Lab03protocol

November 10, 2022

1 Answer the following questions

[ ]:
```
from sklearn.datasets import load_breast_cancer
import numpy as np
x,y = load_breast_cancer(return_X_y = True, as_frame = True)
x.head()
```

[ ]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	\
0	17.99	10.38	122.80	1001.0	0.11840	
1	20.57	17.77	132.90	1326.0	0.08474	
2	19.69	21.25	130.00	1203.0	0.10960	
3	11.42	20.38	77.58	386.1	0.14250	
4	20.29	14.34	135.10	1297.0	0.10030	

	mean compactness	mean concavity	mean concave points	mean symmetry	\
0	0.27760	0.3001	0.14710	0.2419	
1	0.07864	0.0869	0.07017	0.1812	
2	0.15990	0.1974	0.12790	0.2069	
3	0.28390	0.2414	0.10520	0.2597	
4	0.13280	0.1980	0.10430	0.1809	

	mean fractal dimension	…	worst radius	worst texture	worst perimeter	\
0	0.07871	…	25.38	17.33	184.60	
1	0.05667	…	24.99	23.41	158.80	
2	0.05999	…	23.57	25.53	152.50	
3	0.09744	…	14.91	26.50	98.87	
4	0.05883	…	22.54	16.67	152.20	

	worst area	worst smoothness	worst compactness	worst concavity	\
0	2019.0	0.1622	0.6656	0.7119	
1	1956.0	0.1238	0.1866	0.2416	
2	1709.0	0.1444	0.4245	0.4504	
3	567.7	0.2098	0.8663	0.6869	
4	1575.0	0.1374	0.2050	0.4000	

	worst concave points	worst symmetry	worst fractal dimension
0	0.2654	0.4601	0.11890

```
1                0.1860          0.2750               0.08902
2                0.2430          0.3613               0.08758
3                0.2575          0.6638               0.17300
4                0.1625          0.2364               0.07678

[5 rows x 30 columns]
```

1.0.1 How was the data optained?

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

1.0.2 How many classes are there?

[ ]:
```
y.unique()
```

[ ]:
```
array([0, 1])
```

1.0.3 What does each row represent?

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

1.0.4 How many data points are there?

[ ]:
```
x.shape[0]
```

[ ]:
```
569
```

1.0.5 How many features?

[ ]:
```
x.shape[1]
```

[ ]:
```
30
```

1.0.6 Which kind of features are there?

The mean, standard error, and "worst" or largest (mean of the three worst/largest values) of these features were computed for each image, resulting in 30 features. For instance, field 0 is Mean Radius, field 10 is Radius SE, field 20 is Worst Radius.

1.0.7 Which feature(s) have/has the highest absolut values?

[ ]:
```
x.describe().iloc[1,].sort_values(ascending=False)[:5]
```

[ ]:
```
worst area         880.583128
mean area          654.889104
worst perimeter    107.261213
mean perimeter      91.969033
area error          40.337079
Name: mean, dtype: float64
```

The highest values are at the feature worst area, mean area, worst perimeter

1.0.8 Just from the information present, do you expect high or low correlation?

[ ]:
```
x.corr()
```

[ ]:

	mean radius	mean texture	mean perimeter	mean area	\
mean radius	1.000000	0.323782	0.997855	0.987357	
mean texture	0.323782	1.000000	0.329533	0.321086	
mean perimeter	0.997855	0.329533	1.000000	0.986507	
mean area	0.987357	0.321086	0.986507	1.000000	
mean smoothness	0.170581	-0.023389	0.207278	0.177028	
mean compactness	0.506124	0.236702	0.556936	0.498502	
mean concavity	0.676764	0.302418	0.716136	0.685983	
mean concave points	0.822529	0.293464	0.850977	0.823269	
mean symmetry	0.147741	0.071401	0.183027	0.151293	
mean fractal dimension	-0.311631	-0.076437	-0.261477	-0.283110	
radius error	0.679090	0.275869	0.691765	0.732562	
texture error	-0.097317	0.386358	-0.086761	-0.066280	
perimeter error	0.674172	0.281673	0.693135	0.726628	
area error	0.735864	0.259845	0.744983	0.800086	
smoothness error	-0.222600	0.006614	-0.202694	-0.166777	
compactness error	0.206000	0.191975	0.250744	0.212583	
concavity error	0.194204	0.143293	0.228082	0.207660	
concave points error	0.376169	0.163851	0.407217	0.372320	
symmetry error	-0.104321	0.009127	-0.081629	-0.072497	
fractal dimension error	-0.042641	0.054458	-0.005523	-0.019887	
worst radius	0.969539	0.352573	0.969476	0.962746	
worst texture	0.297008	0.912045	0.303038	0.287489	
worst perimeter	0.965137	0.358040	0.970387	0.959120	
worst area	0.941082	0.343546	0.941550	0.959213	
worst smoothness	0.119616	0.077503	0.150549	0.123523	
worst compactness	0.413463	0.277830	0.455774	0.390410	
worst concavity	0.526911	0.301025	0.563879	0.512606	
worst concave points	0.744214	0.295316	0.771241	0.722017	
worst symmetry	0.163953	0.105008	0.189115	0.143570	

```
worst fractal dimension     0.007066      0.119205        0.051019   0.003738

                        mean smoothness  mean compactness  mean concavity  \
mean radius                    0.170581          0.506124        0.676764
mean texture                  -0.023389          0.236702        0.302418
mean perimeter                 0.207278          0.556936        0.716136
mean area                      0.177028          0.498502        0.685983
mean smoothness                1.000000          0.659123        0.521984
mean compactness               0.659123          1.000000        0.883121
mean concavity                 0.521984          0.883121        1.000000
mean concave points            0.553695          0.831135        0.921391
mean symmetry                  0.557775          0.602641        0.500667
mean fractal dimension         0.584792          0.565369        0.336783
radius error                   0.301467          0.497473        0.631925
texture error                  0.068406          0.046205        0.076218
perimeter error                0.296092          0.548905        0.660391
area error                     0.246552          0.455653        0.617427
smoothness error               0.332375          0.135299        0.098564
compactness error              0.318943          0.738722        0.670279
concavity error                0.248396          0.570517        0.691270
concave points error           0.380676          0.642262        0.683260
symmetry error                 0.200774          0.229977        0.178009
fractal dimension error        0.283607          0.507318        0.449301
worst radius                   0.213120          0.535315        0.688236
worst texture                  0.036072          0.248133        0.299879
worst perimeter                0.238853          0.590210        0.729565
worst area                     0.206718          0.509604        0.675987
worst smoothness               0.805324          0.565541        0.448822
worst compactness              0.472468          0.865809        0.754968
worst concavity                0.434926          0.816275        0.884103
worst concave points           0.503053          0.815573        0.861323
worst symmetry                 0.394309          0.510223        0.409464
worst fractal dimension        0.499316          0.687382        0.514930

                        mean concave points  mean symmetry  \
mean radius                        0.822529       0.147741
mean texture                       0.293464       0.071401
mean perimeter                     0.850977       0.183027
mean area                          0.823269       0.151293
mean smoothness                    0.553695       0.557775
mean compactness                   0.831135       0.602641
mean concavity                     0.921391       0.500667
mean concave points                1.000000       0.462497
mean symmetry                      0.462497       1.000000
mean fractal dimension             0.166917       0.479921
radius error                       0.698050       0.303379
texture error                      0.021480       0.128053
```

```
perimeter error                  0.710650       0.313893
area error                       0.690299       0.223970
smoothness error                 0.027653       0.187321
compactness error                0.490424       0.421659
concavity error                  0.439167       0.342627
concave points error             0.615634       0.393298
symmetry error                   0.095351       0.449137
fractal dimension error          0.257584       0.331786
worst radius                     0.830318       0.185728
worst texture                    0.292752       0.090651
worst perimeter                  0.855923       0.219169
