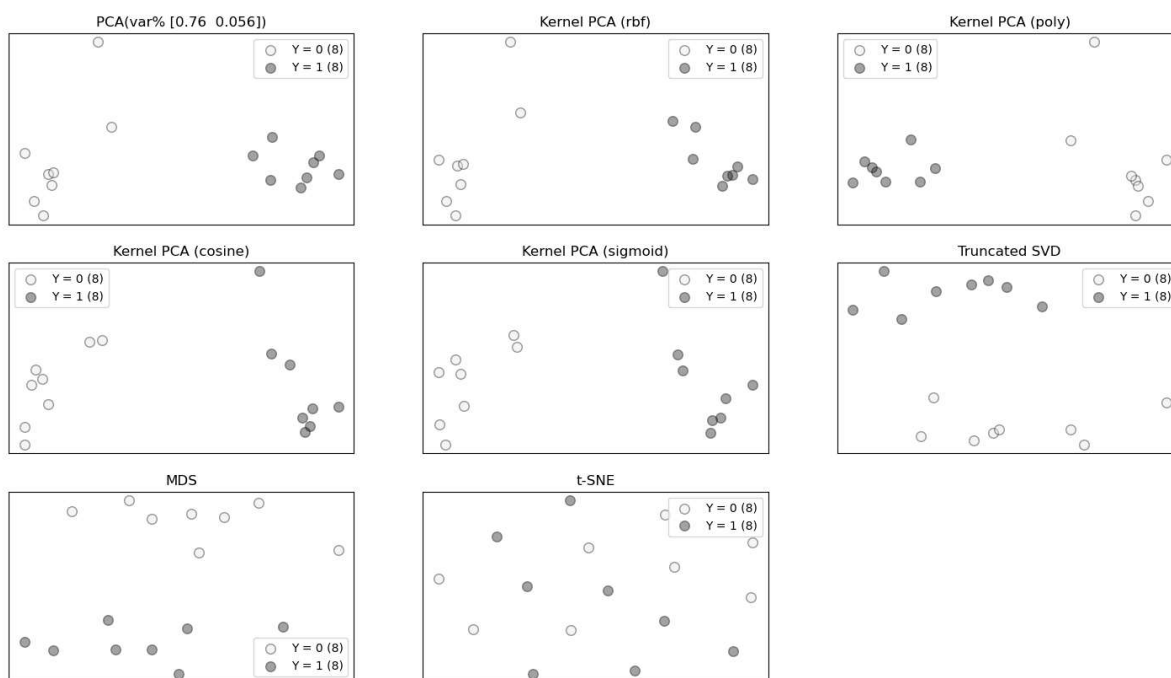
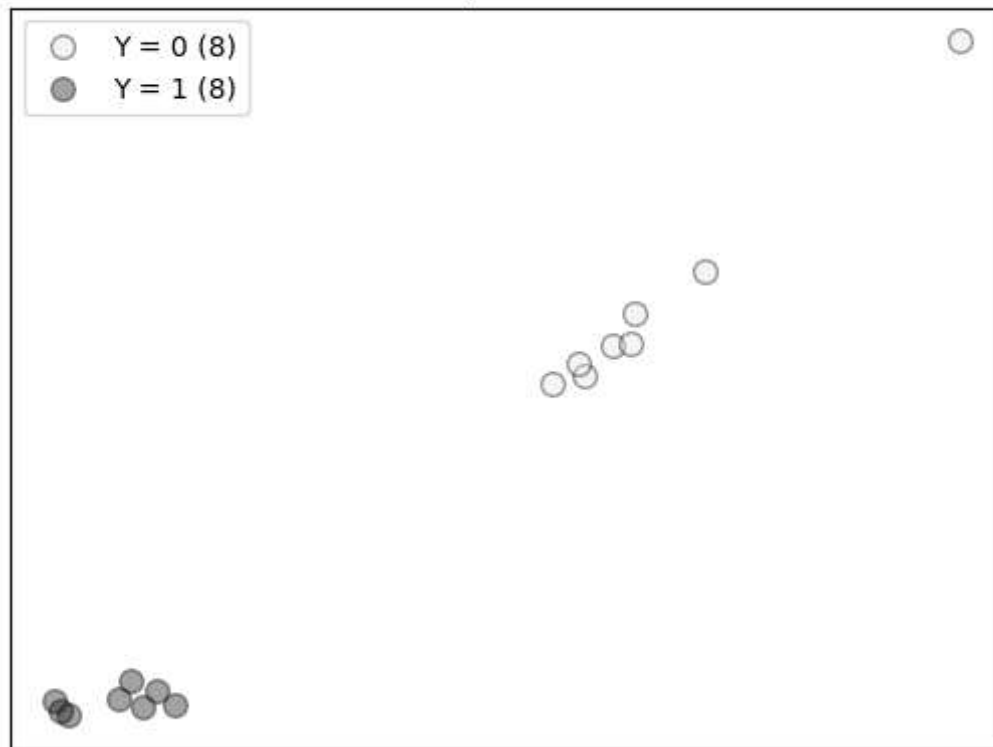
36.92213016 36.62471157 35.36356432 35.28125449 33.35987618

32.83478208 31.58194405 30.9943056 30.8438028 29.52949864

28.83996905 28.69139691 28.412932 28.13648681 27.94115528

27.69828756 27.09928368 26.96948065 26.71179792 26.08880033]

