Classifying diseased trees

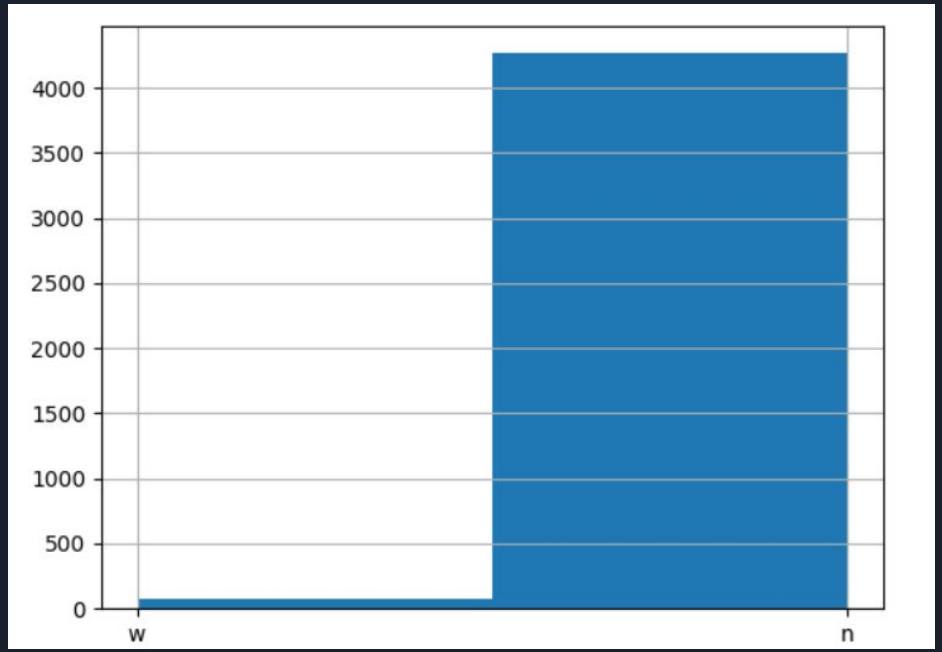
UCI Machine Learning Repository - WILTData Set
MLMT Exam, Med 7 2020-2021

Info about the data set

Wilt data set: high resolution Remote Sensing data set (Quickbird). Small number of training samples of diseased trees and large number for other land cover. Testing data set from stratified random sample of image. The data set consists of image segments, generated by segmenting the pan sharpened image.

Data description

- 4339 entries
- 74 diseased trees, 4265 land cover
- class: 'w' (diseased trees), 'n' (all other land cover)
- GLCM_Pan: GLCM mean texture (Pan band)
- Mean_G: Mean green value
- Mean_R: Mean red value
- Mean_NIR: Mean NIR value
- SD_Pan: Standard deviation (Pan band)
- 1 Status (0,1)
- Task: Classification



Tools/library used

- Python 3.7
  - Pandas
  - Scikit-learn
  - Numpy
  - Matplotlib

Data description

	GLCM_pan	Mean_Green	Mean_Red	Mean_NIR	SD_pan
count	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000
mean	126.831298	233.906908	117.292439	534.104683	24.924588
std	13.735836	60.757687	60.711159	154.495500	11.008303
min	0.000000	164.625000	59.142857	86.500000	0.000000
25%	118.589080	206.000000	91.975244	422.875000	18.009143
50%	127.479167	221.454545	101.727273	528.500000	23.612444
75%	135.043591	241.791304	116.866071	643.087037	29.899148
max	183.281250	955.714286	746.333333	1005.516129	156.508431

Data set info

Class distribution

```
class
n    4265
w      74
dtype: int64
```

Data shape

```
(4339, 6)
```

Variance

```
Variance of GLCM_pan is:  188.62971890416026
Variance of Mean_Green is:  3690.6457771477285
Variance of Mean_Red is:  3690.6457771477285
Variance of Mean_NIR is:  23863.3585170741
Variance of SD_pan is:  121.15480413605562
```