worst area                       0.809630       0.177193
worst smoothness                 0.452753       0.426675
worst compactness                0.667454       0.473200
worst concavity                  0.752399       0.433721
worst concave points             0.910155       0.430297
worst symmetry                   0.375744       0.699826
worst fractal dimension          0.368661       0.438413

                         mean fractal dimension  …  worst radius  \
mean radius                           -0.311631  …      0.969539
mean texture                          -0.076437  …      0.352573
mean perimeter                        -0.261477  …      0.969476
mean area                             -0.283110  …      0.962746
mean smoothness                        0.584792  …      0.213120
mean compactness                       0.565369  …      0.535315
mean concavity                         0.336783  …      0.688236
mean concave points                    0.166917  …      0.830318
mean symmetry                          0.479921  …      0.185728
mean fractal dimension                 1.000000  …     -0.253691
radius error                           0.000111  …      0.715065
texture error                          0.164174  …     -0.111690
perimeter error                        0.039830  …      0.697201
area error                            -0.090170  …      0.757373
smoothness error                       0.401964  …     -0.230691
compactness error                      0.559837  …      0.204607
concavity error                        0.446630  …      0.186904
concave points error                   0.341198  …      0.358127
symmetry error                         0.345007  …     -0.128121
fractal dimension error                0.688132  …     -0.037488
worst radius                          -0.253691  …      1.000000
worst texture                         -0.051269  …      0.359921
worst perimeter                       -0.205151  …      0.993708
worst area                            -0.231854  …      0.984015
worst smoothness                       0.504942  …      0.216574
worst compactness                      0.458798  …      0.475820
worst concavity                        0.346234  …      0.573975
```

```
worst concave points                 0.175325  …      0.787424
worst symmetry                       0.334019  …      0.243529
worst fractal dimension              0.767297  …      0.093492

                         worst texture  worst perimeter  worst area  \
mean radius                   0.297008         0.965137    0.941082
mean texture                  0.912045         0.358040    0.343546
mean perimeter                0.303038         0.970387    0.941550
mean area                     0.287489         0.959120    0.959213
mean smoothness               0.036072         0.238853    0.206718
mean compactness              0.248133         0.590210    0.509604
mean concavity                0.299879         0.729565    0.675987
mean concave points           0.292752         0.855923    0.809630
mean symmetry                 0.090651         0.219169    0.177193
mean fractal dimension       -0.051269        -0.205151   -0.231854
radius error                  0.194799         0.719684    0.751548
texture error                 0.409003        -0.102242   -0.083195
perimeter error               0.200371         0.721031    0.730713
area error                    0.196497         0.761213    0.811408
smoothness error             -0.074743        -0.217304   -0.182195
compactness error             0.143003         0.260516    0.199371
concavity error               0.100241         0.226680    0.188353
concave points error          0.086741         0.394999    0.342271
symmetry error               -0.077473        -0.103753   -0.110343
fractal dimension error      -0.003195        -0.001000   -0.022736
worst radius                  0.359921         0.993708    0.984015
worst texture                 1.000000         0.365098    0.345842
worst perimeter               0.365098         1.000000    0.977578
worst area                    0.345842         0.977578    1.000000
worst smoothness              0.225429         0.236775    0.209145
worst compactness             0.360832         0.529408    0.438296
worst concavity               0.368366         0.618344    0.543331
worst concave points          0.359755         0.816322    0.747419
worst symmetry                0.233027         0.269493    0.209146
worst fractal dimension       0.219122         0.138957    0.079647

                         worst smoothness  worst compactness  worst concavity  \
mean radius                      0.119616           0.413463         0.526911
mean texture                     0.077503           0.277830         0.301025
mean perimeter                   0.150549           0.455774         0.563879
mean area                        0.123523           0.390410         0.512606
mean smoothness                  0.805324           0.472468         0.434926
mean compactness                 0.565541           0.865809         0.816275
mean concavity                   0.448822           0.754968         0.884103
mean concave points              0.452753           0.667454         0.752399
mean symmetry                    0.426675           0.473200         0.433721
mean fractal dimension           0.504942           0.458798         0.346234
```