## Scatter Plot of Top-2 Selected Features



Classification accuracy with the selected features (LogisticRegressionCV) = 1.0

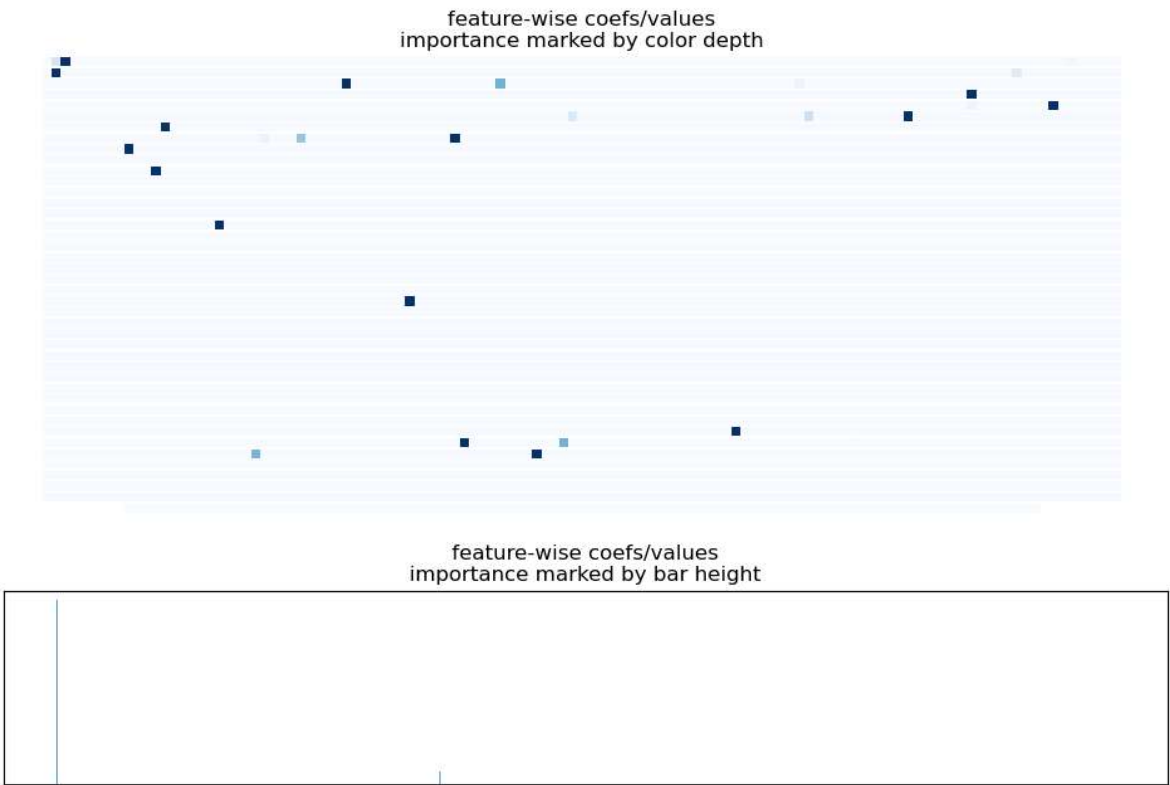
Generally speaking, univariate methods dont generate sparsity and have weak FS effect.

## Model based ranking

Use a machine learning method to build a discriminative model for the response variable using each individual feature, and measure the performance of each model.