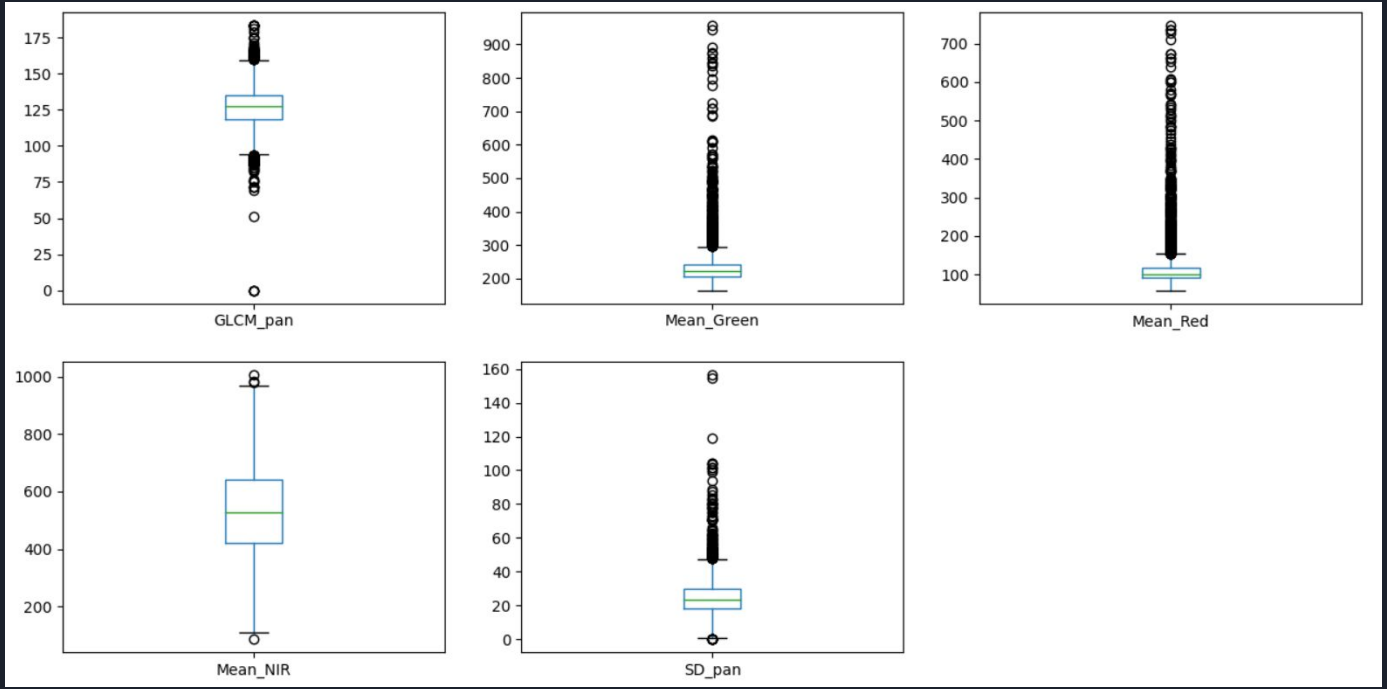
Dimensionality reduction PCA

	0	1	2	3	4
0	-119.485140	-2.428412	6.444780	9.617164	-16.320741
1	-181.834040	0.969881	2.537671	5.222736	-16.212408
2	-59.803322	-16.902237	-8.292051	13.915148	-17.486411
3	-261.535380	-22.151249	0.455989	4.462354	-15.542618
4	-5.858443	-28.976101	-9.621701	10.320577	-20.955928