radius error	0.141919	0.287103	0.380585
texture error	-0.073658	-0.092439	-0.068956
perimeter error	0.130054	0.341919	0.418899
area error	0.125389	0.283257	0.385100
smoothness error	0.314457	-0.055558	-0.058298
compactness error	0.227394	0.678780	0.639147
concavity error	0.168481	0.484858	0.662564
concave points error	0.215351	0.452888	0.549592
symmetry error	-0.012662	0.060255	0.037119
fractal dimension error	0.170568	0.390159	0.379975
worst radius	0.216574	0.475820	0.573975
worst texture	0.225429	0.360832	0.368366
worst perimeter	0.236775	0.529408	0.618344
worst area	0.209145	0.438296	0.543331
worst smoothness	1.000000	0.568187	0.518523
worst compactness	0.568187	1.000000	0.892261
worst concavity	0.518523	0.892261	1.000000
worst concave points	0.547691	0.801080	0.855434
worst symmetry	0.493838	0.614441	0.532520
worst fractal dimension	0.617624	0.810455	0.686511

	worst concave points	worst symmetry	\
mean radius	0.744214	0.163953	
mean texture	0.295316	0.105008	
mean perimeter	0.771241	0.189115	
mean area	0.722017	0.143570	
mean smoothness	0.503053	0.394309	
mean compactness	0.815573	0.510223	
mean concavity	0.861323	0.409464	
mean concave points	0.910155	0.375744	
mean symmetry	0.430297	0.699826	
mean fractal dimension	0.175325	0.334019	
radius error	0.531062	0.094543	
texture error	-0.119638	-0.128215	
perimeter error	0.554897	0.109930	
area error	0.538166	0.074126	
smoothness error	-0.102007	-0.107342	
compactness error	0.483208	0.277878	
concavity error	0.440472	0.197788	
concave points error	0.602450	0.143116	
symmetry error	-0.030413	0.389402	
fractal dimension error	0.215204	0.111094	
worst radius	0.787424	0.243529	
worst texture	0.359755	0.233027	
worst perimeter	0.816322	0.269493	
worst area	0.747419	0.209146	
worst smoothness	0.547691	0.493838	

```
worst compactness                   0.801080        0.614441
worst concavity                     0.855434        0.532520
worst concave points                1.000000        0.502528
worst symmetry                      0.502528        1.000000
worst fractal dimension             0.511114        0.537848
```

```
                        worst fractal dimension
mean radius                            0.007066
mean texture                           0.119205
mean perimeter                         0.051019
mean area                              0.003738
mean smoothness                        0.499316
mean compactness                       0.687382
mean concavity                         0.514930
mean concave points                    0.368661
mean symmetry                          0.438413
mean fractal dimension                 0.767297
radius error                           0.049559
texture error                         -0.045655
perimeter error                        0.085433
area error                             0.017539
smoothness error                       0.101480
compactness error                      0.590973
concavity error                        0.439329
concave points error                   0.310655
symmetry error                         0.078079
fractal dimension error                0.591328
worst radius                           0.093492
worst texture                          0.219122
worst perimeter                        0.138957
worst area                             0.079647
worst smoothness                       0.617624
worst compactness                      0.810455
worst concavity                        0.686511
worst concave points                   0.511114
worst symmetry                         0.537848
worst fractal dimension                1.000000

[30 rows x 30 columns]
```

Yes

2 Classification

1. Split data into a train and test set using a test set size of 15%.

2. Scale the data.

3. Train a KNeighborsClassifier and report the accuracy score on the test set.

Use nearest neighbors.