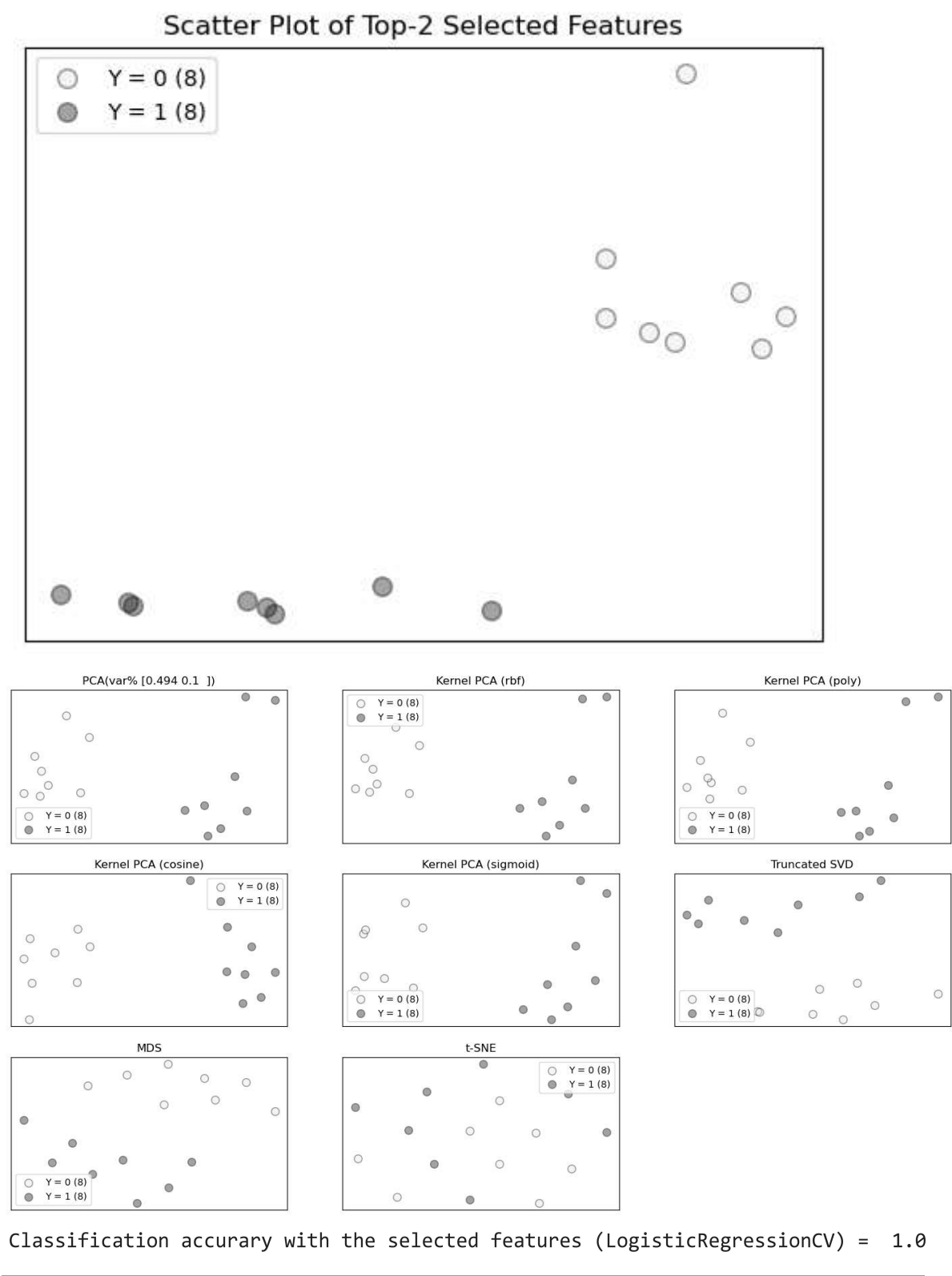
lasso

R2 = 1.0  
LASSO alpha = 0.000156  
Non-zero feature coefficients: 29



Important feature Number: 29  
Important features indices: [ 2 969 121 695 462 733 273 1 885 1212 42 46 290 4374 1819 868 4257 4156 684 591 4343 227 113 658 2680 323 864 582 4289 619]  
Important features names: ['PC(14:1(9Z)/20:2(11Z,14Z))' 'Pimelic acid' 'Zolmitriptan' 'O-propanoyl-D-carnitine' '4-(Nitrosoamino)-1-(3-pyridinyl)-1-butanone' '(E)-1-O-Cinnamoyl-beta-D-glucose' 'N-Acetyl-D-glucosamine' 'PE(P-18:0/20:4(6E,8Z,11Z,14Z)(5OH[S]))' '12-Hydroxydodecanoic acid' 'Alanine' 'Leucokinin I' 'AMINOHYDROXYBUTYRIC ACID' 'Piretanide glucuronide' 'Shikimic acid' 'Plastoquinol-1' 'Ramontoside' 'Phyllanthusol B' 'Osmundalactone' 'Polidocanol' 'Indole-3-carboxylic acid' 'Talaromycin A' 'ANETHOLE' 'Isoquinoline' 'Dihydropanaxacol' '7-Methylinosine' '15-methyl-heptadecanoic acid' '15-octadecynoic acid' 'N6,N6,O-Tridemethylpuromycin-5'-phosphate' '(±)-2-Methylthiazolidine']  
Top-29 feature Importance: [5.06181843e-01 4.56729203e-01 1.26132292e-01 9.58700504e-02 9.12149627e-02 8.66399945e-02 8.12756924e-02 8.00731510e-02 6.93869617e-02 5.29379084e-02 4.93599166e-02 3.90435598e-02 3.72638790e-02 3.55686065e-02 2.71298200e-02 2.37122969e-02 2.17101755e-02 1.97823301e-02 1.81763935e-02 1.73652778e-02 1.50270152e-02 1.49744275e-02 1.28398491e-02 7.50725538e-03 4.55094526e-03 3.92768393e-03 8.85098511e-04 2.52521335e-04 2.70674582e-05]





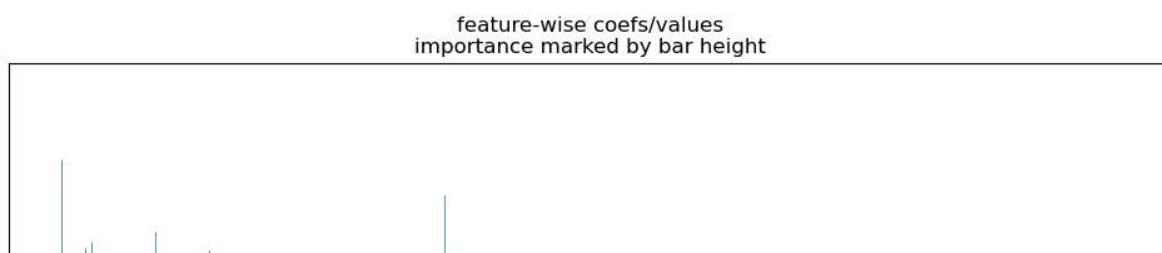
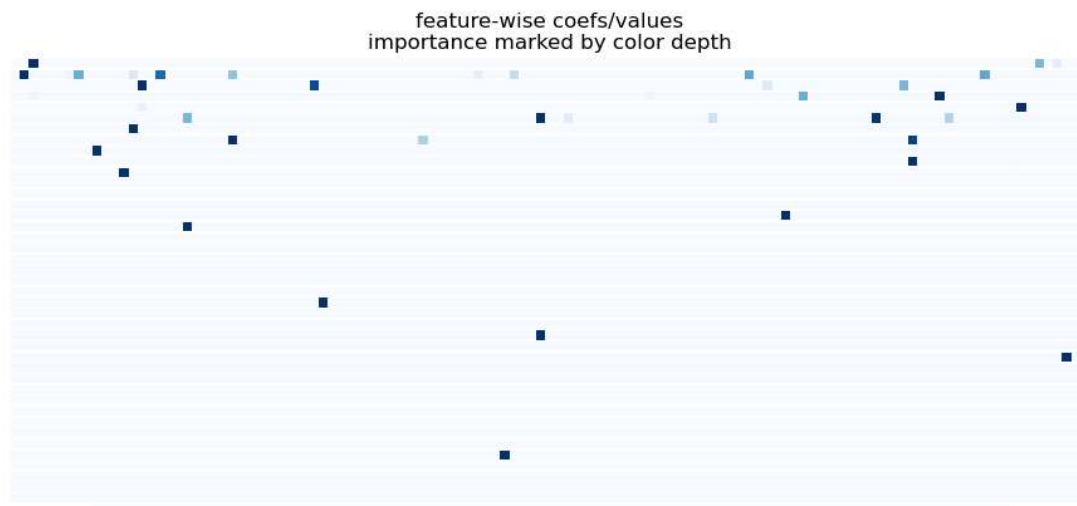
elastic net

LASSO系列基于L1-norm, 特征选择的稀疏性较强烈。ElasticNet(L1+L2) 更为"温和"。Elastic net is "a doubly regularized technique which encourages grouping effect i.e. either selection or omission of the correlated variable together and is particularly useful when the number of covariates (p) is much larger than the number of observations (n). "

