4.0.KNN

May 15, 2023

[112]:
```
# KNN Classification
import numpy as np
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier
%matplotlib inline
from sklearn.model_selection import train_test_split
from scipy.stats import zscore
import seaborn as sns
import matplotlib.pyplot as plt
```

[113]:
```
data=pd.read_csv('wdbc.data',header=None)
headers=['id','diagnosis','mean_radius','mean_texture','mean_perimeter','mean_area','mean_smoo
 ↪points','mean_symmetry','mean_fractal␣
 ↪dimension','SE_radius','SE_texture','SE_perimeter','SE_area','SE_smoothness','SE_compactnes
 ↪points','SE_symmetry','SE_fractal␣
 ↪dimension','worst_radius','worst_texture','worst_perimeter','worst_area','worst_smoothness'
 ↪points','worst_symmetry','worst_fractal dimension']
data.to_csv('labeledData.csv',header=headers,index=False)
data=pd.read_csv('labeledData.csv')
```

data.head()

[114]:
```
data.head()
```

[114]:

	id	diagnosis	mean_radius	mean_texture	mean_perimeter	mean_area	\
0	842302	M	17.99	10.38	122.80	1001.0	\
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	

	mean_smoothness	mean_compactness	mean_concavity	mean_concave points	
0	0.11840	0.27760	0.3001	0.14710	\
1	0.08474	0.07864	0.0869	0.07017	
2	0.10960	0.15990	0.1974	0.12790	
3	0.14250	0.28390	0.2414	0.10520	

```
4          0.10030           0.13280          0.1980               0.10430

   …  worst_radius  worst_texture  worst_perimeter  worst_area
0  …         25.38          17.33           184.60      2019.0  \
1  …         24.99          23.41           158.80      1956.0
2  …         23.57          25.53           152.50      1709.0
3  …         14.91          26.50            98.87       567.7
4  …         22.54          16.67           152.20      1575.0

   worst_smoothness  worst_compactness  worst_concavity  worst_concave points
0            0.1622             0.6656           0.7119                0.2654  \
1            0.1238             0.1866           0.2416                0.1860
2            0.1444             0.4245           0.4504                0.2430
3            0.2098             0.8663           0.6869                0.2575
4            0.1374             0.2050           0.4000                0.1625

   worst_symmetry  worst_fractal dimension
0          0.4601                  0.11890
1          0.2750                  0.08902
2          0.3613                  0.08758
3          0.6638                  0.17300
4          0.2364                  0.07678

[5 rows x 32 columns]
```

[115]:
```
def diag(z):
    if z== 'M':
        return 1
    else:
        return 0

z=data['diagnosis'].apply(diag)
data.diagnosis=z
```

[117]:
```
df = pd.DataFrame(data=data)
df=df.drop('id',axis=1)

x=df.drop('diagnosis',axis=1)
y=df['diagnosis']
```