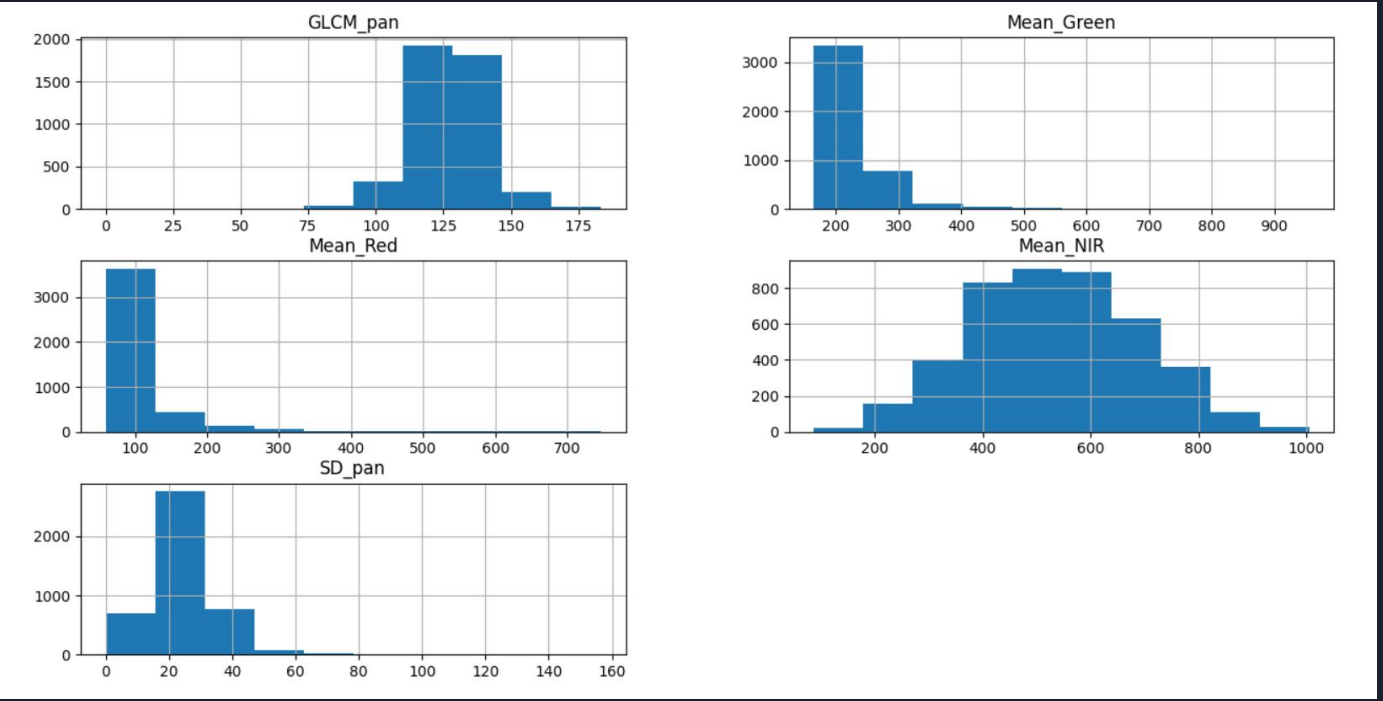
Data normalized

	GLCM_pan	Mean_Green	Mean_Red	Mean_NIR	SD_pan
0	0.656711	0.051669	0.087679	0.359168	0.132110
1	0.680591	0.048256	0.081768	0.291435	0.106749
2	0.734892	0.043814	0.083986	0.425844	0.143741
3	0.698087	0.017373	0.048350	0.208890	0.095697
4	0.738927	0.040925	0.077923	0.485794	0.112481
...	...	...	...	...	...
4334	0.608730	0.049253	0.049654	0.345190	0.169402
4335	0.684387	0.052800	0.049354	0.463289	0.213829
4336	0.720880	0.380133	0.343705	0.449122	0.247673
4337	0.679118	0.063871	0.057250	0.641355	0.182141
4338	0.682950	0.498582	0.446492	0.383858	0.098348