R2 = 1.0

alpha = 0.000313 , L1 ratio = 0.5

Non-zero feature coefficients: 44



Important feature Number: 30

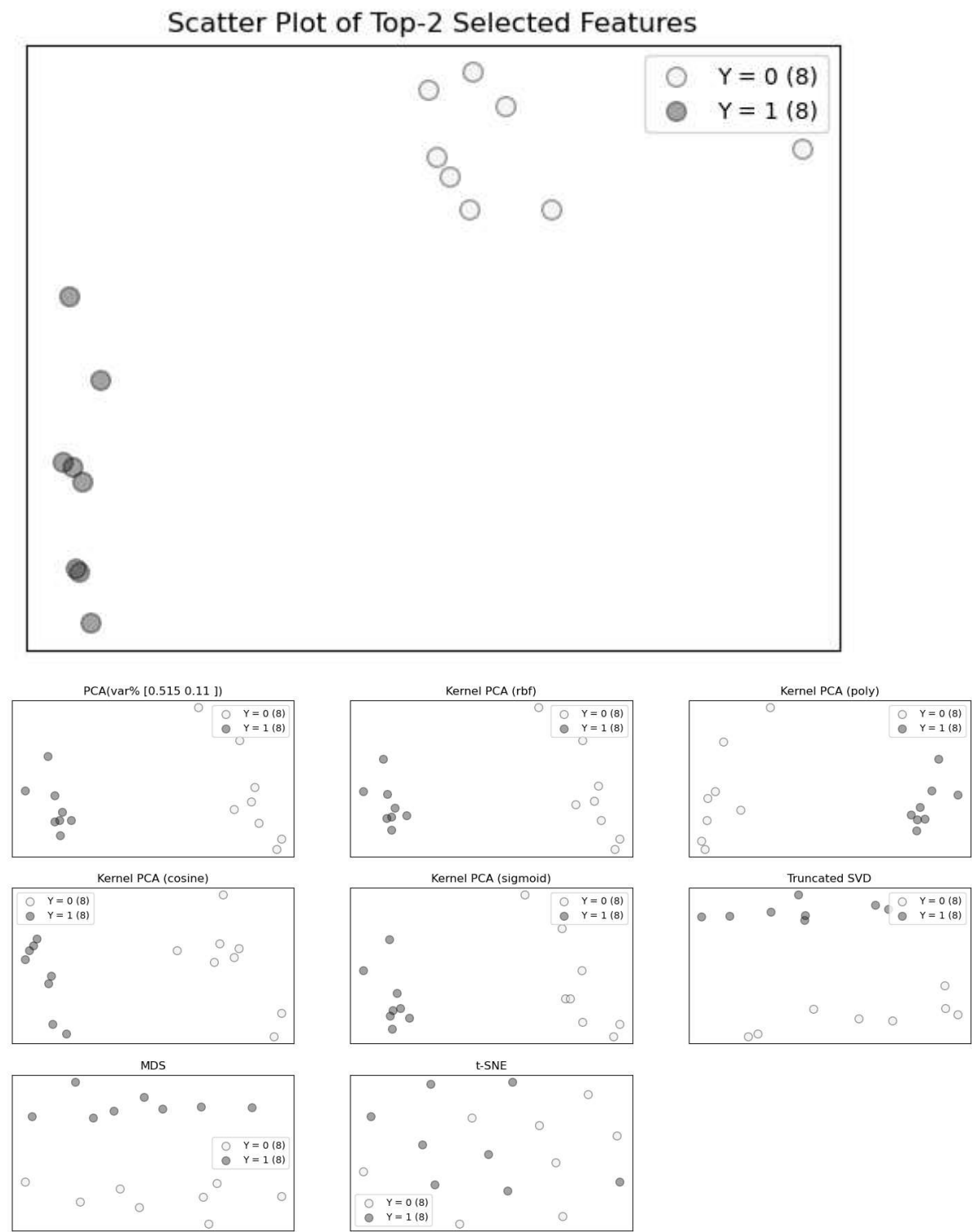
Important features indices: [ 969 2 1212 1819 462 113 121 864 939 254 43  
74 136 273 447  
201 227 658 127 695 144 338 885 619 175 115 703 591 733  
133 677]

Important features names: ['Pimelic acid' 'PC(14:1(9Z)/20:2(11Z,14Z))' 'Alanine'  
'Shikimic acid'

'4-(Nitrosoamino)-1-(3-pyridinyl)-1-butanone' 'ANETHOLE' 'Zolmitriptan'  
'15-methyl-heptadecanoic acid' '5-Hydroxylysine' 'm-Chlorobenzamide'  
'Piretanide glucuronide' 'Nicotyrine' 'N-Acetyl-D-glucosamine'  
'Germacrone-13-al' 'Oxypurinol' 'Talaromycin A' 'Isoquinoline'  
'beta-Alanyl-L-lysine' 'O-propanoyl-D-carnitine' '3-Vinylcatechol'  
'O-glutaryl carnitine' '12-Hydroxydodecanoic acid'  
'(±)-2-Methylthiazolidine' '(1x,2x)-Guaiacylglycerol 3-glucoside'  
'3-hydroxyisovaleryl carnitine' 'S-Acetyldihydrolipoamide-E' 'Polidocanol'  
'(E)-1-O-Cinnamoyl-beta-D-glucose' 'Cryptomeridiol'  
'1-(2-methoxy-13-methyl-tetradecanyl)-sn-glycero-3-phosphoserine']

Top-30 feature Importance: [0.38886748 0.20384667 0.13148465 0.13033233 0.1081301  
8 0.09356189

0.08165362 0.07653693 0.07084296 0.06532448 0.06437457 0.06371077  
0.058302 0.05336823 0.0432795 0.04301016 0.04151844 0.04077066  
0.04058233 0.03279985 0.03031252 0.02476956 0.01913504 0.01908849  
0.01828042 0.01261195 0.01211671 0.01170882 0.01093752 0.00805349]