[118]:
```
x_scaled=x.apply(zscore)
x_scaled.describe()
```

[118]:
```
        mean_radius  mean_texture  mean_perimeter     mean_area
count  5.690000e+02  5.690000e+02    5.690000e+02  5.690000e+02  \
mean  -1.373633e-16  6.868164e-17   -1.248757e-16 -2.185325e-16
std    1.000880e+00  1.000880e+00    1.000880e+00  1.000880e+00
```

```
min   -2.029648e+00 -2.229249e+00   -1.984504e+00 -1.454443e+00
25%   -6.893853e-01 -7.259631e-01   -6.919555e-01 -6.671955e-01
50%   -2.150816e-01 -1.046362e-01   -2.359800e-01 -2.951869e-01
75%    4.693926e-01  5.841756e-01    4.996769e-01  3.635073e-01
max    3.971288e+00  4.651889e+00    3.976130e+00  5.250529e+00

       mean_smoothness  mean_compactness  mean_concavity  mean_concave points
count     5.690000e+02      5.690000e+02    5.690000e+02         5.690000e+02  \
mean     -8.366672e-16      1.873136e-16    4.995028e-17        -4.995028e-17
std       1.000880e+00      1.000880e+00    1.000880e+00         1.000880e+00
min      -3.112085e+00     -1.610136e+00   -1.114873e+00        -1.261820e+00
25%      -7.109628e-01     -7.470860e-01   -7.437479e-01        -7.379438e-01
50%      -3.489108e-02     -2.219405e-01   -3.422399e-01        -3.977212e-01
75%       6.361990e-01      4.938569e-01    5.260619e-01         6.469351e-01
max       4.770911e+00      4.568425e+00    4.243589e+00         3.927930e+00

       mean_symmetry  mean_fractal dimension  …  worst_radius
count   5.690000e+02            5.690000e+02  …  5.690000e+02  \
mean    1.748260e-16            4.745277e-16  … -8.241796e-16
std     1.000880e+00            1.000880e+00  …  1.000880e+00
min    -2.744117e+00           -1.819865e+00  … -1.726901e+00
25%    -7.032397e-01           -7.226392e-01  … -6.749213e-01
50%    -7.162650e-02           -1.782793e-01  … -2.690395e-01
75%     5.307792e-01            4.709834e-01  …  5.220158e-01
max     4.484751e+00            4.910919e+00  …  4.094189e+00

       worst_texture  worst_perimeter  worst_area  worst_smoothness
count   5.690000e+02     5.690000e+02  569.000000      5.690000e+02  \
mean    1.248757e-17    -3.746271e-16    0.000000     -2.372638e-16
std     1.000880e+00     1.000880e+00    1.000880      1.000880e+00
min    -2.223994e+00    -1.693361e+00   -1.222423     -2.682695e+00
25%    -7.486293e-01    -6.895783e-01   -0.642136     -6.912304e-01
50%    -4.351564e-02    -2.859802e-01   -0.341181     -4.684277e-02
75%     6.583411e-01     5.402790e-01    0.357589      5.975448e-01
max     3.885905e+00     4.287337e+00    5.930172      3.955374e+00

       worst_compactness  worst_concavity  worst_concave points
count       5.690000e+02     5.690000e+02          5.690000e+02  \
mean       -3.371644e-16     7.492542e-17          2.247763e-16
std         1.000880e+00     1.000880e+00          1.000880e+00
min        -1.443878e+00    -1.305831e+00         -1.745063e+00
25%        -6.810833e-01    -7.565142e-01         -7.563999e-01
50%        -2.695009e-01    -2.182321e-01         -2.234689e-01
75%         5.396688e-01     5.311411e-01          7.125100e-01
max         5.112877e+00     4.700669e+00          2.685877e+00

       worst_symmetry  worst_fractal dimension
```

```
count    5.690000e+02             5.690000e+02
mean     2.622390e-16            -5.744282e-16
std      1.000880e+00             1.000880e+00
min     -2.160960e+00            -1.601839e+00
25%     -6.418637e-01            -6.919118e-01
50%     -1.274095e-01            -2.164441e-01
75%      4.501382e-01             4.507624e-01
max      6.046041e+00             6.846856e+00

[8 rows x 30 columns]
```

[119]:
```
num_folds=10
kfold=KFold(n_splits=num_folds)
model=KNeighborsClassifier()
results=cross_val_score(model,x_scaled,y,cv=kfold)
print(results.mean())
```

```
0.9666040100250626
```

[120]:
```
x_train,x_test,y_train,y_test=train_test_split(x_scaled,y,test_size=0.
↪3,random_state=42)
```

[121]:
```
knn=KNeighborsClassifier(n_neighbors=5,weights='distance')
knn.fit(x_train,y_train)
```

