Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Технологии машинного обучения»
на тему
«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-61Б
Агличеев М. С.

Москва —2022 г.

1. Разведочный анализ данных. Исследование и визуализация данных.

1.1. 1) Текстовое описание набора данных

В качестве набора данных мы будем использовать набор данных о диагностике рака в штате Висконсин - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.

Он хорошо подходит для задачи классификации - сведения гистологических данных (размера клетки из биопсии, размера её ядра и т.д.) в 2 класса - злокачественной и доброкачественной ткани.

Датасет состоит из одного файла - обучающей выборки

Этот файл содержит следующие колонки: - radius - среднее расстояние от центра ядра клетки до её периметра - texture - стандартное отклонение цвета после преобразования изображения в чёрно-белый формат - perimeter - длина периметра клетки - area - площадь клетки - smoothness - местная вариация радиуса - compactness (perimeter^2 / area - 1.0) - concavity - насколько сильны впуклые части контура - concave points - количество впуклых частей - symmetry - симметричность клетки - fractal dimension ("coastline approximation" - 1) Для каждого из этих параметров вычисляли среднее, стандартное отклонение и наихудшее значение (среднее трёх наибольших значений) для всех клеток из гистологического образца, таким образом получая 30 параметров - class - либо WDBC-Malignant (злокачественность), либо WDBC-Benign (доброкачественность)

2. Импорт библиотек

Импортируем библиотеки с помощью команды import. Как правило, все команды import размещают в первой ячейке ноутбука, но мы в этом примере будем подключать все библиотеки последовательно, по мере их использования.

[53]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

3. Загрузка данных

Загрузим файлы датасета в помощью библиотеки Pandas.

Используемый датасет уже присутствует в составе библиотеки sklearn, поэтому его надо просто преобразовать в Pandas DataFrame

[54]:
```
from sklearn.datasets import *
```

[55]:
```
cancer = load_breast_cancer()
data = pd.DataFrame(data= np.c_[cancer['data'], cancer['target']],
                    columns= cancer['feature_names'].tolist() +␣
 ↪['malignancy'])
```

[56]:
```
data
```

[56]:

	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
..	…	…	…	…	…
564	21.56	22.39	142.00	1479.0	0.11100
565	20.13	28.25	131.20	1261.0	0.09780
566	16.60	28.08	108.30	858.1	0.08455
567	20.60	29.33	140.10	1265.0	0.11780
568	7.76	24.54	47.92	181.0	0.05263

	mean compactness	mean concavity	mean concave points	mean symmetry ↪\
0	0.27760	0.30010	0.14710	0.2419
1	0.07864	0.08690	0.07017	0.1812
2	0.15990	0.19740	0.12790	0.2069
3	0.28390	0.24140	0.10520	0.2597
4	0.13280	0.19800	0.10430	0.1809
..	…	…	…	…
564	0.11590	0.24390	0.13890	0.1726
565	0.10340	0.14400	0.09791	0.1752
566	0.10230	0.09251	0.05302	0.1590
567	0.27700	0.35140	0.15200	0.2397
568	0.04362	0.00000	0.00000	0.1587

	mean fractal dimension	…	worst texture	worst perimeter	worst area ↪ \
0	0.07871	…	17.33	184.60	2019.0
1	0.05667	…	23.41	158.80	1956.0
2	0.05999	…	25.53	152.50	1709.0
3	0.09744	…	26.50	98.87	567.7
4	0.05883	…	16.67	152.20	1575.0
..	…	…	…	…	…
564	0.05623	…	26.40	166.10	2027.0
565	0.05533	…	38.25	155.00	1731.0
566	0.05648	…	34.12	126.70	1124.0
567	0.07016	…	39.42	184.60	1821.0
568	0.05884	…	30.37	59.16	268.6

	worst smoothness	worst compactness	worst concavity \
0	0.16220	0.66560	0.7119
1	0.12380	0.18660	0.2416
2	0.14440	0.42450	0.4504
3	0.20980	0.86630	0.6869
4	0.13740	0.20500	0.4000

```
..                ...               ...               ...
564           0.14100           0.21130           0.4107
565           0.11660           0.19220           0.3215
566           0.11390           0.30940           0.3403
567           0.16500           0.86810           0.9387
568           0.08996           0.06444           0.0000

     worst concave points  worst symmetry  worst fractal dimension ␣
 ↪malignancy
0                  0.2654          0.4601                  0.11890 ␣
 ↪0.0
1                  0.1860          0.2750                  0.08902 ␣
 ↪0.0
2                  0.2430          0.3613                  0.08758 ␣
 ↪0.0
3                  0.2575          0.6638                  0.17300 ␣
 ↪0.0
4                  0.1625          0.2364                  0.07678 ␣
 ↪0.0
..                    ...             ...                      ...       ...
564                0.2216          0.2060                  0.07115 ␣
 ↪0.0
565                0.1628          0.2572                  0.06637 ␣
 ↪0.0
566                0.1418          0.2218                  0.07820 ␣
 ↪0.0
567                0.2650          0.4087                  0.12400 ␣
 ↪0.0
568                0.0000          0.2871                  0.07039 ␣
 ↪1.0

[569 rows x 31 columns]
```

4. 2) Основные характеристики датасета

[57]:
```
# Первые 5 строк датасета
data.head()
```

[57]:
```
   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 texture  worst perimeter  worst area  \
0                 0.07871  …          17.33           184.60      2019.0
1                 0.05667  …          23.41           158.80      1956.0
2                 0.05999  …          25.53           152.50      1709.0
3                 0.09744  …          26.50            98.87       567.7
4                 0.05883  …          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  malignancy
0          0.4601                  0.11890         0.0
1          0.2750                  0.08902         0.0
2          0.3613                  0.08758         0.0
3          0.6638                  0.17300         0.0
4          0.2364                  0.07678         0.0

[5 rows x 31 columns]
```

[58]:
```
# Размер датасета - 569 строк, 31 колонка
data.shape
```

[58]:
```
(569, 31)
```

[59]:
```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 569
```

[60]:
```
# Список колонок
data.columns
```

[60]:
```
Index(['mean radius', 'mean texture', 'mean perimeter', 'mean area',
       'mean smoothness', 'mean compactness', 'mean concavity',
       'mean concave points', 'mean symmetry', 'mean fractal dimension',
       'radius error', 'texture error', 'perimeter error', 'area error',
```

```
       'smoothness error', 'compactness error', 'concavity error',
       'concave points error', 'symmetry error', 'fractal dimension error',
       'worst radius', 'worst texture', 'worst perimeter', 'worst area',
       'worst smoothness', 'worst compactness', 'worst concavity',
       'worst concave points', 'worst symmetry', 'worst fractal dimension',
       'malignancy'],
      dtype='object')
```

[61]:
```
# Список колонок с типами данных
data.dtypes
```

[61]:
```
mean radius                float64
mean texture               float64
mean perimeter             float64
mean area                  float64
mean smoothness            float64
mean compactness           float64
mean concavity             float64
mean concave points        float64
mean symmetry              float64
mean fractal dimension     float64
radius error               float64
texture error              float64
perimeter error            float64
area error                 float64
smoothness error           float64
compactness error          float64
concavity error            float64
concave points error       float64
symmetry error             float64
fractal dimension error    float64
worst radius               float64
worst texture              float64
worst perimeter            float64
worst area                 float64
worst smoothness           float64
worst compactness          float64
worst concavity            float64
worst concave points       float64
worst symmetry             float64
worst fractal dimension    float64
malignancy                 float64
dtype: object
```

[62]:
```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
mean radius - 0
mean texture - 0
mean perimeter - 0
mean area - 0
mean smoothness - 0
mean compactness - 0
mean concavity - 0
mean concave points - 0
mean symmetry - 0
mean fractal dimension - 0
radius error - 0
texture error - 0
perimeter error - 0
area error - 0
smoothness error - 0
compactness error - 0
concavity error - 0
concave points error - 0
symmetry error - 0
fractal dimension error - 0
worst radius - 0
worst texture - 0
worst perimeter - 0
worst area - 0
worst smoothness - 0
worst compactness - 0
worst concavity - 0
worst concave points - 0
worst symmetry - 0
worst fractal dimension - 0
malignancy - 0
```

[63]:
```
# Основные статистические характеристки набора данных
data.describe()
```

[63]:
```
       mean radius  mean texture  mean perimeter    mean area  \
count   569.000000    569.000000      569.000000   569.000000
mean     14.127292     19.289649       91.969033   654.889104
std       3.524049      4.301036       24.298981   351.914129
min       6.981000      9.710000       43.790000   143.500000
25%      11.700000     16.170000       75.170000   420.300000
50%      13.370000     18.840000       86.240000   551.100000
75%      15.780000     21.800000      104.100000   782.700000
max      28.110000     39.280000      188.500000  2501.000000

       mean smoothness  mean compactness  mean concavity  mean concave␣
 ↪points  \
count       569.000000        569.000000      569.000000           569.
 ↪000000
```

```
mean          0.096360          0.104341        0.088799             0.
↪048919
std           0.014064          0.052813        0.079720             0.
↪038803
min           0.052630          0.019380        0.000000             0.
↪000000
25%           0.086370          0.064920        0.029560             0.
↪020310
50%           0.095870          0.092630        0.061540             0.
↪033500
75%           0.105300          0.130400        0.130700             0.
↪074000
max           0.163400          0.345400        0.426800             0.
↪201200
```

```
       mean symmetry  mean fractal dimension  …  worst texture  \
count     569.000000              569.000000  …     569.000000
mean        0.181162                0.062798  …      25.677223
std         0.027414                0.007060  …       6.146258
min         0.106000                0.049960  …      12.020000
25%         0.161900                0.057700  …      21.080000
50%         0.179200                0.061540  …      25.410000
75%         0.195700                0.066120  …      29.720000
max         0.304000                0.097440  …      49.540000
```

```
       worst perimeter   worst area  worst smoothness  worst compactness  \
count       569.000000   569.000000        569.000000         569.000000
mean        107.261213   880.583128          0.132369           0.254265
std          33.602542   569.356993          0.022832           0.157336
min          50.410000   185.200000          0.071170           0.027290
25%          84.110000   515.300000          0.116600           0.147200
50%          97.660000   686.500000          0.131300           0.211900
75%         125.400000  1084.000000          0.146000           0.339100
max         251.200000  4254.000000          0.222600           1.058000
```

```
       worst concavity  worst concave points  worst symmetry  \
count       569.000000            569.000000      569.000000
mean          0.272188              0.114606        0.290076
std           0.208624              0.065732        0.061867
min           0.000000              0.000000        0.156500
25%           0.114500              0.064930        0.250400
50%           0.226700              0.099930        0.282200
75%           0.382900              0.161400        0.317900
max           1.252000              0.291000        0.663800
```

```
       worst fractal dimension  malignancy
count               569.000000  569.000000
mean                  0.083946    0.627417
std                   0.018061    0.483918
```

```
min                     0.055040    0.000000
25%                     0.071460    0.000000
50%                     0.080040    1.000000
75%                     0.092080    1.000000
max                     0.207500    1.000000

[8 rows x 31 columns]
```

[64]:
```
# Определим уникальные значения для целевого признака
data['malignancy'].unique()
```

[64]:
```
array([0., 1.])
```

Целевой признак является бинарным и содержит только значения 0 и 1.