Classification accuracy with the selected features (LogisticRegressionCV) = 1.0

---

adaptive lasso

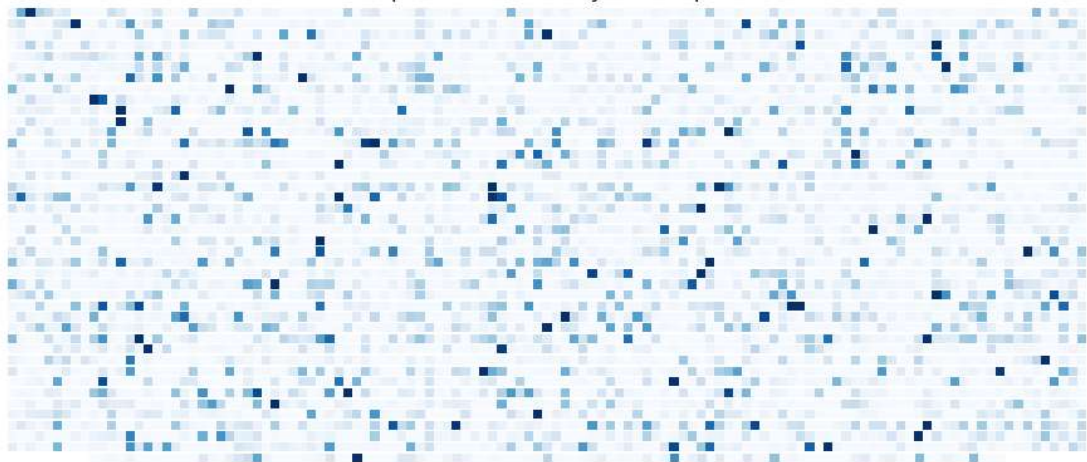
$$\alpha = 0.1$$

100% |

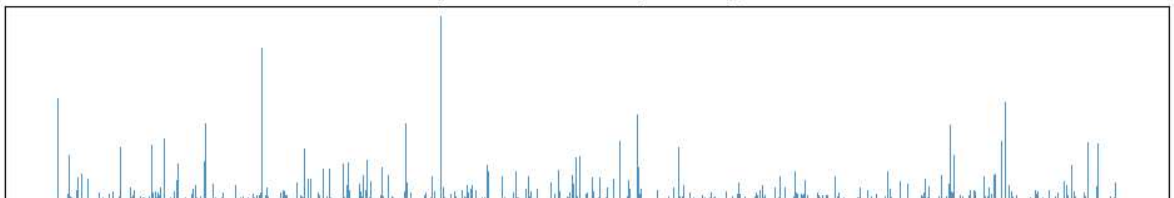
1/1 [00:00&lt;00:00, 5.74it/s]

```
Non-zero feature coefficients (eps = 0): 2556
```

feature-wise coefs/values  
importance marked by color depth



feature-wise coefs/values  
importance marked by bar height

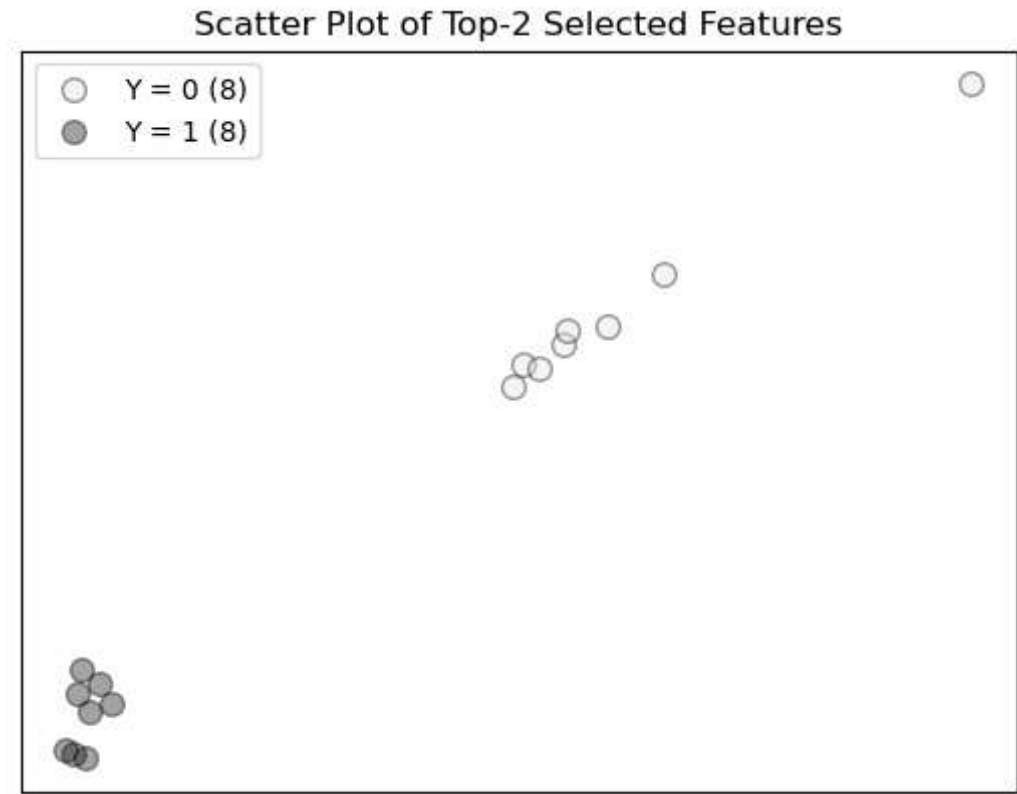


Important feature Number: 30

Important features indices: [ 969 1819 970 4349 4780 1212 3954 4949 2495 2 44 98 2752 2472 2236 1653 3853 703 4237 4693 2381 2076 1092 2554 127 2674 2093 4279 3735 4984 4703]

Important features names: ['Pimelic acid' 'Shikimic acid' 'Methylsuccinic acid' 'L-3-Hydroxykynurenine' 'Guanine' 'Alanine' '28-Homobrassinolide' 'Isoguanosine' 'N-phenethyltridecanamide' 'PC(14:1(9Z)/20:2(11Z,14Z))' 'Guanosine' 'Tris(butoxyethyl)phosphate' 'Adefovir' 'Alginic acid' '1-O-(2R-methoxy-hexadecyl)-sn-glycerol' 'Ethyl isopropyl disulfide' 'S-Acetyldihydrolipoamide-E' 'MG(0:0/22:4(7Z,10Z,13Z,16Z)/0:0)' '3-Hydroxyphenylacetic acid' 'beta-Geraniol' 'Kikkanol A' 'β-Alanine' 'D-Cysteine' 'beta-Alanyl-L-lysine' 'Serylleucine' 'S1P Lyase Fluorogenic Substrate' 'PS(0-16:0/19:1(9Z))' 'bismuth subgallate' '5-Acetoxy-tetradeca-2E,4E,6E-trienoic acid' 'Tetralin']