[121]:
```
KNeighborsClassifier(weights='distance')
```

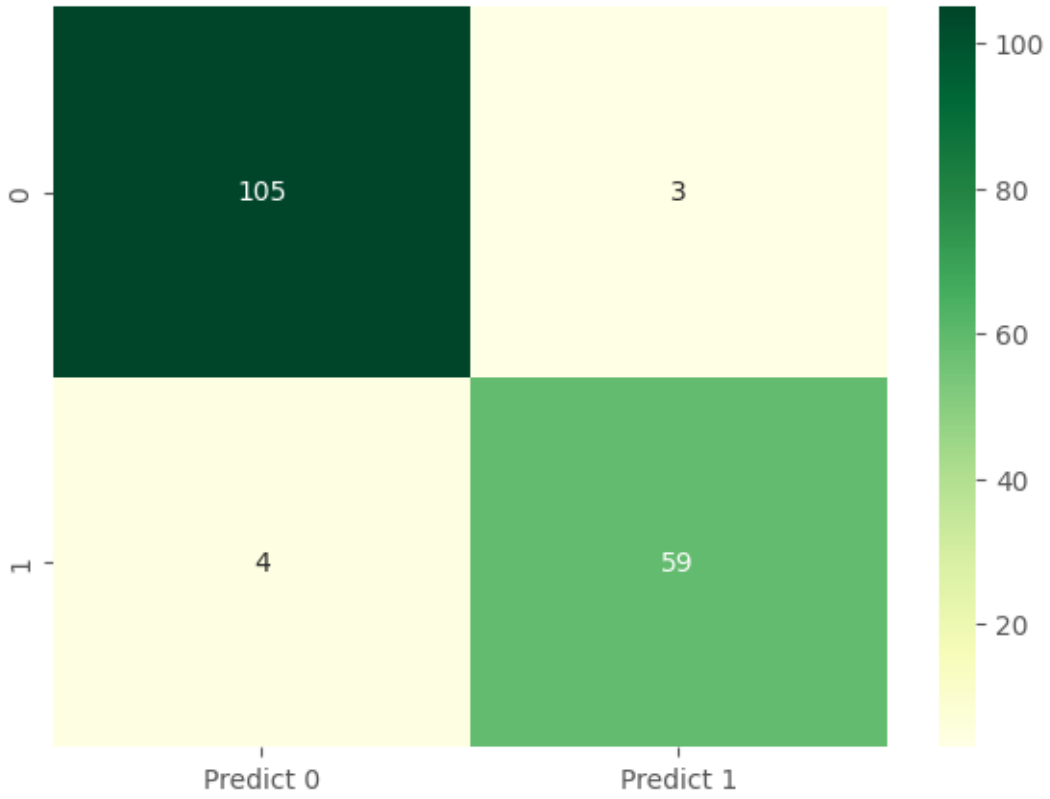
[122]:
```
predicted_labels=knn.predict(x_test)
knn.score(x_test,y_test)
```

[122]:
```
0.9590643274853801
```

[123]:
```
from sklearn import metrics
print('Confusion Matrix')
cm=metrics.confusion_matrix(y_test,predicted_labels,labels=[0,1])
df_cm=pd.DataFrame(cm,index=[i for i in [0,1]],
                   columns=[i for i in ['Predict 0','Predict 1']])
plt.figure(figsize=(7,5))
sns.heatmap(df_cm,annot=True,fmt='.5g',cmap='YlGn')
```

```
Confusion Matrix
```

[123]:
```
<Axes: >
```



[137]:
```
from sklearn.metrics import␣
↪roc_auc_score,roc_curve,classification_report,ConfusionMatrixDisplay


print(classification_report(y_test,predicted_labels))
```

```
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       108
           1       0.95      0.94      0.94        63

    accuracy                           0.96       171
   macro avg       0.96      0.95      0.96       171
weighted avg       0.96      0.96      0.96       171
```

[138]:
```
from sklearn.model_selection import GridSearchCV
#Hyperparameters to be tuned
leaf_size=list(range(1,50))
n_neighbors=list(range(1,30))
p=[1,2]
```

```
hyperparameters=dict(leaf_size=leaf_size,n_neighbors=n_neighbors,p=p)
#create a new KNN object
knn_2=KNeighborsClassifier()
clf=GridSearchCV(knn_2,hyperparameters,cv=10)
#fit the model
best_model=clf.fit(x_scaled,y)
print('Best leaf_size:',best_model.best_estimator_.get_params()['leaf_size'])
print('Best p:',best_model.best_estimator_.get_params()['p'])
print('Best n_neighbors:',best_model.best_estimator_.
 ↪get_params()['n_neighbors'])
```

```
Best leaf_size: 1
Best p: 1
Best n_neighbors: 3
```