5. 3) Визуальное исследование датасета

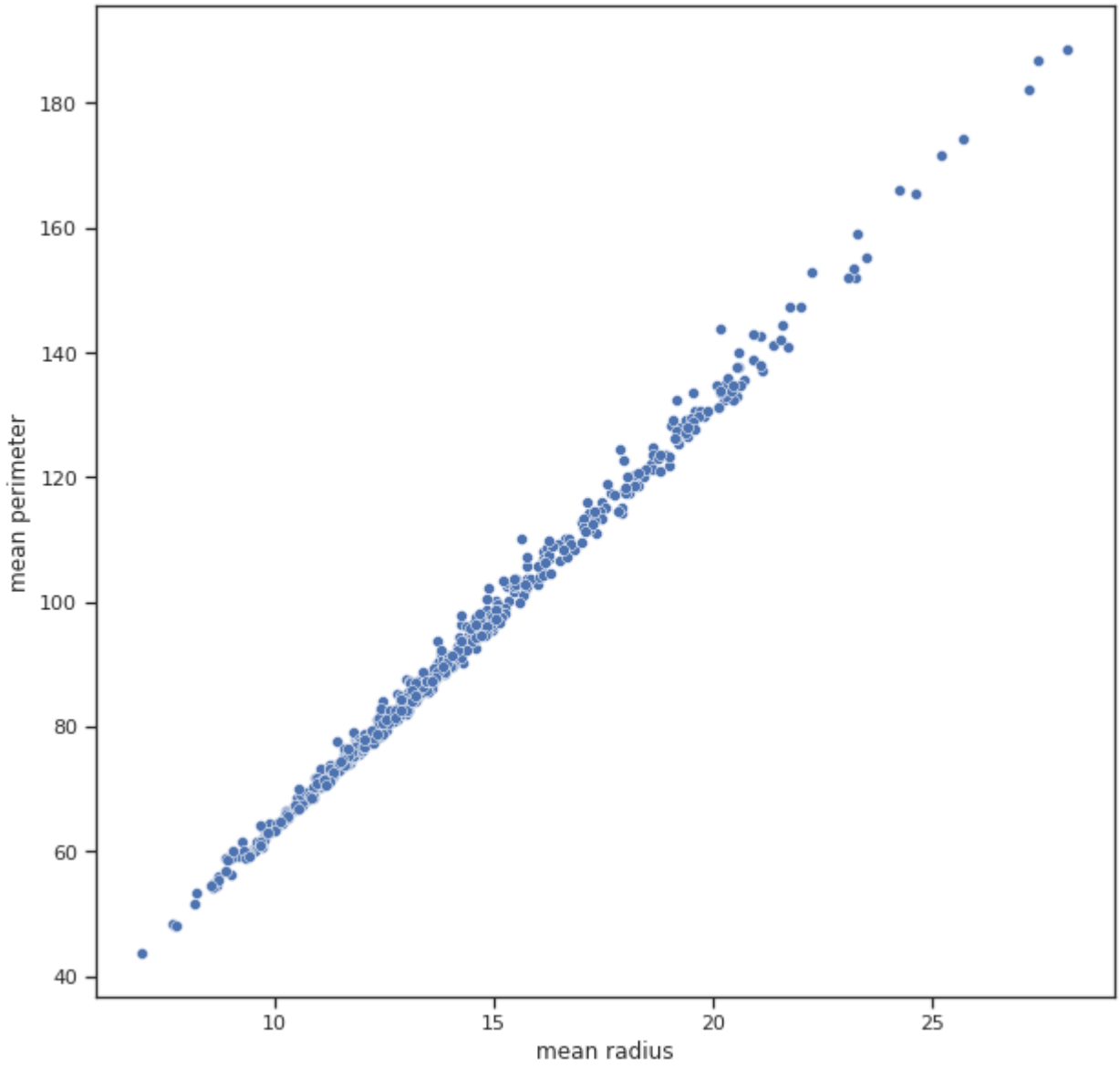
Для визуального исследования могут быть использованы различные виды диаграмм, мы построим только некоторые варианты диаграмм, которые используются достаточно часто.

5.0.1. Диаграмма рассеяния

Позволяет построить распределение двух колонок данных и визуально обнаружить наличие зависимости. Не предполагается, что значения упорядочены (например, по времени).

[65]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='mean radius', y='mean perimeter', data=data)
```

[65]:
```
<AxesSubplot:xlabel='mean radius', ylabel='mean perimeter'>
```

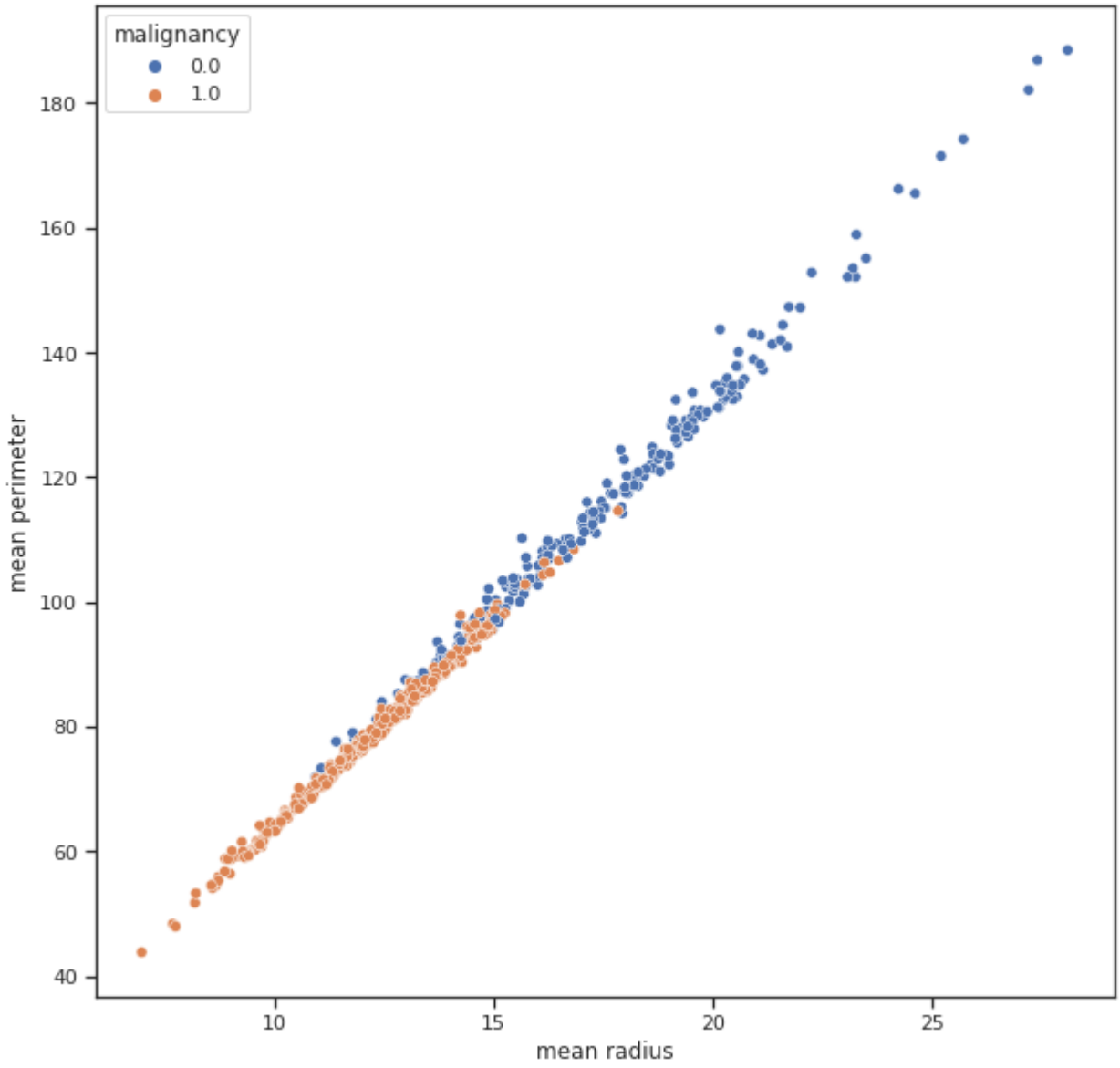


Можно видеть что между полями “mean radius” и “mean perimeter” пристутствует почти линейная зависимость.

Посмотрим насколько на эту зависимость влияет целевой признак.

[66]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='mean radius', y='mean perimeter', data=data,␣
↪hue='malignancy')
```

[66]:
```
<AxesSubplot:xlabel='mean radius', ylabel='mean perimeter'>
```



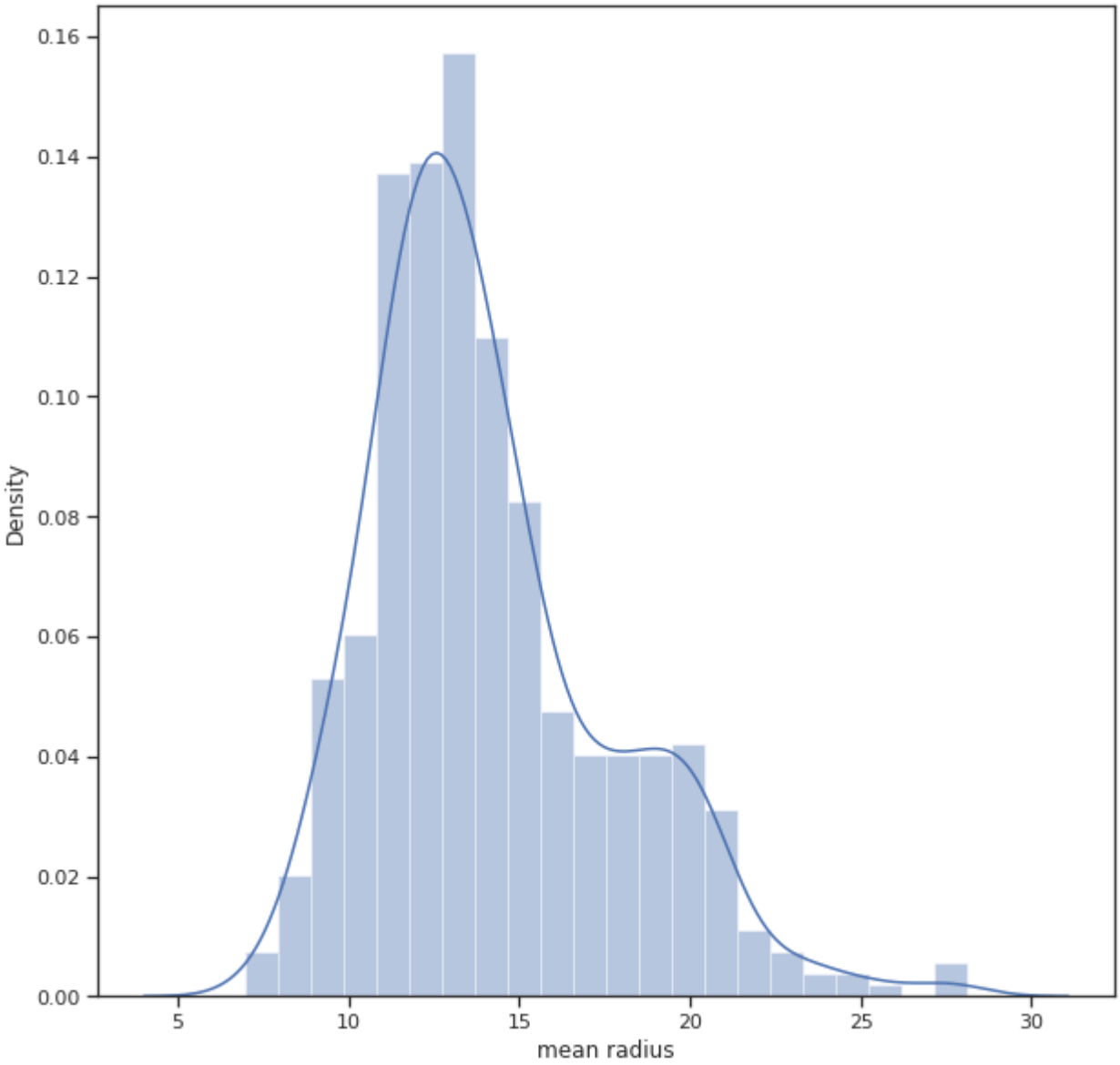
5.1. Гистограмма

Позволяет оценить плотность вероятности распределения данных.

[67]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['mean radius'])
```

```
/home/mikhail/.local/lib/python3.8/site-packages/seaborn/distributions.py:
↪2619:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a␣
↪figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

[67]: <AxesSubplot:xlabel='mean radius', ylabel='Density'>

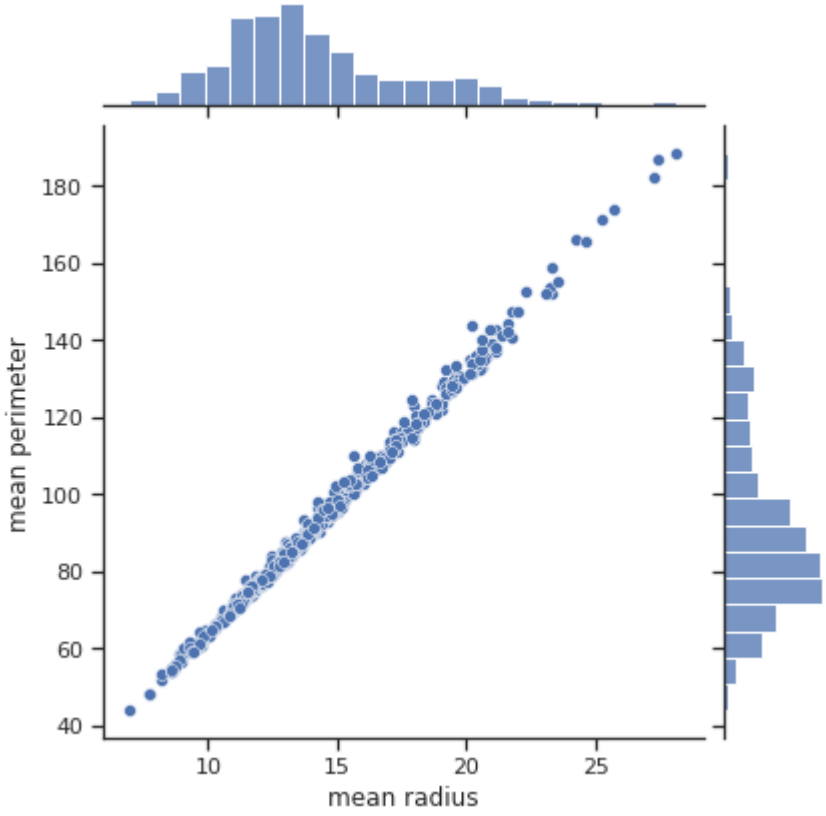


5.1.1. Jointplot

Комбинация гистограмм и диаграмм рассеивания.