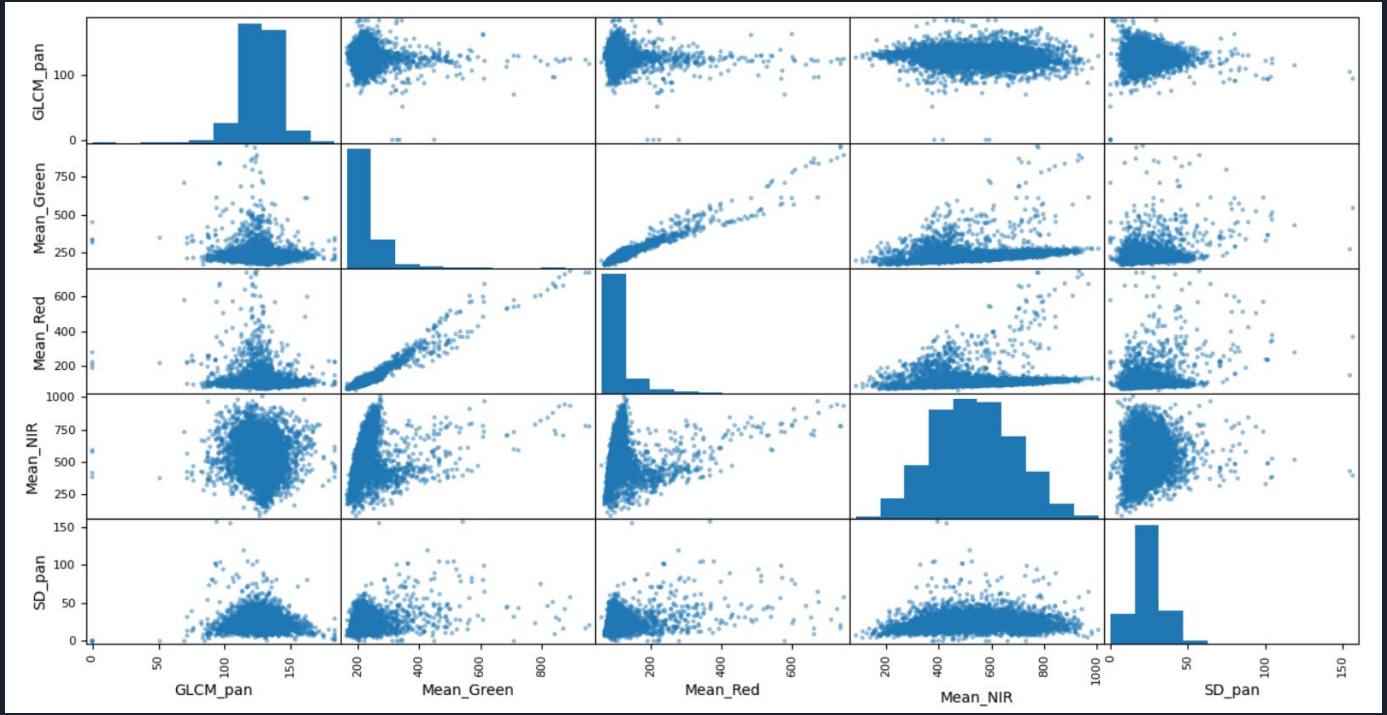
Univariate plots



Histogram plot each variable



Scatter Matrix Plot



Distance between clusters

```
[[  0.         311.71113755 428.46914489 165.62614863 460.66988544]
 [311.71113755   0.         446.83074026 147.44978217 149.76370023]
 [428.46914489 446.83074026   0.         423.99214992 539.41212303]
 [165.62614863 147.44978217 423.99214992   0.         295.4199153 ]
 [460.66988544 149.76370023 539.41212303 295.4199153    0.        ]]
```