Top-30 feature Importance: [5.52585341e-15 5.52162356e-15 4.56450291e-15 3.54105426e-15 3.34701342e-15 3.27361357e-15 3.24419613e-15 3.21838678e-15 3.21210767e-15 3.04522072e-15 2.94742088e-15 2.56612146e-15 2.49929879e-15 2.48020937e-15 2.29764981e-15 2.29712586e-15 2.28834825e-15 2.28166714e-15 2.26321067e-15 2.20693532e-15 2.19689906e-15 2.17774781e-15 2.16940262e-15 2.16470342e-15 2.16363781e-15 2.15083213e-15 2.05502989e-15 2.04272045e-15 2.04073389e-15 2.03063852e-15]

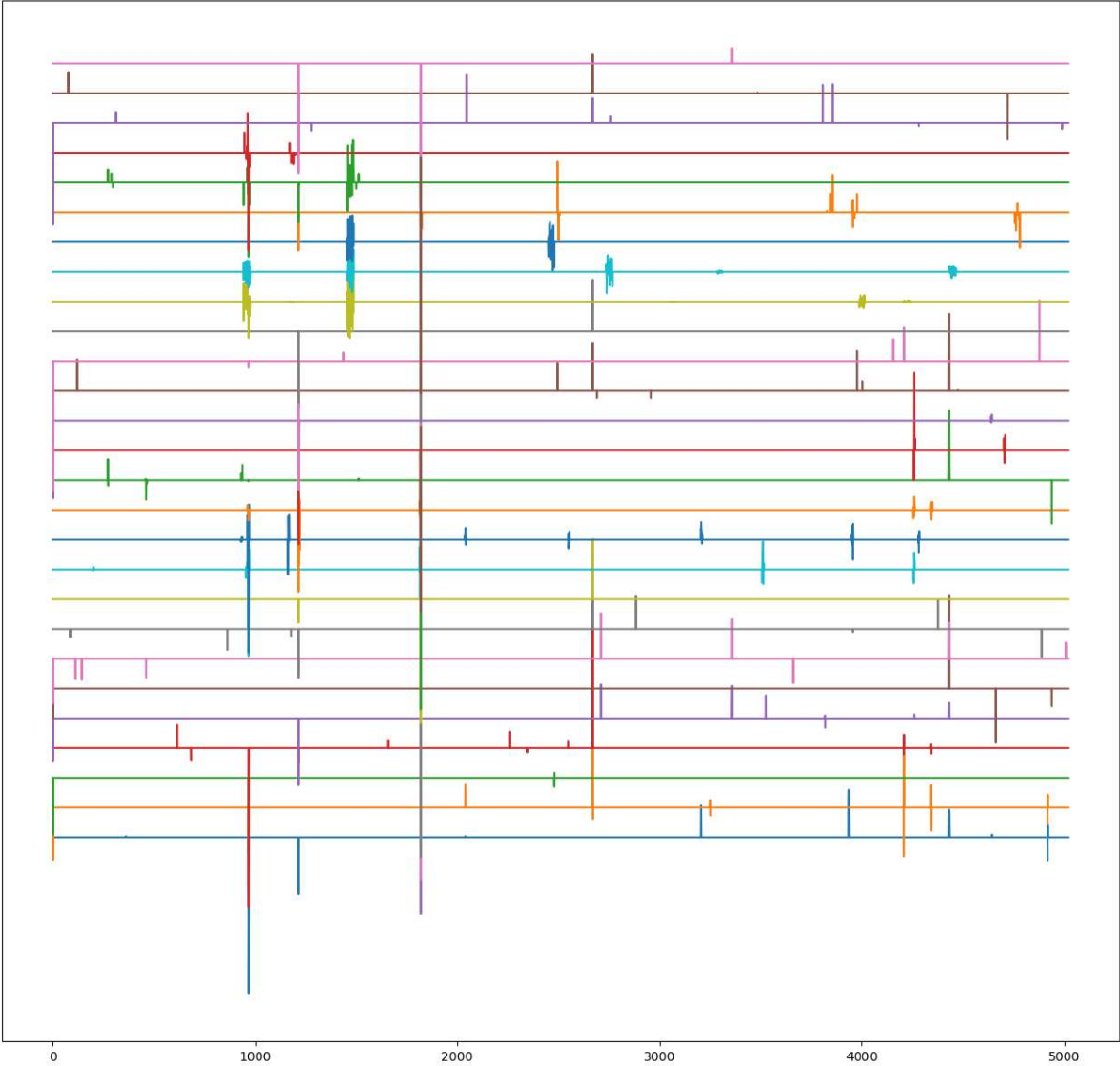




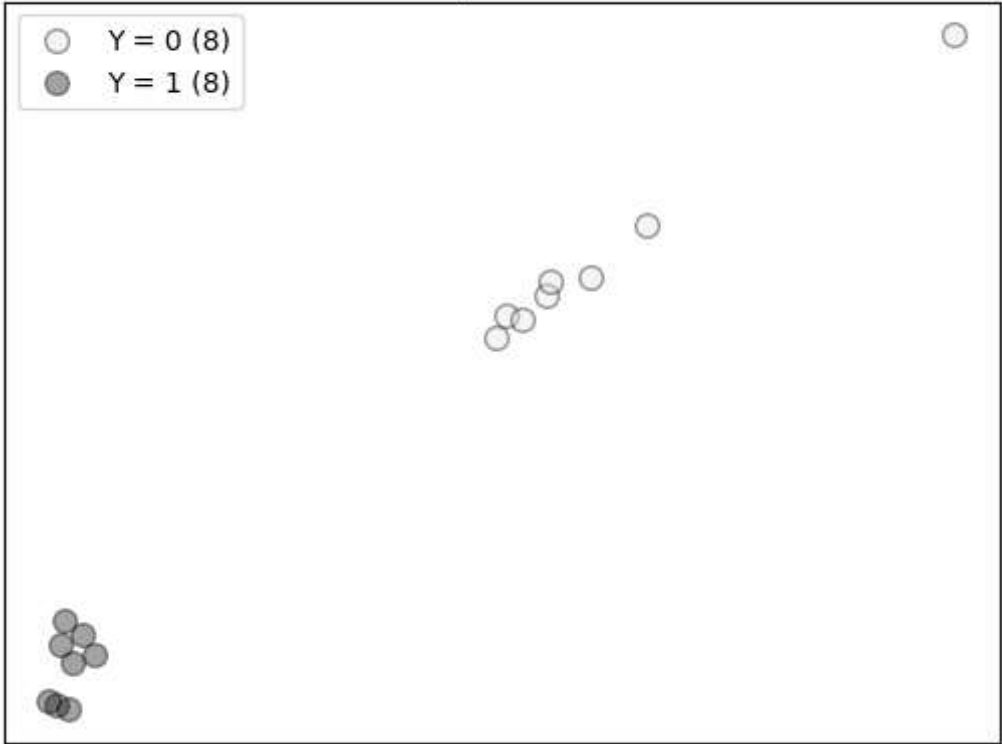


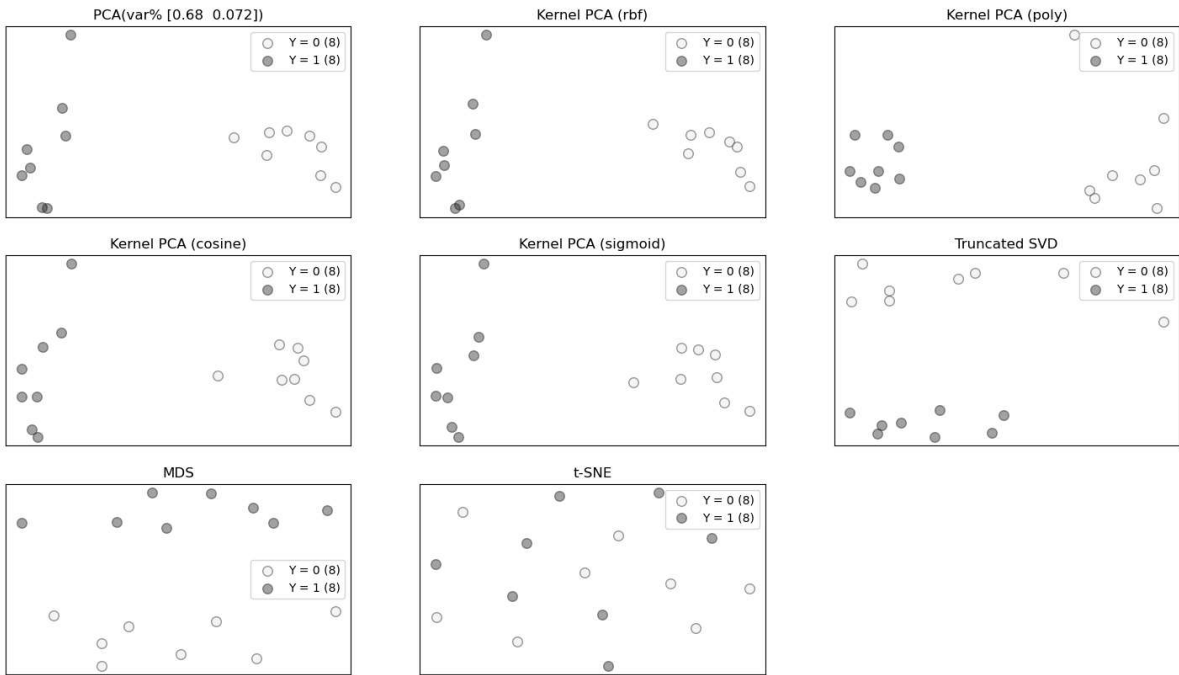
The group lasso regulariser is a well known method to achieve structured sparsity in machine learning and statistics. An extension of the group lasso regulariser is the sparse group lasso regulariser [2], which imposes both group-wise sparsity and coefficient-wise sparsity. This is done by combining the group lasso penalty with the traditional lasso penalty. Reference: [/py/machine learning/source/19. Kernel/Group Lasso.ipynb](#) This method implements the sparse group lasso, which is a linear combination between lasso and group lasso, so it provides solutions that are both between and within group sparse.