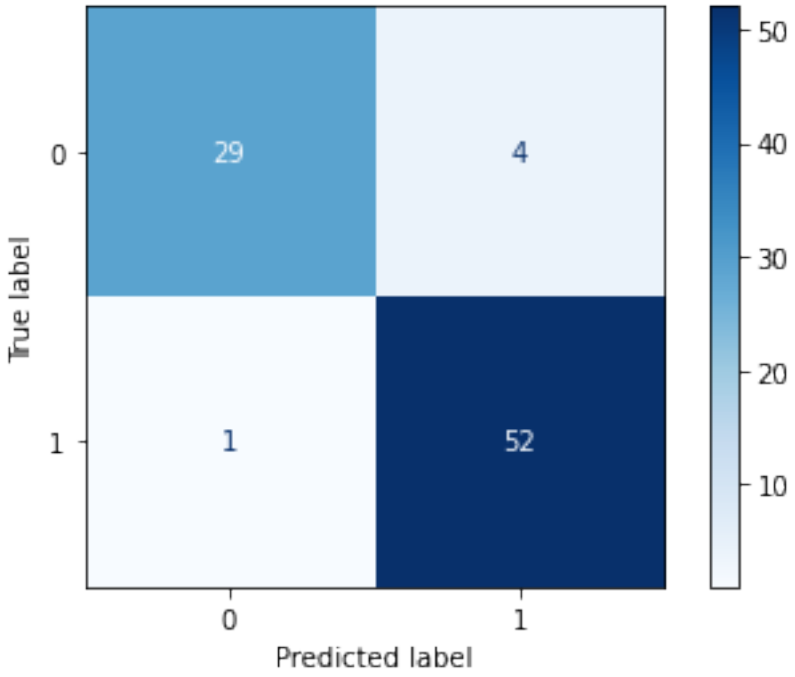
[ ]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
```

[ ]:
```
x_train, x_test, y_train, y_test = train_test_split(x ,y, train_size=0.85)

scaler = StandardScaler(copy=True)
xTrain_scaled = scaler.fit_transform(x_train, y_train)
minDis = KNeighborsClassifier(n_neighbors=7)
minDis.fit(xTrain_scaled, y_train)
xTest_scaled = scaler.transform(x_test)

minDis.score(xTest_scaled, y_test)
```

[ ]:
```
0.9418604651162791
```

3 Confusion Matrix

[ ]:
```
import matplotlib.pyplot as plt
from sklearn.metrics import ConfusionMatrixDisplay
ConfusionMatrixDisplay.from_estimator(minDis, xTest_scaled, y_test,␣
 ↪cmap='Blues')
plt.show()
```



[ ]:
```
manual_calc_score = (29 + 52)/(29+4+1+52)
manual_calc_score
```

[ ]: 0.9418604651162791

3.0.1 Benign data points are represented by which class number?

1

3.0.2 Malignant data points are represented by which class number?

0

3.0.3 How many data points are correctly classified as benign?

from the plot (1,1) –> 52

3.0.4 How many data points are correctly classified as malignant?

from the plot (0,0) –> 29

3.0.5 How many data points are classified as malignant, although being benign?

from the plot (0,1) –> 4

4 Performance Measures

Given the output from the confusion matrix, compute precision, recall and F1 score by hand for both labels.

[ ]:
```
TP = 29
TN = 4
FP = 1
FN = 52
precision = TP / (TP + FP)
recall = TP / (TP + TN)
F1_score = 2 * (precision * recall)/(precision + recall)
print("precision=", precision)
print("recall=", recall)
print("F1-score=", F1_score)
```

```
precision= 0.9666666666666667
recall= 0.8787878787878788
F1-score= 0.9206349206349207
```

Only then, use the methods implemented in scikit-learn for verification. Again, make sure to compute the values for both classes.

5 Classification Report

This is not the same as the calulated values by hand because the methods do interpred TP, TN, FP, FN not the same

The classification report provides a very good summary of all scores. By comparing to your own computed

values, make sure to be able to read and interpret the report.

[ ]:
```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred, digits=4))
```