Training set

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

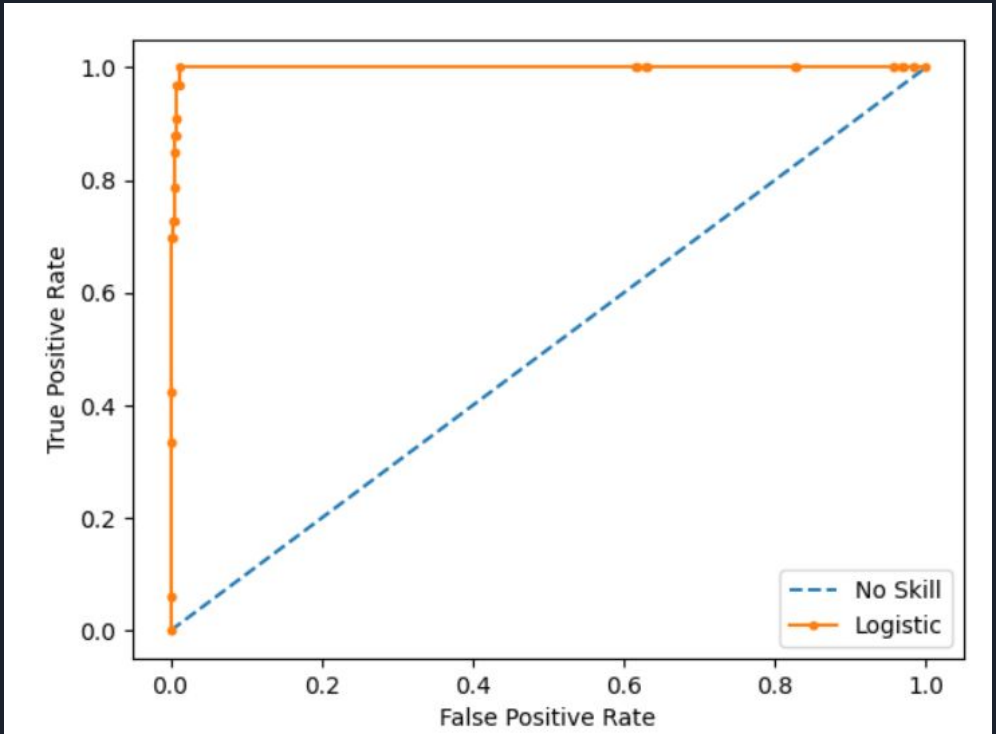
Logistic regression

Accuracy of Logistic Regression: 0.97926267281105990

	precision	recall	f1-score	support
n	0.98	1.00	0.99	851
w	0.33	0.06	0.10	17
accuracy			0.98	868
macro avg	0.66	0.53	0.54	868
weighted avg	0.97	0.98	0.97	868

Logistic regression ROC curve

AUC score =0.892



SVC

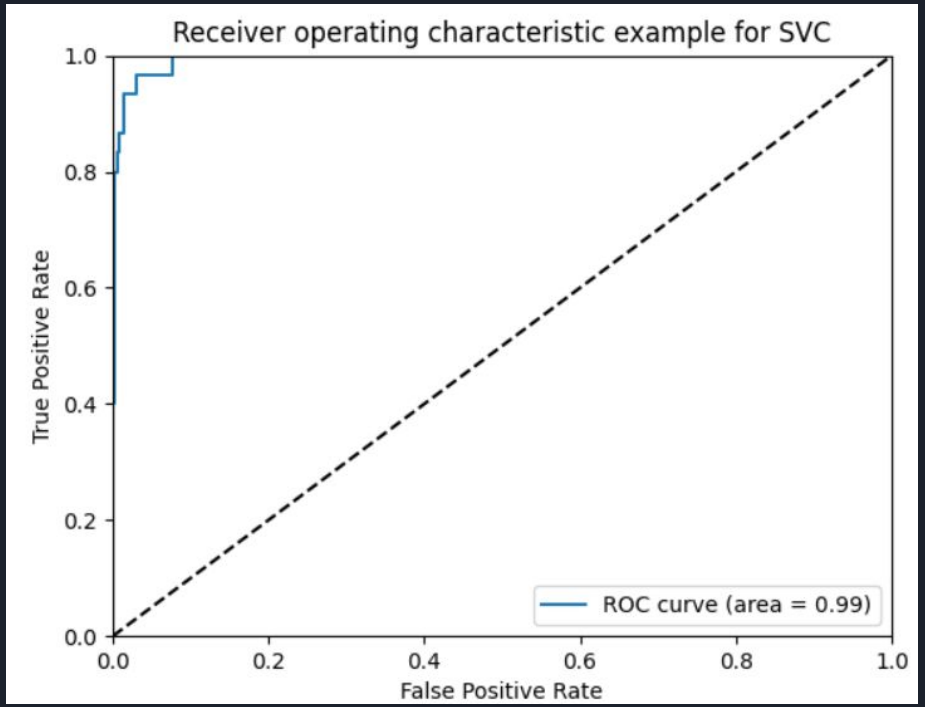
```
SVC Accuracy is: 0.9923195084485407
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      1285
           1       1.00      0.41      0.58        17

    accuracy                           0.99      1302
   macro avg       1.00      0.71      0.79      1302
weighted avg       0.99      0.99      0.99      1302
```

SVC ROC Curve

AUC score =0.912



Naive Bayes - Gaussian

```
naive_bayes = GaussianNB()
naive_bayes.fit(X_train, y_train)
y_predicted = naive_bayes.predict(X_test)
```

```
Number of mislabeled points out of a total 1302 points : 23
```

```
Accuracy of GNB classifier on training set: 0.98
Accuracy of GNB classifier on test set: 0.98
```

K fold cross vald linear regression

```
Fold:1, Train set: 3471, Test set:868
Fold:2, Train set: 3471, Test set:868
Fold:3, Train set: 3471, Test set:868
Fold:4, Train set: 3471, Test set:868
Fold:5, Train set: 3472, Test set:867
Scores for each fold: [-0.01497696 -0.01152074 -0.00691244 -0.00691244 -0.00576701]
rmse= 0.10
```