```
top-30 common features and their frequencies: [(969, 25), (1819, 21), (1212, 20),  
(970, 15), (2, 15), (4431, 14), (2669, 14), (1, 12), (3356, 10), (3853, 10), (148  
0, 10), (2479, 9), (2495, 9), (4780, 9), (4938, 8), (1459, 8), (4342, 7), (4720,  
7), (1510, 7), (1487, 7), (4279, 6), (4256, 6), (3953, 6), (4257, 6), (4498, 6),  
(3735, 6), (1172, 6), (4949, 6), (4374, 6), (965, 6)]
```



Scatter Plot of Top-2 Selected Features



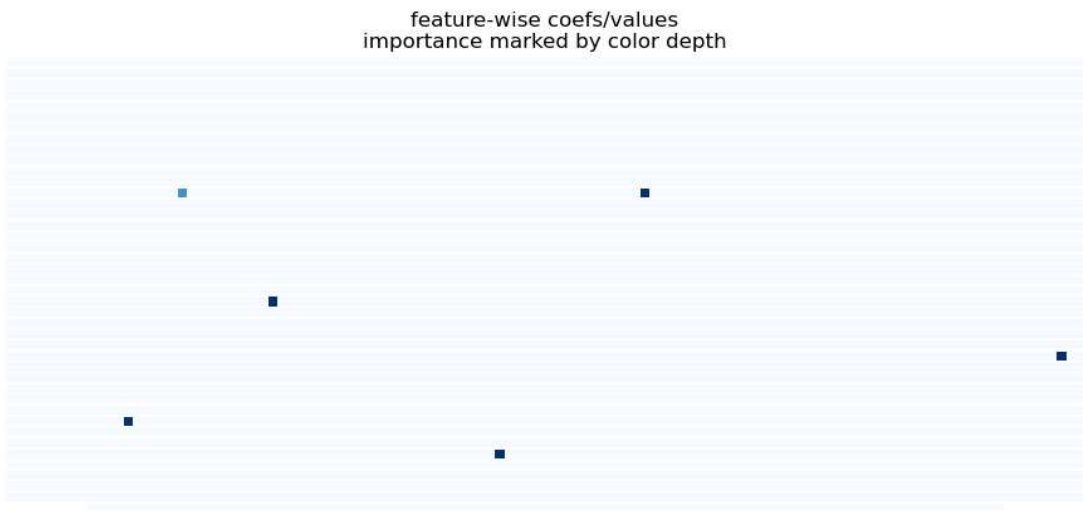


Classification accuracy with the selected features (LogisticRegressionCV) = 1.0

## adaptive elastic net

" The adaptive elastic-net combines the strengths of the quadratic regularization and the adaptively weighted lasso shrinkage. Under weak regularity conditions, we establish the oracle property of the adaptive elastic-net. We show by simulations that the adaptive elastic-net deals with the collinearity problem better than the other oracle-like methods, thus enjoying much improved finite sample performance."