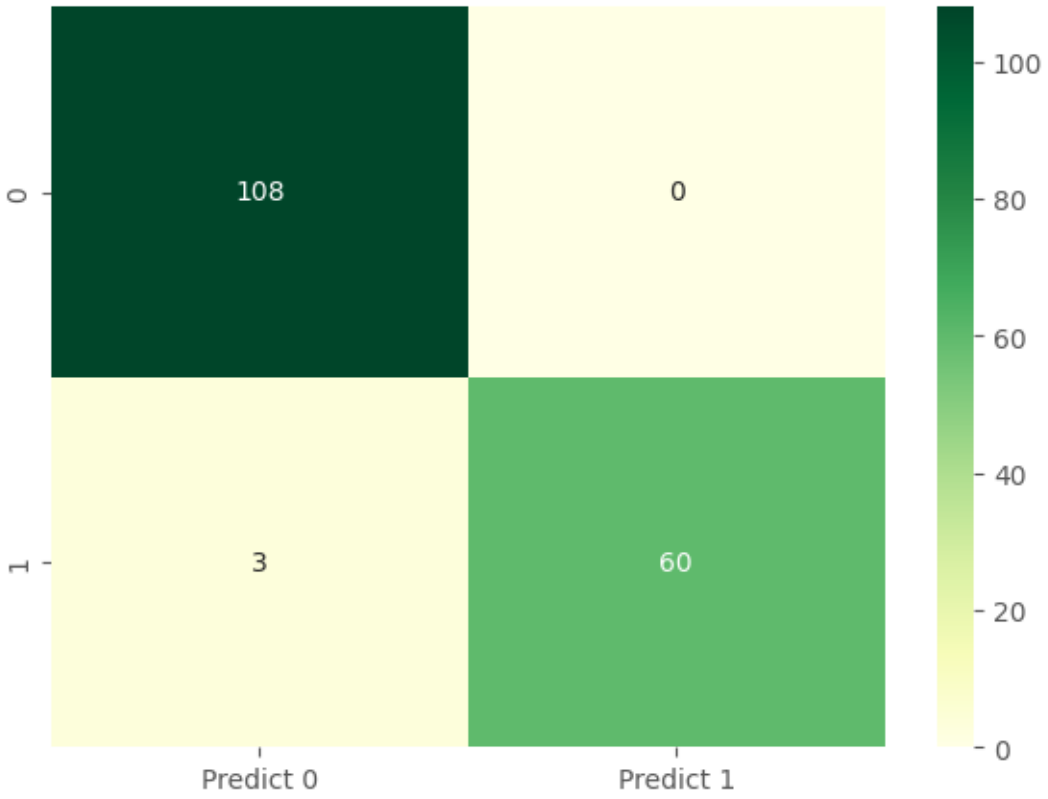
[139]:
```
y_pred=best_model.predict(x_test)
best_model.score(x_test,y_test)
```

[139]:
```
0.9824561403508771
```

[140]:
```
from sklearn import metrics
print('Confusion Matrix')
cm=metrics.confusion_matrix(y_test,y_pred,labels=[0,1])
df_cm=pd.DataFrame(cm,index=[i for i in [0,1]],
                   columns=[i for i in ['Predict 0','Predict 1']])
plt.figure(figsize=(7,5))
sns.heatmap(df_cm,annot=True,fmt='.5g',cmap='YlGn')
```

```
Confusion Matrix
```

[140]:
```
<Axes: >
```



[141]:
```
false_negatives=np.logical_and(y_test!=predicted_labels1,predicted_labels1==0)
x_test[false_negatives]
```

[141]:

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	\
73	-0.092956	-0.814392	-0.063393	-0.201331	0.308838	\
255	-0.047513	-0.521181	-0.022203	-0.149284	0.942210	
414	0.284783	2.448156	0.195281	0.183760	-0.936557	

	mean_compactness	mean_concavity	mean_concave points	mean_symmetry	\
73	0.448373	-0.136966	0.045677	-0.546249	\
255	0.446478	0.114133	0.091333	0.351883	
414	-1.104700	-0.526547	-0.555322	0.147430	

	mean_fractal dimension	…	worst_radius	worst_texture	\
73	0.405774	…	0.062293	-0.784455	\
255	-0.212302	…	0.025018	-0.587414	
414	-1.397419	…	0.205179	1.829188	

	worst_perimeter	worst_area	worst_smoothness	worst_compactness	\
73	0.090513	-0.119860	0.382749	0.635726	\
255	0.024984	-0.095952	0.825491	0.457607	

```
414         0.084556    0.089332         -0.770135          -0.989865

     worst_concavity  worst_concave points  worst_symmetry
73          0.027401              0.360776       -0.504352  \
255         0.233695              0.347072        0.270565
414        -0.563654             -0.743914        0.537498

     worst_fractal dimension
73                  1.055903
255                -0.242489
414                -1.235541

[3 rows x 30 columns]
```

[142]:
```
true_negatives=np.logical_and(y_test==predicted_labels1,predicted_labels1==0)
frames=[x_test[false_negatives],x_test[true_negatives]]
pred_neg=pd.concat(frames)
pred_neg
```

[142]:
```
     mean_radius  mean_texture  mean_perimeter  mean_area  mean_smoothness
73     -0.092956     -0.814392       -0.063393  -0.201331         0.308838  \
255    -0.047513     -0.521181       -0.022203  -0.149284         0.942210
414     0.284783      2.448156        0.195281   0.183760        -0.936557
204    -0.470694     -0.160486       -0.448110  -0.491999         0.234114
431    -0.490575     -0.374576       -0.432457  -0.532101         0.643316
..           …             …               …          …                …
426    -1.035883     -1.002884       -1.008296  -0.913779         0.128078
69     -0.382650     -0.651497       -0.436576  -0.433410         0.138753
542     0.174018      1.426574        0.112489   0.038995        -0.968582
176    -1.199475     -0.286147       -1.127336  -1.002515         0.044814
247    -0.351408     -1.205339       -0.289115  -0.405822        -0.623429

     mean_compactness  mean_concavity  mean_concave points  mean_symmetry
73           0.448373       -0.136966             0.045677      -0.546249  \
255          0.446478        0.114133             0.091333       0.351883
414         -1.104700       -0.526547            -0.555322       0.147430
204          0.027651       -0.109847            -0.276232       0.413949
431          0.516599       -0.142993            -0.539846      -0.002259
..                …               …                    …              …
426         -0.057631       -0.319515            -0.689709       0.413949
69          -0.985496       -0.656240            -0.523080      -0.809117
542         -0.610256       -0.599491            -0.481036       0.103619
176          0.474905        0.526062            -0.303315      -0.520693
247          0.573453        0.610180            -0.235219      -0.787211

     mean_fractal dimension  …  worst_radius  worst_texture
73                 0.405774  …      0.062293      -0.784455  \
```

```
255               -0.212302  …      0.025018      -0.587414
414               -1.397419  …      0.205179       1.829188
204                0.132176  …     -0.269040      -0.168905
431                1.165609  …     -0.701842      -0.450625
..                      …  …           …              …
426                0.900517  …     -0.857154      -0.668836
69                -0.888499  …     -0.581734      -0.963583
542               -0.850224  …      0.049868       1.076850
176                2.603060  …     -1.037316      -0.209616
247                0.183210  …     -0.389147      -1.299041

     worst_perimeter  worst_area  worst_smoothness  worst_compactness
73          0.090513   -0.119860          0.382749           0.635726  \
255         0.024984   -0.095952          0.825491           0.457607
414         0.084556    0.089332         -0.770135          -0.989865
204        -0.333935   -0.356299          0.448503          -0.104741
431        -0.525756   -0.641257          0.553709           0.054930
..               …           …                 …                 …
426        -0.770000   -0.773804          0.014527           0.288394
69         -0.643112   -0.572523         -0.121364          -1.168303
542         0.004134   -0.095249         -1.155891          -0.742153
176        -1.018414   -0.862051         -0.099446           0.259131
247        -0.067352   -0.424506         -0.305475           2.103300

     worst_concavity  worst_concave points  worst_symmetry
73          0.027401              0.360776       -0.504352  \
255         0.233695              0.347072        0.270565
414        -0.563654             -0.743914        0.537498
204        -0.024412             -0.199563        0.183204
431        -0.152986             -0.622863       -0.557739
..               …                     …               …
426         0.104162             -0.327467        0.192911
69         -0.807368             -0.849434       -0.837615
542        -0.532950             -0.077750       -0.289188
176         0.366586             -0.236107       -0.463908
247         2.401216              0.631809       -0.423463

     worst_fractal dimension
73                  1.055903
255                -0.242489
414                -1.235541
204                 0.196958
431                 0.534440
..                       …
426                 0.693484
69                 -1.099772
542                -0.797202
```

```
176                 1.787392
247                 1.876057

[111 rows x 30 columns]
```

[143]:
```
stacks=[y_test[false_negatives],y_test[true_negatives]]
y_labels=np.hstack(stacks)
y_labels.shape
print(y_labels)
```

```
[1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

[146]:
```
new_df=pd.DataFrame(data=pred_neg)
new_df['diagnosis']=y_labels
new_df.shape
new_df.head()
```

[146]:

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	\
73	-0.092956	-0.814392	-0.063393	-0.201331	0.308838	\
255	-0.047513	-0.521181	-0.022203	-0.149284	0.942210	
414	0.284783	2.448156	0.195281	0.183760	-0.936557	
204	-0.470694	-0.160486	-0.448110	-0.491999	0.234114	
431	-0.490575	-0.374576	-0.432457	-0.532101	0.643316	

	mean_compactness	mean_concavity	mean_concave points	mean_symmetry	\
73	0.448373	-0.136966	0.045677	-0.546249	\
255	0.446478	0.114133	0.091333	0.351883	
414	-1.104700	-0.526547	-0.555322	0.147430	
204	0.027651	-0.109847	-0.276232	0.413949	
431	0.516599	-0.142993	-0.539846	-0.002259	

	mean_fractal dimension	…	worst_texture	worst_perimeter	worst_area	\
73	0.405774	…	-0.784455	0.090513	-0.119860	\
255	-0.212302	…	-0.587414	0.024984	-0.095952	
414	-1.397419	…	1.829188	0.084556	0.089332	
204	0.132176	…	-0.168905	-0.333935	-0.356299	
431	1.165609	…	-0.450625	-0.525756	-0.641257	

	worst_smoothness	worst_compactness	worst_concavity	\
73	0.382749	0.635726	0.027401	\
255	0.825491	0.457607	0.233695	
414	-0.770135	-0.989865	-0.563654	
204	0.448503	-0.104741	-0.024412	
431	0.553709	0.054930	-0.152986	

```
     worst_concave points  worst_symmetry  worst_fractal dimension  diagnosis
73                0.360776       -0.504352                 1.055903          1
255               0.347072        0.270565                -0.242489          1
414              -0.743914        0.537498                -1.235541          1
204              -0.199563        0.183204                 0.196958          0
431              -0.622863       -0.557739                 0.534440          0

[5 rows x 31 columns]
```

[147]:
```
new_df['diagnosis'].value_counts()
```

[147]:
```
diagnosis
0    108
1      3
Name: count, dtype: int64
```

[148]:
```
new_df_corr=new_df.corr()['diagnosis'].abs().sort_values(ascending=False)
new_df_corr
```

[148]:
```
diagnosis                 1.000000
worst_area                0.316577
worst_radius              0.289529
worst_perimeter           0.286102
SE_area                   0.230159
mean_area                 0.229837
mean_perimeter            0.216750
mean_radius               0.211266
worst_concave points      0.169396
mean_concave points       0.169167
SE_radius                 0.131486
worst_compactness         0.130878
mean_concavity            0.115360
mean_compactness          0.113136
SE_perimeter              0.112116
mean_texture              0.108400
worst_concavity           0.087137
worst_smoothness          0.081663
worst_texture             0.081404
mean_fractal dimension    0.080556
SE_smoothness             0.070546
worst_symmetry            0.068493
SE_texture                0.062043
SE_fractal dimension      0.059387
mean_smoothness           0.051853
SE_concave points         0.031789
worst_fractal dimension   0.030647
mean_symmetry             0.024369
```

```
SE_compactness             0.005433
SE_symmetry                0.002118
SE_concavity               0.001798
Name: diagnosis, dtype: float64
```