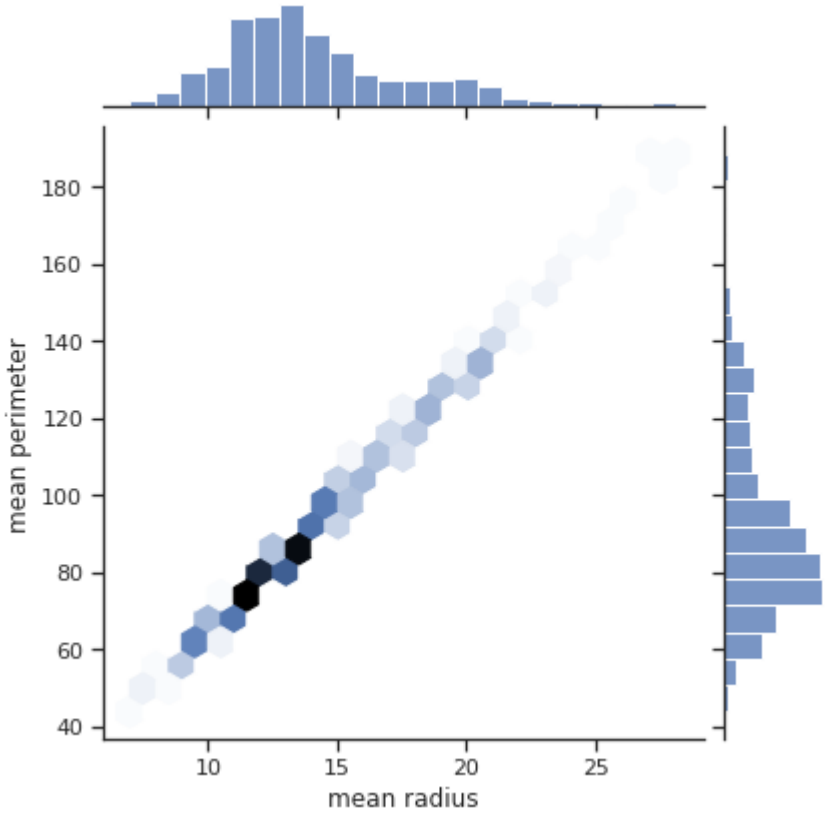
[68]:
```
sns.jointplot(x='mean radius', y='mean perimeter', data=data)
```

[68]: <seaborn.axisgrid.JointGrid at 0x7f3e1e6c8d00>



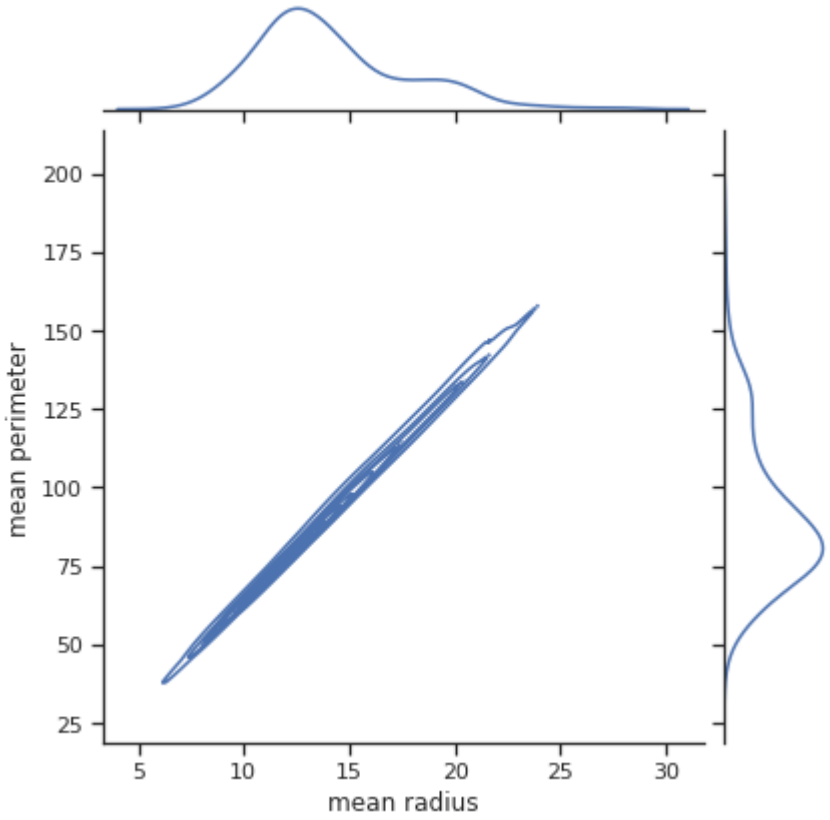
[69]:
```
sns.jointplot(x='mean radius', y='mean perimeter', data=data, kind="hex")
```

[69]:
```
<seaborn.axisgrid.JointGrid at 0x7f3e1e450dc0>
```



[70]:
```
sns.jointplot(x='mean radius', y='mean perimeter', data=data, kind="kde")
```

[70]:
```
<seaborn.axisgrid.JointGrid at 0x7f3e1e2d3af0>
```



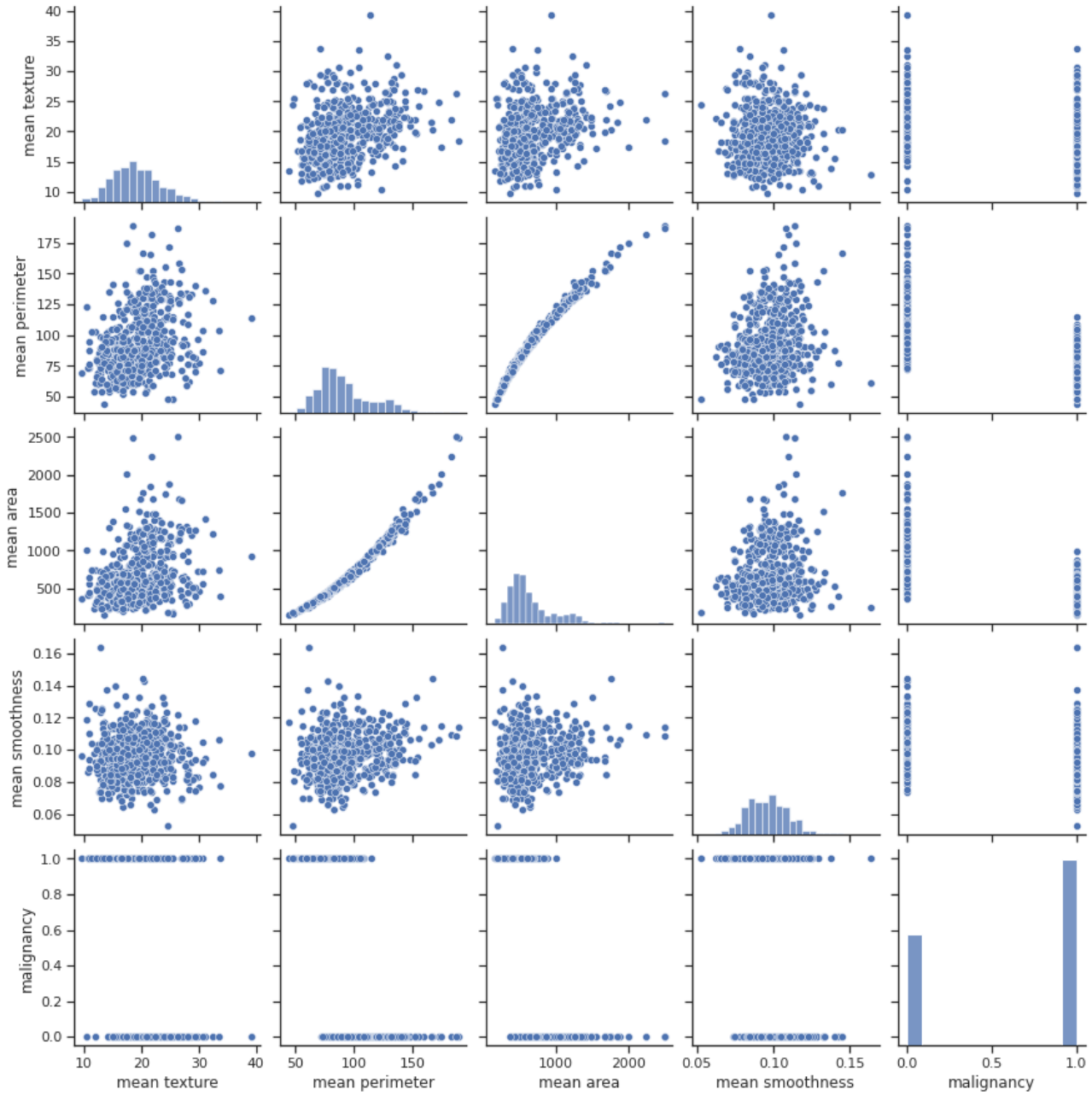
5.1.2. “Парные диаграммы”

Комбинация гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответствуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

[71]:
```
cols_to_plot = data.columns[1:5].tolist() + ['malignancy']
sns.pairplot(data[cols_to_plot])
```

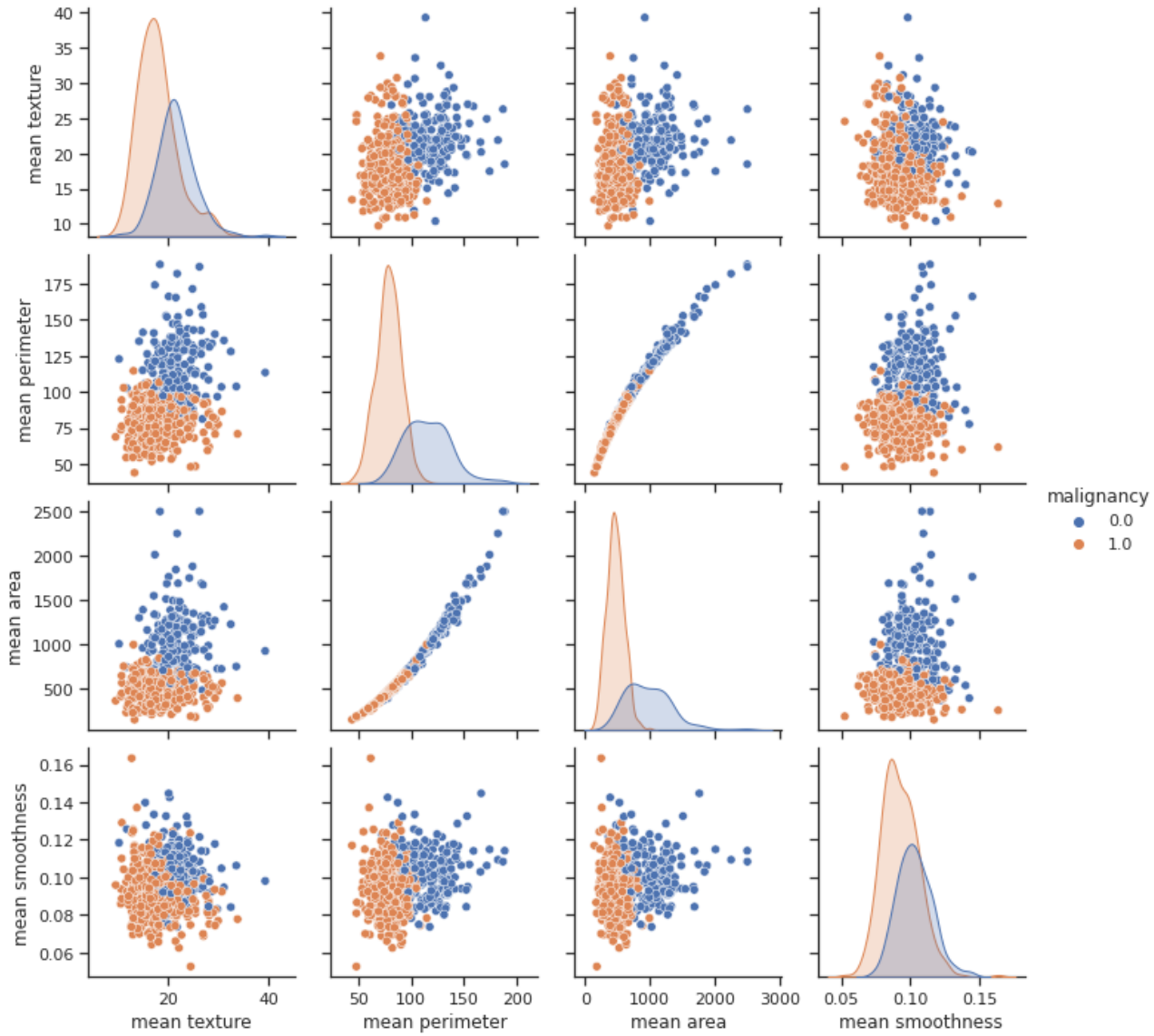
[71]:
```
<seaborn.axisgrid.PairGrid at 0x7f3e1e06d490>
```



С помощью параметра “hue” возможна группировка по значениям какого-либо признака.

[72]:
```
sns.pairplot(data[cols_to_plot], hue="malignancy")
```

[72]:
```
<seaborn.axisgrid.PairGrid at 0x7f3e1e2e2790>
```

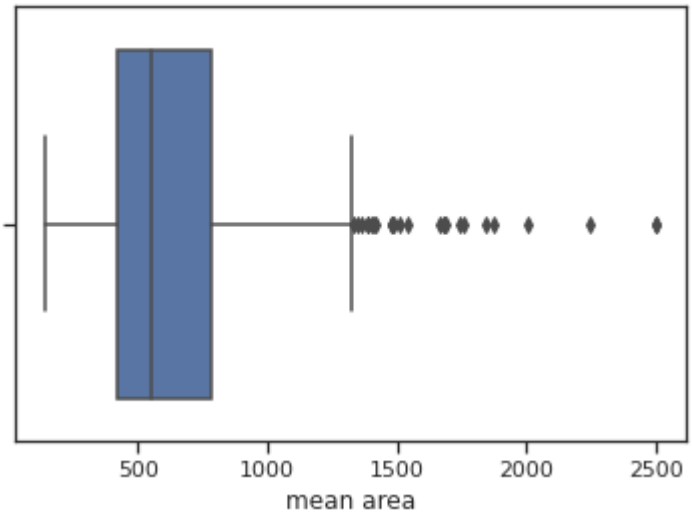


5.1.3. Ящик с усами

Отображает одномерное распределение вероятности.

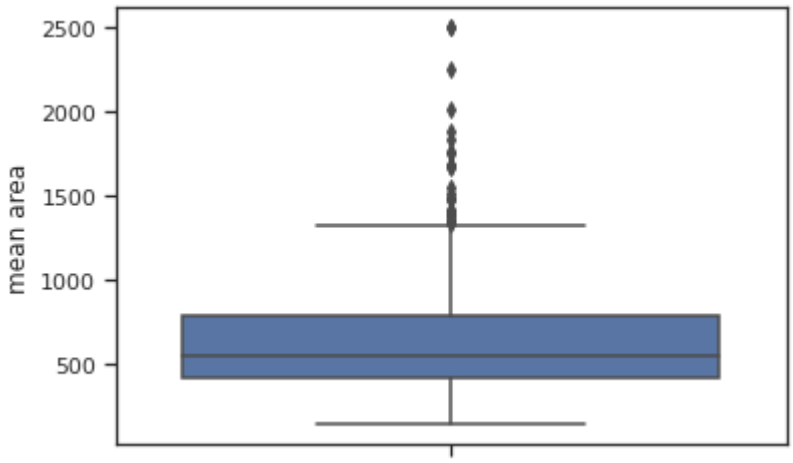
[73]:
```
sns.boxplot(x=data['mean area'])
```

[73]:
```
<AxesSubplot:xlabel='mean area'>
```



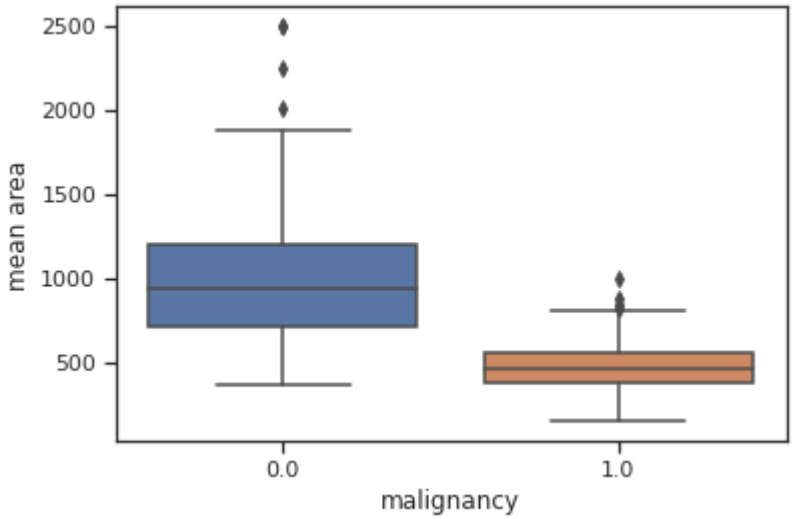
[74]:
```
# По вертикали
sns.boxplot(y=data['mean area'])
```

[74]:
```
<AxesSubplot:ylabel='mean area'>
```



[75]:
```
# Распределение параметра mean area сгруппированные по malignancy.
sns.boxplot(x='malignancy', y='mean area', data=data)
```

[75]:
```
<AxesSubplot:xlabel='malignancy', ylabel='mean area'>
```

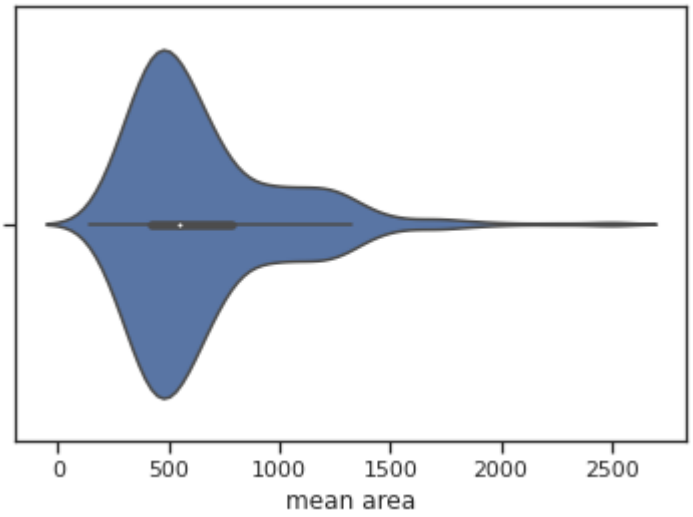


5.1.4. Violin plot

Похоже на предыдущую диаграмму, но по краям отображаются распределения плотности - https://en.wikipedia.org/wiki/Kernel_density_estimation

[76]:
```
sns.violinplot(x=data['mean area'])
```

[76]:
```
<AxesSubplot:xlabel='mean area'>
```

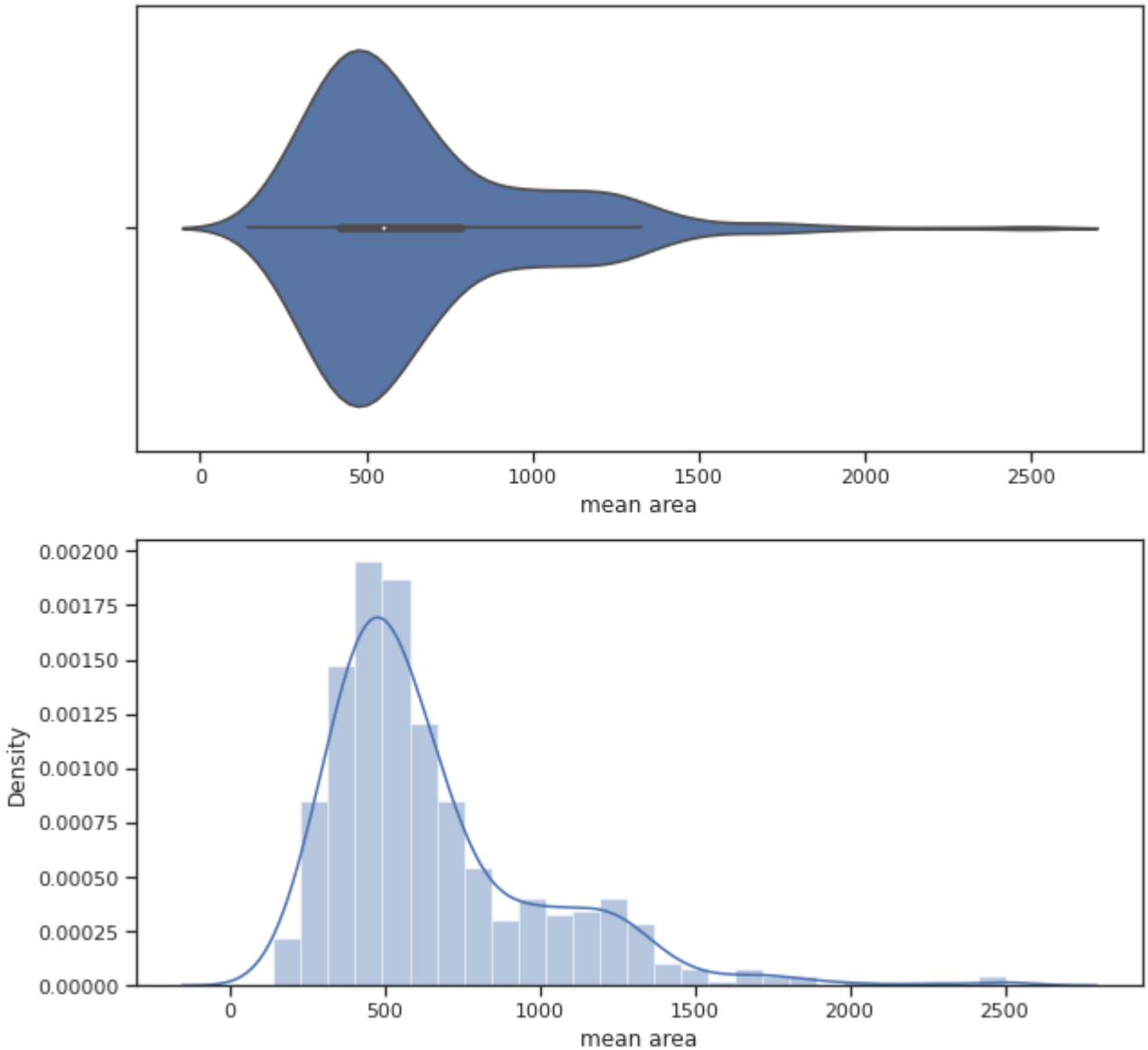


[77]:
```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['mean area'])
sns.distplot(data['mean area'], ax=ax[1])
```

```
/home/mikhail/.local/lib/python3.8/site-packages/seaborn/distributions.py:
↪2619:
FutureWarning: `distplot` is a deprecated function and will be removed in a
```

```
future version. Please adapt your code to use either `displot` (a␣
 ↪figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

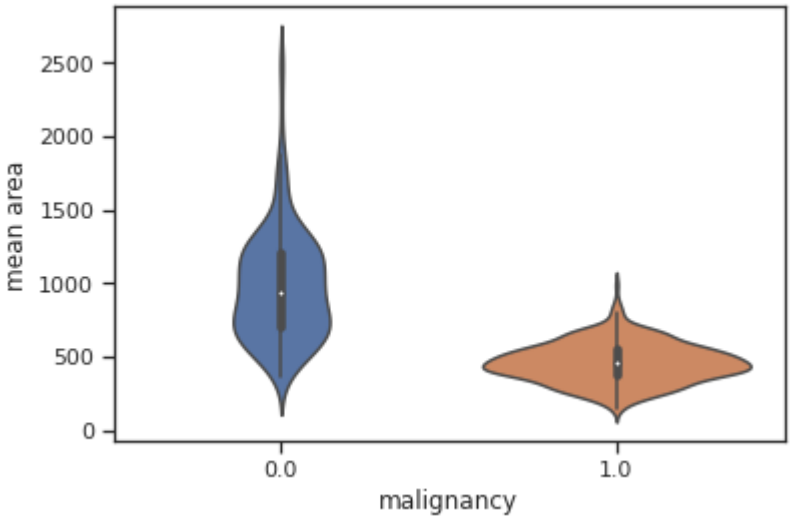
[77]: <AxesSubplot:xlabel='mean area', ylabel='Density'>



Из приведенных графиков видно, что violinplot действительно показывает распределение плотности.

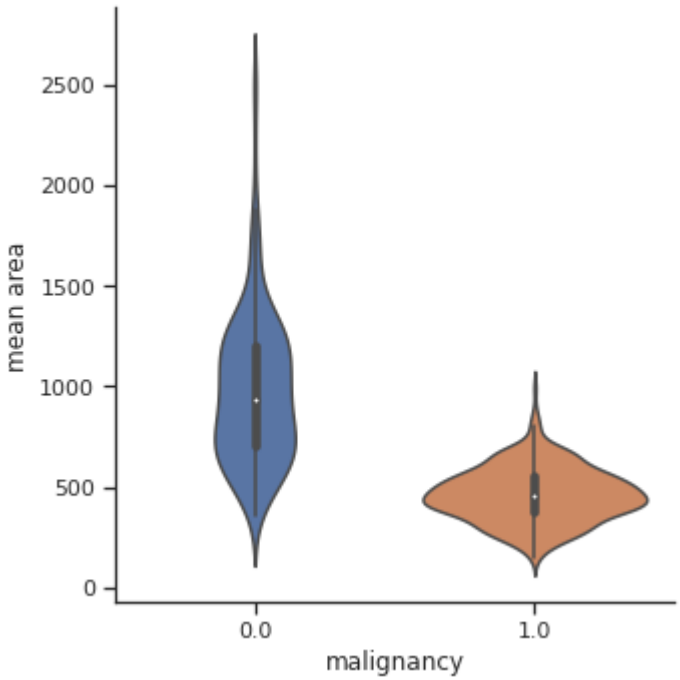
[78]:
```
# Распределение параметра mean area сгруппированные по malignancy.
sns.violinplot(x='malignancy', y='mean area', data=data)
```

[78]: <AxesSubplot:xlabel='malignancy', ylabel='mean area'>



[79]:
```
sns.catplot(x='malignancy', y='mean area', data=data, kind="violin",␣
↪split=True)
```

[79]:
```
<seaborn.axisgrid.FacetGrid at 0x7f3e2618a340>
```



6. 4) Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи: 1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере это колонка “malignancy”). Именно эти признаки будут наиболее

информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели. 1. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

[80]:
```
np.abs(data.corr())
```

[80]:

	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
malignancy                    0.730029      0.415185        0.742636   0.
↪708984
```

```
                         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
malignancy	0.358560	0.596534	0.696360

	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	
malignancy	0.776614	0.330499	

	mean fractal dimension	…	worst texture	\
mean radius	0.311631	…	0.297008	
mean texture	0.076437	…	0.912045	
mean perimeter	0.261477	…	0.303038	
mean area	0.283110	…	0.287489	

mean smoothness	0.584792	…	0.036072
mean compactness	0.565369	…	0.248133
mean concavity	0.336783	…	0.299879
mean concave points	0.166917	…	0.292752
mean symmetry	0.479921	…	0.090651
mean fractal dimension	1.000000	…	0.051269
radius error	0.000111	…	0.194799
texture error	0.164174	…	0.409003
perimeter error	0.039830	…	0.200371
area error	0.090170	…	0.196497
smoothness error	0.401964	…	0.074743
compactness error	0.559837	…	0.143003
concavity error	0.446630	…	0.100241
concave points error	0.341198	…	0.086741
symmetry error	0.345007	…	0.077473
fractal dimension error	0.688132	…	0.003195
worst radius	0.253691	…	0.359921
worst texture	0.051269	…	1.000000
worst perimeter	0.205151	…	0.365098
worst area	0.231854	…	0.345842
worst smoothness	0.504942	…	0.225429
worst compactness	0.458798	…	0.360832
worst concavity	0.346234	…	0.368366
worst concave points	0.175325	…	0.359755
worst symmetry	0.334019	…	0.233027
worst fractal dimension	0.767297	…	0.219122
malignancy	0.012838	…	0.456903

	worst perimeter	worst area	worst smoothness	\
mean radius	0.965137	0.941082	0.119616	
mean texture	0.358040	0.343546	0.077503	
mean perimeter	0.970387	0.941550	0.150549	
mean area	0.959120	0.959213	0.123523	
mean smoothness	0.238853	0.206718	0.805324	
mean compactness	0.590210	0.509604	0.565541	
mean concavity	0.729565	0.675987	0.448822	
mean concave points	0.855923	0.809630	0.452753	
mean symmetry	0.219169	0.177193	0.426675	
mean fractal dimension	0.205151	0.231854	0.504942	
radius error	0.719684	0.751548	0.141919	
texture error	0.102242	0.083195	0.073658	
perimeter error	0.721031	0.730713	0.130054	
area error	0.761213	0.811408	0.125389	
smoothness error	0.217304	0.182195	0.314457	
compactness error	0.260516	0.199371	0.227394	
concavity error	0.226680	0.188353	0.168481	
concave points error	0.394999	0.342271	0.215351	
symmetry error	0.103753	0.110343	0.012662	
fractal dimension error	0.001000	0.022736	0.170568	
worst radius	0.993708	0.984015	0.216574	

worst texture	0.365098	0.345842	0.225429
worst perimeter	1.000000	0.977578	0.236775
worst area	0.977578	1.000000	0.209145
worst smoothness	0.236775	0.209145	1.000000
worst compactness	0.529408	0.438296	0.568187
worst concavity	0.618344	0.543331	0.518523
worst concave points	0.816322	0.747419	0.547691
worst symmetry	0.269493	0.209146	0.493838
worst fractal dimension	0.138957	0.079647	0.617624
malignancy	0.782914	0.733825	0.421465

	worst compactness	worst concavity	\
mean radius	0.413463	0.526911	
mean texture	0.277830	0.301025	
mean perimeter	0.455774	0.563879	
mean area	0.390410	0.512606	
mean smoothness	0.472468	0.434926	
mean compactness	0.865809	0.816275	
mean concavity	0.754968	0.884103	
mean concave points	0.667454	0.752399	
mean symmetry	0.473200	0.433721	
mean fractal dimension	0.458798	0.346234	
radius error	0.287103	0.380585	
texture error	0.092439	0.068956	
perimeter error	0.341919	0.418899	
area error	0.283257	0.385100	
smoothness error	0.055558	0.058298	
compactness error	0.678780	0.639147	
concavity error	0.484858	0.662564	
concave points error	0.452888	0.549592	
symmetry error	0.060255	0.037119	
fractal dimension error	0.390159	0.379975	
worst radius	0.475820	0.573975	
worst texture	0.360832	0.368366	
worst perimeter	0.529408	0.618344	
worst area	0.438296	0.543331	
worst smoothness	0.568187	0.518523	
worst compactness	1.000000	0.892261	
worst concavity	0.892261	1.000000	
worst concave points	0.801080	0.855434	
worst symmetry	0.614441	0.532520	
worst fractal dimension	0.810455	0.686511	
malignancy	0.590998	0.659610	

	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
malignancy	0.793566	0.416294

	worst fractal dimension	malignancy
mean radius	0.007066	0.730029
mean texture	0.119205	0.415185
mean perimeter	0.051019	0.742636
mean area	0.003738	0.708984
mean smoothness	0.499316	0.358560
mean compactness	0.687382	0.596534
mean concavity	0.514930	0.696360
mean concave points	0.368661	0.776614
mean symmetry	0.438413	0.330499
mean fractal dimension	0.767297	0.012838
radius error	0.049559	0.567134
texture error	0.045655	0.008303
perimeter error	0.085433	0.556141
area error	0.017539	0.548236
smoothness error	0.101480	0.067016
compactness error	0.590973	0.292999
concavity error	0.439329	0.253730
concave points error	0.310655	0.408042
symmetry error	0.078079	0.006522
fractal dimension error	0.591328	0.077972
worst radius	0.093492	0.776454
worst texture	0.219122	0.456903

```
worst perimeter                  0.138957    0.782914
worst area                       0.079647    0.733825
worst smoothness                 0.617624    0.421465
worst compactness                0.810455    0.590998
worst concavity                  0.686511    0.659610
worst concave points             0.511114    0.793566
worst symmetry                   0.537848    0.416294
worst fractal dimension          1.000000    0.323872
malignancy                       0.323872    1.000000

[31 rows x 31 columns]
```

Видны очень сильные корреляции между признаками mean area, mean radius и mean perimeter, следовательно, в модели стоит использовать только какой-то один из них. Аналогично дела обстоят с тройкой worst area - worst radius - worst perimeter. Целевой параметр malignancy хорошо (коэффициент > 0.7) коррелирует с параметрами mean perimeter, mean concave points, worst perimeter, worst concave points. Их и следует выбирать для построения дальнейшей модели и решения задачи классификации.

Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков.

Корреляционная матрица симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой).

Описание метода corr - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html

По умолчанию при построении матрицы используется коэффициент корреляции Пирсона. Возможно также построить корреляционную матрицу на основе коэффициентов корреляции Кендалла и Спирмена. На практике три метода редко дают значимые различия.

[81]:
```
np.abs(data.corr(method='pearson'))
```

[81]:
```
                         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
malignancy                   0.730029        0.415185        0.742636    0.
↪708984
```

	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
malignancy	0.358560	0.596534	0.696360

	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
malignancy	0.776614	0.330499

	mean fractal dimension	…	worst texture	\
mean radius	0.311631	…	0.297008	
mean texture	0.076437	…	0.912045	
mean perimeter	0.261477	…	0.303038	
mean area	0.283110	…	0.287489	
mean smoothness	0.584792	…	0.036072	
mean compactness	0.565369	…	0.248133	
mean concavity	0.336783	…	0.299879	
mean concave points	0.166917	…	0.292752	
mean symmetry	0.479921	…	0.090651	
mean fractal dimension	1.000000	…	0.051269	
radius error	0.000111	…	0.194799	
texture error	0.164174	…	0.409003	
perimeter error	0.039830	…	0.200371	
area error	0.090170	…	0.196497	
smoothness error	0.401964	…	0.074743	
compactness error	0.559837	…	0.143003	
concavity error	0.446630	…	0.100241	
concave points error	0.341198	…	0.086741	
symmetry error	0.345007	…	0.077473	
fractal dimension error	0.688132	…	0.003195	
worst radius	0.253691	…	0.359921	
worst texture	0.051269	…	1.000000	
worst perimeter	0.205151	…	0.365098	
worst area	0.231854	…	0.345842	
worst smoothness	0.504942	…	0.225429	
worst compactness	0.458798	…	0.360832	
worst concavity	0.346234	…	0.368366	
worst concave points	0.175325	…	0.359755	
worst symmetry	0.334019	…	0.233027	
worst fractal dimension	0.767297	…	0.219122	
malignancy	0.012838	…	0.456903	

	worst perimeter	worst area	worst smoothness	\
mean radius	0.965137	0.941082	0.119616	
mean texture	0.358040	0.343546	0.077503	
mean perimeter	0.970387	0.941550	0.150549	
mean area	0.959120	0.959213	0.123523	
mean smoothness	0.238853	0.206718	0.805324	
mean compactness	0.590210	0.509604	0.565541	
mean concavity	0.729565	0.675987	0.448822	
mean concave points	0.855923	0.809630	0.452753	
mean symmetry	0.219169	0.177193	0.426675	
mean fractal dimension	0.205151	0.231854	0.504942	
radius error	0.719684	0.751548	0.141919	
texture error	0.102242	0.083195	0.073658	
perimeter error	0.721031	0.730713	0.130054	
area error	0.761213	0.811408	0.125389	
smoothness error	0.217304	0.182195	0.314457	
compactness error	0.260516	0.199371	0.227394	
concavity error	0.226680	0.188353	0.168481	
concave points error	0.394999	0.342271	0.215351	
symmetry error	0.103753	0.110343	0.012662	
fractal dimension error	0.001000	0.022736	0.170568	
worst radius	0.993708	0.984015	0.216574	
worst texture	0.365098	0.345842	0.225429	
worst perimeter	1.000000	0.977578	0.236775	
worst area	0.977578	1.000000	0.209145	
worst smoothness	0.236775	0.209145	1.000000	
worst compactness	0.529408	0.438296	0.568187	
worst concavity	0.618344	0.543331	0.518523	
worst concave points	0.816322	0.747419	0.547691	
worst symmetry	0.269493	0.209146	0.493838	
worst fractal dimension	0.138957	0.079647	0.617624	
malignancy	0.782914	0.733825	0.421465	

	worst compactness	worst concavity	\
mean radius	0.413463	0.526911	
mean texture	0.277830	0.301025	
mean perimeter	0.455774	0.563879	
mean area	0.390410	0.512606	
mean smoothness	0.472468	0.434926	
mean compactness	0.865809	0.816275	
mean concavity	0.754968	0.884103	
mean concave points	0.667454	0.752399	
mean symmetry	0.473200	0.433721	
mean fractal dimension	0.458798	0.346234	
radius error	0.287103	0.380585	
texture error	0.092439	0.068956	
perimeter error	0.341919	0.418899	
area error	0.283257	0.385100	
smoothness error	0.055558	0.058298	

compactness error	0.678780	0.639147
concavity error	0.484858	0.662564
concave points error	0.452888	0.549592
symmetry error	0.060255	0.037119
fractal dimension error	0.390159	0.379975
worst radius	0.475820	0.573975
worst texture	0.360832	0.368366
worst perimeter	0.529408	0.618344
worst area	0.438296	0.543331
worst smoothness	0.568187	0.518523
worst compactness	1.000000	0.892261
worst concavity	0.892261	1.000000
worst concave points	0.801080	0.855434
worst symmetry	0.614441	0.532520
worst fractal dimension	0.810455	0.686511
malignancy	0.590998	0.659610

	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	
malignancy	0.793566	0.416294	

```
                         worst fractal dimension  malignancy
mean radius                             0.007066    0.730029
mean texture                            0.119205    0.415185
mean perimeter                          0.051019    0.742636
mean area                               0.003738    0.708984
mean smoothness                         0.499316    0.358560
mean compactness                        0.687382    0.596534
mean concavity                          0.514930    0.696360
mean concave points                     0.368661    0.776614
mean symmetry                           0.438413    0.330499
mean fractal dimension                  0.767297    0.012838
radius error                            0.049559    0.567134
texture error                           0.045655    0.008303
perimeter error                         0.085433    0.556141
area error                              0.017539    0.548236
smoothness error                        0.101480    0.067016
compactness error                       0.590973    0.292999
concavity error                         0.439329    0.253730
concave points error                    0.310655    0.408042
symmetry error                          0.078079    0.006522
fractal dimension error                 0.591328    0.077972
worst radius                            0.093492    0.776454
worst texture                           0.219122    0.456903
worst perimeter                         0.138957    0.782914
worst area                              0.079647    0.733825
worst smoothness                        0.617624    0.421465
worst compactness                       0.810455    0.590998
worst concavity                         0.686511    0.659610
worst concave points                    0.511114    0.793566
worst symmetry                          0.537848    0.416294
worst fractal dimension                 1.000000    0.323872
malignancy                              0.323872    1.000000

[31 rows x 31 columns]
```

[82]:
```
np.abs(data.corr(method='kendall'))
```

[82]:
```
                         mean radius  mean texture  mean perimeter  mean␣
 ↪area  \
mean radius                 1.000000      0.229159        0.963320   0.
 ↪985565
mean texture                0.229159      1.000000        0.234353   0.
 ↪230829
mean perimeter              0.963320      0.234353        1.000000   0.
 ↪956965
mean area                   0.985565      0.230829        0.956965   1.
 ↪000000
mean smoothness             0.099549      0.017135        0.122434   0.
 ↪092541
```

```
mean compactness             0.340020      0.184220      0.375540   0.
↪333534
mean concavity               0.465087      0.236079      0.497587   0.
↪461843
mean concave points          0.566917      0.209629      0.596716   0.
↪561760
mean symmetry                0.081502      0.075493      0.101781   0.
↪076963
mean fractal dimension       0.246220      0.039255      0.214661   0.
↪252131
radius error                 0.384712      0.247098      0.393716   0.
↪387101
texture error                0.095861      0.309294      0.091170   0.
↪094179
perimeter error              0.398999      0.264142      0.414246   0.
↪400951
area error                   0.549079      0.270659      0.557127   0.
↪552309
smoothness error             0.226065      0.024484      0.215165   0.
↪226658
compactness error            0.181153      0.180535      0.211562   0.
↪177378
concavity error              0.255945      0.197820      0.283680   0.
↪254040
concave points error         0.286742      0.162570      0.310856   0.
↪283280
symmetry error               0.168407      0.007114      0.158856   0.
↪169859
fractal dimension error      0.005832      0.099663      0.021156   0.
↪008733
worst radius                 0.882063      0.246993      0.886941   0.
↪883811
worst texture                0.210302      0.741293      0.216032   0.
↪212051
worst perimeter              0.862763      0.254406      0.881245   0.
↪863409
worst area                   0.882035      0.248096      0.885246   0.
↪885906
worst smoothness             0.082284      0.069336      0.102428   0.
↪077793
worst compactness            0.332970      0.199777      0.365026   0.
↪328257
worst concavity              0.421100      0.235242      0.449549   0.
↪418822
worst concave points         0.532569      0.217743      0.559819   0.
↪528375
worst symmetry               0.114141      0.081482      0.130283   0.
↪112032
```

```
worst fractal dimension     0.024103      0.078289        0.053103   0.
 ↪020265
malignancy                  0.599082      0.377644        0.611775   0.
 ↪599992

                         mean smoothness  mean compactness  mean concavity␣
 ↪ \
mean radius                     0.099549          0.340020        0.465087
mean texture                    0.017135          0.184220        0.236079
mean perimeter                  0.122434          0.375540        0.497587
mean area                       0.092541          0.333534        0.461843
mean smoothness                 1.000000          0.491072        0.364105
mean compactness                0.491072          1.000000        0.719194
mean concavity                  0.364105          0.719194        1.000000
mean concave points             0.398511          0.653022        0.775266
mean symmetry                   0.381515          0.392366        0.311354
mean fractal dimension          0.417070          0.345923        0.173476
radius error                    0.223561          0.351001        0.405383
texture error                   0.061806          0.032274        0.035346
perimeter error                 0.220743          0.409292        0.459626
area error                      0.196414          0.372483        0.460080
smoothness error                0.230123          0.084992        0.048788
compactness error               0.274218          0.619218        0.559088
concavity error                 0.247024          0.570663        0.678378
concave points error            0.305479          0.537138        0.581517
symmetry error                  0.101045          0.066251        0.015244
fractal dimension error         0.286167          0.437206        0.351399
worst radius                    0.136219          0.372046        0.492914
worst texture                   0.041361          0.174449        0.227797
worst perimeter                 0.152343          0.412628        0.531750
worst area                      0.128881          0.363980        0.488281
worst smoothness                0.608475          0.412274        0.338902
worst compactness               0.335814          0.724842        0.654070
worst concavity                 0.300847          0.647942        0.787870
worst concave points            0.349195          0.633090        0.734047
worst symmetry                  0.270602          0.312991        0.263490
worst fractal dimension         0.357792          0.498231        0.377455
malignancy                      0.304033          0.497971        0.599449

                         mean concave points  mean symmetry  \
mean radius                         0.566917       0.081502
mean texture                        0.209629       0.075493
mean perimeter                      0.596716       0.101781
mean area                           0.561760       0.076963
mean smoothness                     0.398511       0.381515
mean compactness                    0.653022       0.392366
mean concavity                      0.775266       0.311354
mean concave points                 1.000000       0.291970
mean symmetry                       0.291970       1.000000
```

```
mean fractal dimension             0.094885       0.297681
radius error                       0.454104       0.225514
texture error                      0.005238       0.095181
perimeter error                    0.490678       0.237409
area error                         0.532389       0.191490
smoothness error                   0.012047       0.136064
compactness error                  0.427162       0.307858
concavity error                    0.497183       0.256674
concave points error               0.568956       0.263132
symmetry error                     0.019819       0.262925
fractal dimension error            0.254332       0.281801
worst radius                       0.590400       0.111969
worst texture                      0.200509       0.081521
worst perimeter                    0.620728       0.129378
worst area                         0.583169       0.104677
worst smoothness                   0.339906       0.291003
worst compactness                  0.557303       0.308333
worst concavity                    0.635728       0.275028
worst concave points               0.783933       0.276077
worst symmetry                     0.240447       0.520061
worst fractal dimension            0.285028       0.281841
malignancy                         0.635873       0.271924

                        mean fractal dimension  …  worst texture  \
mean radius                           0.246220  …       0.210302
mean texture                          0.039255  …       0.741293
mean perimeter                        0.214661  …       0.216032
mean area                             0.252131  …       0.212051
mean smoothness                       0.417070  …       0.041361
mean compactness                      0.345923  …       0.174449
mean concavity                        0.173476  …       0.227797
mean concave points                   0.094885  …       0.200509
mean symmetry                         0.297681  …       0.081521
mean fractal dimension                1.000000  …       0.031406
radius error                          0.002996  …       0.191005
texture error                         0.105633  …       0.348027
perimeter error                       0.038443  …       0.205631
area error                            0.081134  …       0.221482
smoothness error                      0.276804  …       0.024264
compactness error                     0.333806  …       0.142983
concavity error                       0.232522  …       0.158908
concave points error                  0.196318  …       0.105170
symmetry error                        0.214262  …       0.070404
fractal dimension error               0.498389  …       0.055635
worst radius                          0.208569  …       0.250127
worst texture                         0.031406  …       1.000000
worst perimeter                       0.174835  …       0.257984
worst area                            0.215901  …       0.251341
worst smoothness                      0.348018  …       0.148002
worst compactness                     0.274333  …       0.234597
```

worst concavity	0.162783	…	0.266734
worst concave points	0.092549	…	0.247731
worst symmetry	0.204803	…	0.154258
worst fractal dimension	0.566543	…	0.130159
malignancy	0.021173	…	0.389654

	worst perimeter	worst area	worst smoothness	\
mean radius	0.862763	0.882035	0.082284	
mean texture	0.254406	0.248096	0.069336	
mean perimeter	0.881245	0.885246	0.102428	
mean area	0.863409	0.885906	0.077793	
mean smoothness	0.152343	0.128881	0.608475	
mean compactness	0.412628	0.363980	0.412274	
mean concavity	0.531750	0.488281	0.338902	
mean concave points	0.620728	0.583169	0.339906	
mean symmetry	0.129378	0.104677	0.291003	
mean fractal dimension	0.174835	0.215901	0.348018	
radius error	0.425010	0.427399	0.135621	
texture error	0.094206	0.096678	0.016133	
perimeter error	0.455179	0.434218	0.131253	
area error	0.585220	0.592259	0.123439	
smoothness error	0.211233	0.223126	0.255487	
compactness error	0.237664	0.190570	0.224289	
concavity error	0.305778	0.265934	0.213257	
concave points error	0.316127	0.282978	0.204449	
symmetry error	0.171452	0.184874	0.028584	
fractal dimension error	0.042561	0.006084	0.216414	
worst radius	0.939026	0.979299	0.143146	
worst texture	0.257984	0.251341	0.148002	
worst perimeter	1.000000	0.931692	0.158814	
worst area	0.931692	1.000000	0.137600	
worst smoothness	0.158814	0.137600	1.000000	
worst compactness	0.426479	0.375985	0.396455	
worst concavity	0.507132	0.463746	0.364315	
worst concave points	0.615378	0.575373	0.382131	
worst symmetry	0.186325	0.163272	0.346625	
worst fractal dimension	0.113309	0.072342	0.444001	
malignancy	0.650879	0.643117	0.347952	

	worst compactness	worst concavity	\
mean radius	0.332970	0.421100	
mean texture	0.199777	0.235242	
mean perimeter	0.365026	0.449549	
mean area	0.328257	0.418822	
mean smoothness	0.335814	0.300847	
mean compactness	0.724842	0.647942	
mean concavity	0.654070	0.787870	
mean concave points	0.557303	0.635728	
mean symmetry	0.308333	0.275028	
mean fractal dimension	0.274333	0.162783	

radius error	0.226239	0.269353
texture error	0.060713	0.046471
perimeter error	0.295248	0.329696
area error	0.276771	0.338184
smoothness error	0.031356	0.039025
compactness error	0.597656	0.511201
concavity error	0.534558	0.629947
concave points error	0.411078	0.447413
symmetry error	0.055109	0.081353
fractal dimension error	0.365405	0.295379
worst radius	0.382581	0.468060
worst texture	0.234597	0.266734
worst perimeter	0.426479	0.507132
worst area	0.375985	0.463746
worst smoothness	0.396455	0.364315
worst compactness	1.000000	0.744722
worst concavity	0.744722	1.000000
worst concave points	0.648276	0.729468
worst symmetry	0.373628	0.332061
worst fractal dimension	0.569104	0.445621
malignancy	0.495958	0.576907

	worst concave points	worst symmetry	\
mean radius	0.532569	0.114141	
mean texture	0.217743	0.081482	
mean perimeter	0.559819	0.130283	
mean area	0.528375	0.112032	
mean smoothness	0.349195	0.270602	
mean compactness	0.633090	0.312991	
mean concavity	0.734047	0.263490	
mean concave points	0.783933	0.240447	
mean symmetry	0.276077	0.520061	
mean fractal dimension	0.092549	0.204803	
radius error	0.350102	0.097869	
texture error	0.065526	0.080052	
perimeter error	0.396463	0.110611	
area error	0.434106	0.099148	
smoothness error	0.048716	0.046370	
compactness error	0.415710	0.185053	
concavity error	0.482037	0.159541	
concave points error	0.513067	0.091843	
symmetry error	0.097753	0.196914	
fractal dimension error	0.222986	0.120269	
worst radius	0.581723	0.169804	
worst texture	0.247731	0.154258	
worst perimeter	0.615378	0.186325	
worst area	0.575373	0.163272	
worst smoothness	0.382131	0.346625	
worst compactness	0.648276	0.373628	
worst concavity	0.729468	0.332061	

```
worst concave points                1.000000        0.319261
worst symmetry                      0.319261        1.000000
worst fractal dimension             0.357549        0.343083
malignancy                          0.639090        0.324380

                         worst fractal dimension  malignancy
mean radius                             0.024103    0.599082
mean texture                            0.078289    0.377644
mean perimeter                          0.053103    0.611775
mean area                               0.020265    0.599992
mean smoothness                         0.357792    0.304033
mean compactness                        0.498231    0.497971
mean concavity                          0.377455    0.599449
mean concave points                     0.285028    0.635873
mean symmetry                           0.281841    0.271924
mean fractal dimension                  0.566543    0.021173
radius error                            0.073594    0.504197
texture error                           0.031352    0.015872
perimeter error                         0.122593    0.515243
area error                              0.057602    0.583722
smoothness error                        0.087709    0.042656
compactness error                       0.430792    0.311115
concavity error                         0.351748    0.384486
concave points error                    0.244971    0.399546
symmetry error                          0.007769    0.075451
fractal dimension error                 0.528457    0.164674
worst radius                            0.078108    0.644195
worst texture                           0.130159    0.389654
worst perimeter                         0.113309    0.650879
worst area                              0.072342    0.643117
worst smoothness                        0.444001    0.347952
worst compactness                       0.569104    0.495958
worst concavity                         0.445621    0.576907
worst concave points                    0.357549    0.639090
worst symmetry                          0.343083    0.324380
worst fractal dimension                 1.000000    0.254571
malignancy                              0.254571    1.000000

[31 rows x 31 columns]
```

[83]:
```
np.abs(data.corr(method='spearman'))
```

[83]:
```
                         mean radius  mean texture  mean perimeter  mean␣
 ↪area  \
mean radius                 1.000000      0.340956        0.997802   0.
 ↪999602
mean texture                0.340956      1.000000        0.348142   0.
 ↪344145
mean perimeter              0.997802      0.348142        1.000000   0.
 ↪997068
```

```
mean area                0.999602   0.344145   0.997068   1.
↪000000
mean smoothness          0.148510   0.024649   0.182923   0.
↪138053
mean compactness         0.497578   0.266499   0.543925   0.
↪488988
mean concavity           0.645728   0.342646   0.681958   0.
↪642557
mean concave points      0.759702   0.306891   0.788629   0.
↪755165
mean symmetry            0.120242   0.110130   0.150049   0.
↪113928
mean fractal dimension   0.349931   0.059303   0.304891   0.
↪358425
radius error             0.550247   0.363621   0.560326   0.
↪553388
texture error            0.144499   0.450720   0.137578   0.
↪142469
perimeter error          0.565520   0.386813   0.582789   0.
↪568237
area error               0.738077   0.395139   0.745824   0.
↪741518
smoothness error         0.326385   0.037048   0.311147   0.
↪327431
compactness error        0.264904   0.263591   0.308620   0.
↪260362
concavity error          0.364555   0.287188   0.402277   0.
↪362308
concave points error     0.410576   0.238610   0.441996   0.
↪406468
symmetry error           0.241376   0.008945   0.228187   0.
↪243507
fractal dimension error  0.008411   0.147605   0.032429   0.
↪012688
worst radius             0.978604   0.366547   0.981244   0.
↪979258
worst texture            0.314911   0.909218   0.323109   0.
↪318178
worst perimeter          0.971555   0.375273   0.978980   0.
↪971822
worst area               0.978863   0.368335   0.980864   0.
↪980264
worst smoothness         0.125789   0.101401   0.156611   0.
↪119712
worst compactness        0.491357   0.290917   0.534565   0.
↪485813
worst concavity          0.596043   0.339725   0.632106   0.
↪593736
```

```
worst concave points         0.727265      0.319235        0.757526   0.
↪723390
worst symmetry               0.174698      0.120693        0.199007   0.
↪170860
worst fractal dimension      0.044564      0.116144        0.088961   0.
↪038758
malignancy                   0.732785      0.461971        0.748496   0.
↪734122
```

	mean smoothness	mean compactness	mean concavity
mean radius	0.148510	0.497578	0.645728
mean texture	0.024649	0.266499	0.342646
mean perimeter	0.182923	0.543925	0.681958
mean area	0.138053	0.488988	0.642557
mean smoothness	1.000000	0.678806	0.518511
mean compactness	0.678806	1.000000	0.896518
mean concavity	0.518511	0.896518	1.000000
mean concave points	0.565172	0.848295	0.927352
mean symmetry	0.542228	0.552203	0.446793
mean fractal dimension	0.588465	0.499195	0.258174
radius error	0.334282	0.506582	0.575277
texture error	0.091283	0.047766	0.051318
perimeter error	0.331360	0.583520	0.646199
area error	0.296059	0.539511	0.644344
smoothness error	0.338692	0.127381	0.070321
compactness error	0.392455	0.817875	0.761230
concavity error	0.354730	0.772283	0.858306
concave points error	0.438826	0.732425	0.774656
symmetry error	0.150740	0.098388	0.022753
fractal dimension error	0.413429	0.621121	0.513593
worst radius	0.203453	0.542626	0.682316
worst texture	0.060645	0.255305	0.335866
worst perimeter	0.226345	0.592254	0.722424
worst area	0.191735	0.531590	0.676628
worst smoothness	0.796085	0.578902	0.488775
worst compactness	0.481384	0.901029	0.849985
worst concavity	0.429107	0.837921	0.938543
worst concave points	0.498868	0.825473	0.904938
worst symmetry	0.393579	0.450333	0.383667
worst fractal dimension	0.511457	0.688986	0.541838
malignancy	0.371892	0.609288	0.733308

	mean concave points	mean symmetry
mean radius	0.759702	0.120242
mean texture	0.306891	0.110130
mean perimeter	0.788629	0.150049
mean area	0.755165	0.113928
mean smoothness	0.565172	0.542228

mean compactness	0.848295	0.552203
mean concavity	0.927352	0.446793
mean concave points	1.000000	0.423767
mean symmetry	0.423767	1.000000
mean fractal dimension	0.142659	0.428467
radius error	0.635054	0.337912
texture error	0.008710	0.139124
perimeter error	0.679841	0.354888
area error	0.726982	0.288322
smoothness error	0.016798	0.206106
compactness error	0.608388	0.435714
concavity error	0.674668	0.367637
concave points error	0.758438	0.382736
symmetry error	0.028353	0.384123
fractal dimension error	0.378374	0.402630
worst radius	0.787411	0.164552
worst texture	0.300562	0.118890
worst perimeter	0.813960	0.190526
worst area	0.780395	0.154462
worst smoothness	0.490035	0.424230
worst compactness	0.758309	0.440828
worst concavity	0.827281	0.394481
worst concave points	0.937075	0.397477
worst symmetry	0.355477	0.710359
worst fractal dimension	0.421110	0.410069
malignancy	0.777877	0.332567

	mean fractal dimension	…	worst texture	\
mean radius	0.349931	…	0.314911	
mean texture	0.059303	…	0.909218	
mean perimeter	0.304891	…	0.323109	
mean area	0.358425	…	0.318178	
mean smoothness	0.588465	…	0.060645	
mean compactness	0.499195	…	0.255305	
mean concavity	0.258174	…	0.335866	
mean concave points	0.142659	…	0.300562	
mean symmetry	0.428467	…	0.118890	
mean fractal dimension	1.000000	…	0.047791	
radius error	0.001477	…	0.283581	
texture error	0.157103	…	0.496551	
perimeter error	0.055309	…	0.302553	
area error	0.120333	…	0.327857	
smoothness error	0.401530	…	0.036290	
compactness error	0.481139	…	0.209979	
concavity error	0.344007	…	0.235945	
concave points error	0.286393	…	0.157304	
symmetry error	0.314165	…	0.104702	
fractal dimension error	0.683800	…	0.083174	
worst radius	0.294540	…	0.371230	
worst texture	0.047791	…	1.000000	

worst perimeter	0.247456	…	0.381022
worst area	0.304927	…	0.372376
worst smoothness	0.493474	…	0.217799
worst compactness	0.403653	…	0.342319
worst concavity	0.242611	…	0.387009
worst concave points	0.139152	…	0.365309
worst symmetry	0.295046	…	0.226816
worst fractal dimension	0.760771	…	0.193191
malignancy	0.025903	…	0.476720

	worst perimeter	worst area	worst smoothness	\
mean radius	0.971555	0.978863	0.125789	
mean texture	0.375273	0.368335	0.101401	
mean perimeter	0.978980	0.980864	0.156611	
mean area	0.971822	0.980264	0.119712	
mean smoothness	0.226345	0.191735	0.796085	
mean compactness	0.592254	0.531590	0.578902	
mean concavity	0.722424	0.676628	0.488775	
mean concave points	0.813960	0.780395	0.490035	
mean symmetry	0.190526	0.154462	0.424230	
mean fractal dimension	0.247456	0.304927	0.493474	
radius error	0.592509	0.595732	0.203760	
texture error	0.142855	0.147786	0.023095	
perimeter error	0.626896	0.605163	0.197899	
area error	0.768336	0.775662	0.188777	
smoothness error	0.308749	0.323724	0.372247	
compactness error	0.344865	0.278844	0.320500	
concavity error	0.432895	0.377836	0.305368	
concave points error	0.448363	0.403587	0.294840	
symmetry error	0.246712	0.266705	0.042873	
fractal dimension error	0.063012	0.007312	0.312293	
worst radius	0.993548	0.998891	0.218616	
worst texture	0.381022	0.372376	0.217799	
worst perimeter	1.000000	0.992433	0.241172	
worst area	0.992433	1.000000	0.210063	
worst smoothness	0.241172	0.210063	1.000000	
worst compactness	0.613070	0.550007	0.560156	
worst concavity	0.700572	0.651120	0.519490	
worst concave points	0.812983	0.773945	0.543982	
worst symmetry	0.281383	0.248358	0.501230	
worst fractal dimension	0.179003	0.118734	0.614796	
malignancy	0.796319	0.786902	0.425513	

	worst compactness	worst concavity	\
mean radius	0.491357	0.596043	
mean texture	0.290917	0.339725	
mean perimeter	0.534565	0.632106	
mean area	0.485813	0.593736	
mean smoothness	0.481384	0.429107	
mean compactness	0.901029	0.837921	

mean concavity	0.849985	0.938543
mean concave points	0.758309	0.827281
mean symmetry	0.440828	0.394481
mean fractal dimension	0.403653	0.242611
radius error	0.339725	0.404431
texture error	0.090069	0.070625
perimeter error	0.438416	0.490340
area error	0.413658	0.500307
smoothness error	0.049245	0.063848
compactness error	0.789431	0.701251
concavity error	0.731517	0.811327
concave points error	0.585929	0.624555
symmetry error	0.082054	0.118144
fractal dimension error	0.526899	0.431677
worst radius	0.558316	0.655942
worst texture	0.342319	0.387009
worst perimeter	0.613070	0.700572
worst area	0.550007	0.651120
worst smoothness	0.560156	0.519490
worst compactness	1.000000	0.914894
worst concavity	0.914894	1.000000
worst concave points	0.844454	0.902301
worst symmetry	0.527102	0.476179
worst fractal dimension	0.762247	0.623128
malignancy	0.606810	0.705734

	worst concave points	worst symmetry	\
mean radius	0.727265	0.174698	
mean texture	0.319235	0.120693	
mean perimeter	0.757526	0.199007	
mean area	0.723390	0.170860	
mean smoothness	0.498868	0.393579	
mean compactness	0.825473	0.450333	
mean concavity	0.904938	0.383667	
mean concave points	0.937075	0.355477	
mean symmetry	0.397477	0.710359	
mean fractal dimension	0.139152	0.295046	
radius error	0.508662	0.147213	
texture error	0.097025	0.119890	
perimeter error	0.569428	0.166606	
area error	0.619539	0.154415	
smoothness error	0.076503	0.067149	
compactness error	0.587471	0.265987	
concavity error	0.656814	0.230690	
concave points error	0.692071	0.132782	
symmetry error	0.140795	0.283201	
fractal dimension error	0.331206	0.172926	
worst radius	0.780632	0.257165	
worst texture	0.365309	0.226816	
worst perimeter	0.812983	0.281383	

```
worst area                          0.773945        0.248358
worst smoothness                    0.543982        0.501230
worst compactness                   0.844454        0.527102
worst concavity                     0.902301        0.476179
worst concave points                1.000000        0.460711
worst symmetry                      0.460711        1.000000
worst fractal dimension             0.516664        0.488439
malignancy                          0.781674        0.396843

                        worst fractal dimension  malignancy
mean radius                            0.044564    0.732785
mean texture                           0.116144    0.461971
mean perimeter                         0.088961    0.748496
mean area                              0.038758    0.734122
mean smoothness                        0.511457    0.371892
mean compactness                       0.688986    0.609288
mean concavity                         0.541838    0.733308
mean concave points                    0.421110    0.777877
mean symmetry                          0.410069    0.332567
mean fractal dimension                 0.760771    0.025903
radius error                           0.111043    0.616912
texture error                          0.048143    0.019419
perimeter error                        0.185534    0.630411
area error                             0.091670    0.714184
smoothness error                       0.129752    0.052193
compactness error                      0.604844    0.380666
concavity error                        0.505962    0.470338
concave points error                   0.357090    0.488717
symmetry error                         0.011133    0.092303
fractal dimension error                0.712771    0.201492
worst radius                           0.127449    0.787933
worst texture                          0.193191    0.476720
worst perimeter                        0.179003    0.796319
worst area                             0.118734    0.786902
worst smoothness                       0.614796    0.425513
worst compactness                      0.762247    0.606810
worst concavity                        0.623128    0.705734
worst concave points                   0.516664    0.781674
worst symmetry                         0.488439    0.396843
worst fractal dimension                1.000000    0.311477
malignancy                             0.311477    1.000000

[31 rows x 31 columns]
```

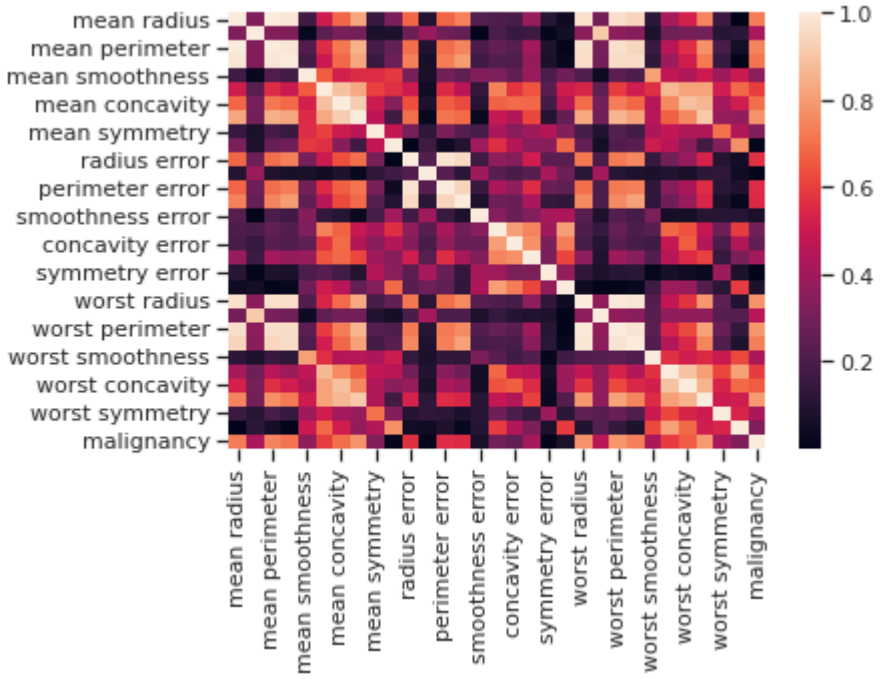
В случае большого количества признаков анализ числовой корреляционной матрицы становится неудобен.

Для визуализации корреляционной матрицы будем использовать “тепловую карту” heatmap которая показывает степень корреляции различными цветами.

Используем метод heatmap библиотеки seaborn - https://seaborn.pydata.org/generated/seaborn.heatmap.html

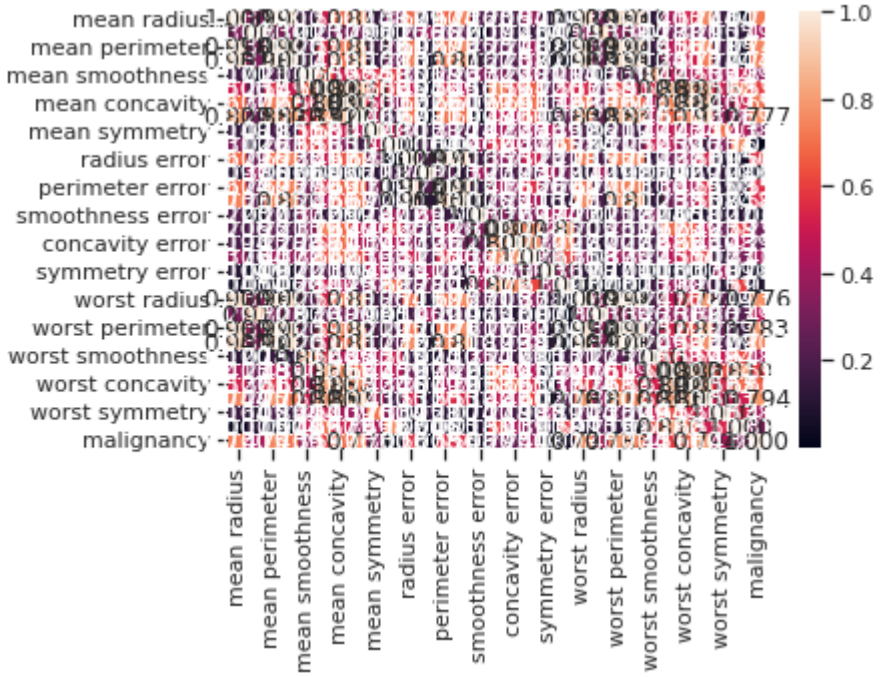
[84]:
```
sns.heatmap(np.abs(data.corr()))
```

[84]: <AxesSubplot:>



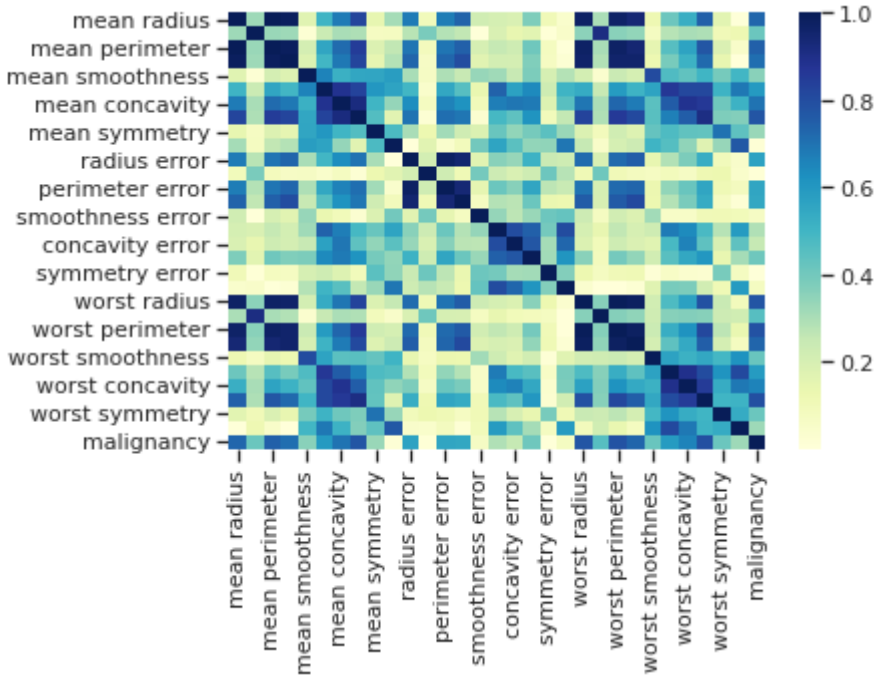
[85]:
```
# Вывод значений в ячейках
sns.heatmap(np.abs(data.corr()), annot=True, fmt='.3f')
```

[85]: <AxesSubplot:>



[86]:
```
# Изменение цветовой гаммы
sns.heatmap(np.abs(data.corr()), cmap='YlGnBu')
```

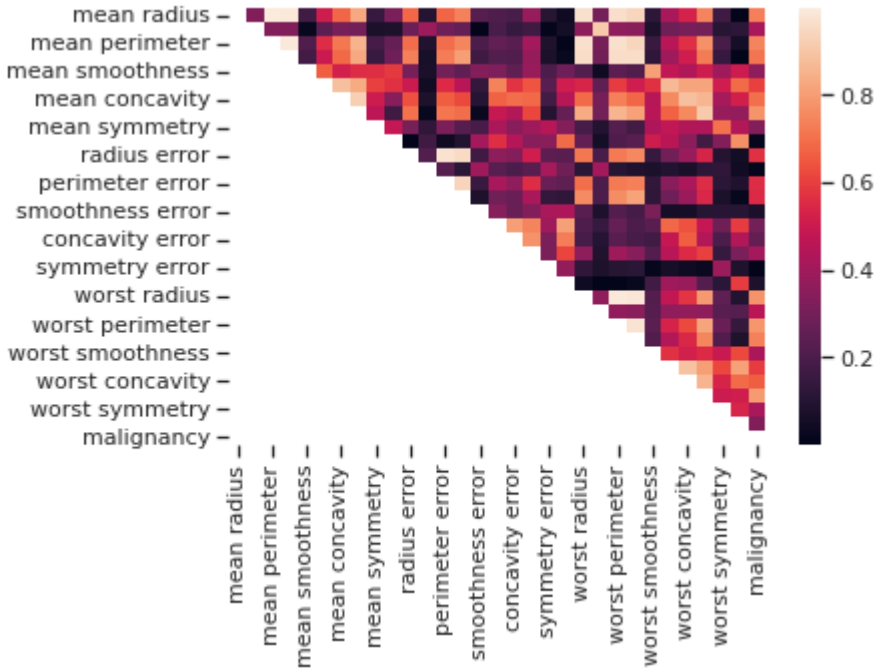
[86]: <AxesSubplot:>



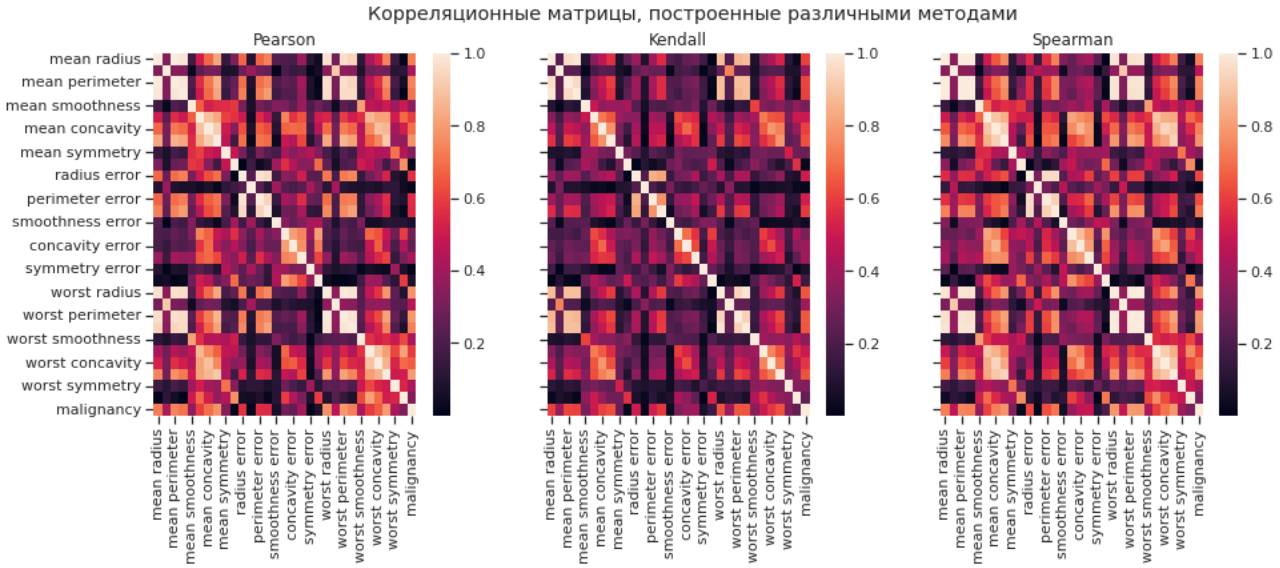
[87]:
```
# Треугольный вариант матрицы
mask = np.zeros_like(np.abs(data.corr()), dtype=np.bool)
# чтобы оставить нижнюю часть матрицы
# mask[np.triu_indices_from(mask)] = True
# чтобы оставить верхнюю часть матрицы
mask[np.tril_indices_from(mask)] = True
sns.heatmap(np.abs(data.corr()), mask=mask)
```

```
/tmp/ipykernel_7045/1998021473.py:2: DeprecationWarning: `np.bool` is a
deprecated alias for the builtin `bool`. To silence this warning, use␣
 ↪`bool` by
itself. Doing this will not modify any behavior and is safe. If you␣
 ↪specifically
wanted the numpy scalar type, use `np.bool_` here.
Deprecated in NumPy 1.20; for more details and guidance:
https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  mask = np.zeros_like(np.abs(data.corr()), dtype=np.bool)
```

[87]: <AxesSubplot:>



[88]:
```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))
sns.heatmap(np.abs(data.corr(method='pearson')), ax=ax[0])
sns.heatmap(np.abs(data.corr(method='kendall')), ax=ax[1])
sns.heatmap(np.abs(data.corr(method='spearman')), ax=ax[2])
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```



Необходимо отметить, что тепловая карта не очень хорошо подходит для определения корреляции нецелевых признаков между собой.

В примере тепловая карта помогает определить значимую корреляцию между признаками mean radius, mean perimeter и mean area, следовательно только один из этих признаков нужно включать в модель.

Но в реальной модели могут быть сотни признаков и коррелирующие признаки могут образовывать группы, состояшие более чем из двух признаков. Увидеть такие группы с помощью тепловой карты сложно.

Для решения задачи предлагается новый вариант визуализации - “Солнечная корреляционная карта” Solar correlation map.

К сожалению, данная библиотека пока работает только через файловый интерфейс и не предназначена для встраивания в ноутбук.

Примеры статей с описанием работы библиотеки: - https://www.oreilly.com/learning/a-new-visualization-to-beautifully-explore-correlations - https://www.mtab.com/the-puzzle-of-visualizing-correlations/

7. Дополнительные ссылки на обучающие ноутбуки

The Best Tutorial for Beginners (Kaggle)