Be patient. Your data is high-dimensional. It will take long time.  
R2 = -0.194  
Non-zero feature coefficients: 6

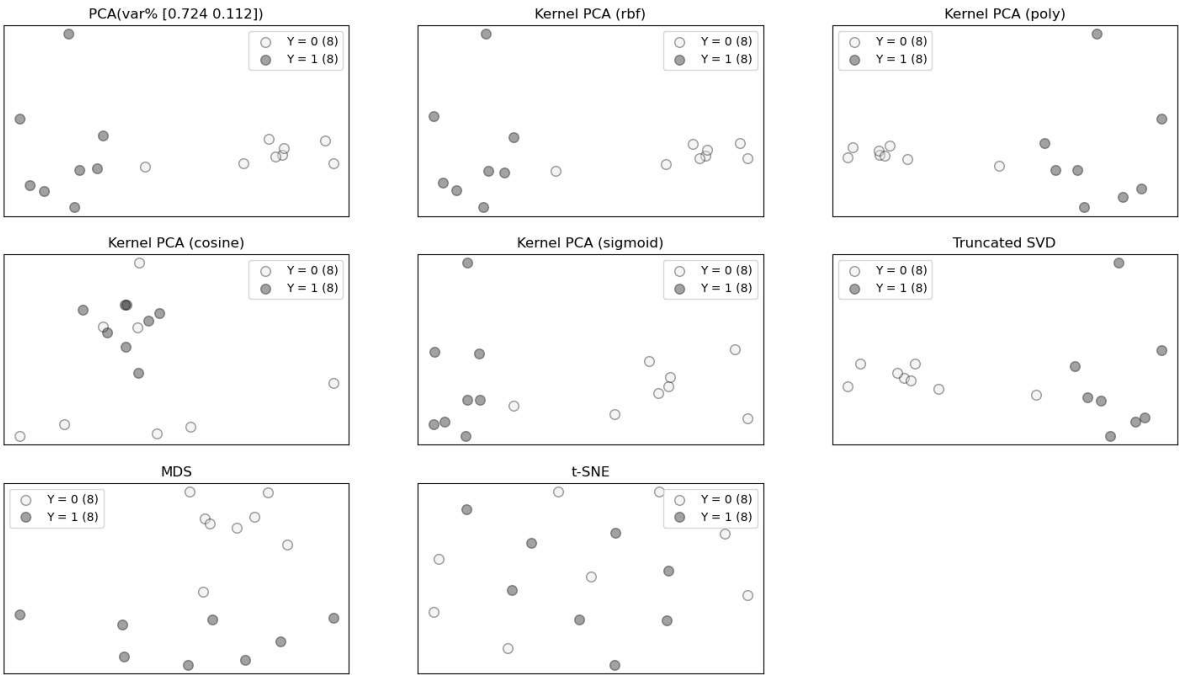
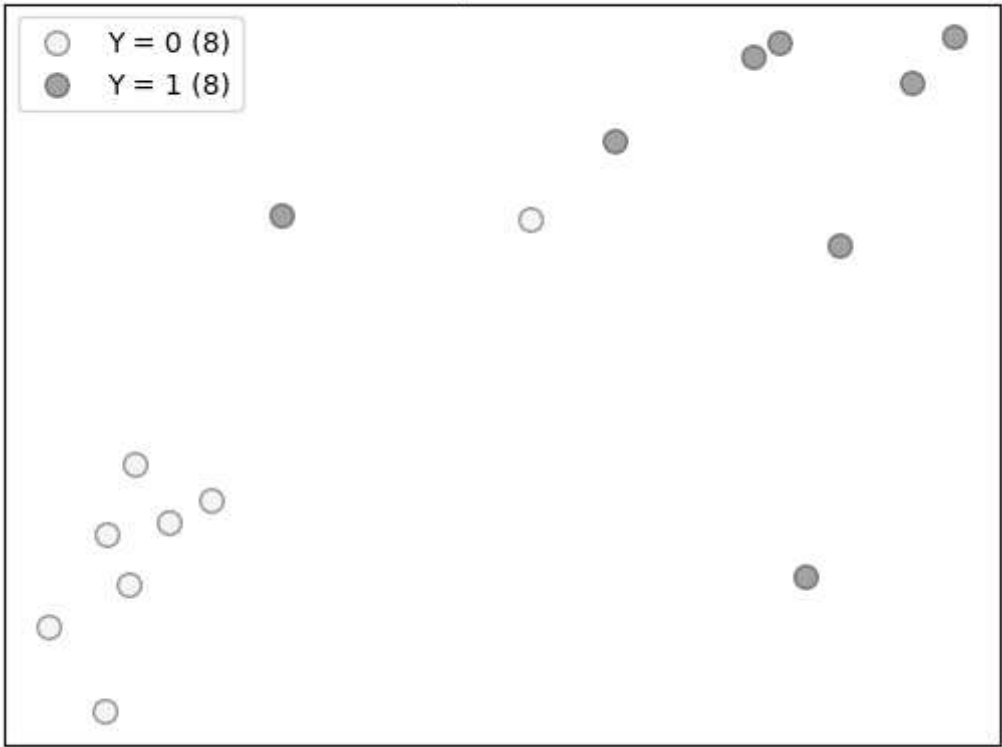


feature-wise coefs/values  
importance marked by bar height



Important feature Number: 6  
Important features indices: [2669 3356 1510 1459 4374 3973]  
Important features names: ['3-Indoxyl phosphate' 'Cromoglicic acid' 'G? 6983'  
'2-Phenylpropionaldehyde dimethyl acetal' 'Piretanide glucuronide'  
'L-902,688']  
Top-6 feature Importance: [0.15772273 0.14534578 0.12601856 0.07804902 0.0395402  
0.03280003]

Scatter Plot of Top-2 Selected Features



Classification accuracy with the selected features (LogisticRegressionCV) = 0.938

In [ ]:

In [ ]: