Набор данных для прогнозирования сердечной недостаточности

Выполнила:
Синица И.И

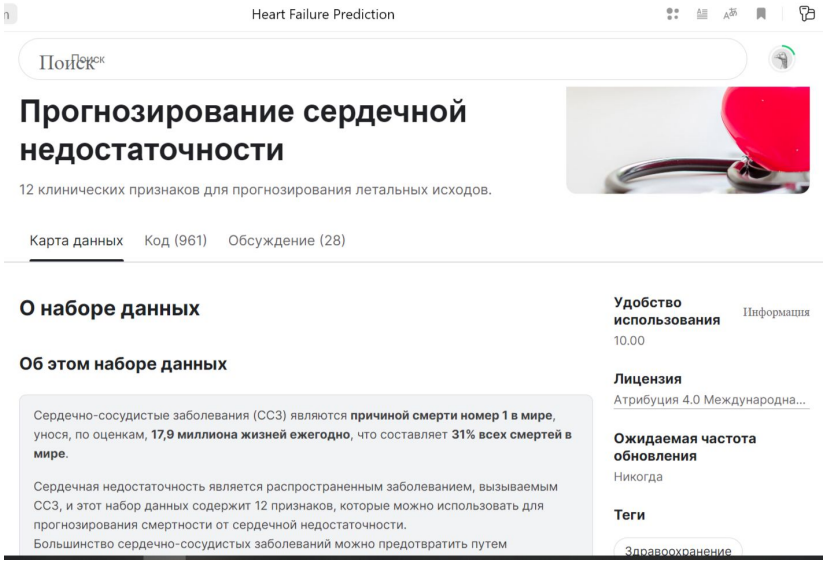
Описание источника данных

Набор данных размещен по адресу:

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data

Сердечно-сосудистые заболевания (ССЗ) являются причиной смерти номер 1 в мире, унося, по оценкам, 17,9 миллиона жизней ежегодно, что составляет 31% всех смертей в мире.

Сердечная недостаточность является распространенным заболеванием, вызываемым ССЗ, и этот набор данных содержит 12 признаков, которые можно использовать для прогнозирования смертности от сердечной недостаточности.
Большинство сердечно-сосудистых заболеваний можно предотвратить путем устранения поведенческих факторов риска, таких как употребление табака, нездоровое питание и ожирение, отсутствие физической активности и вредное употребление алкоголя, с использованием общегосударственных стратегий.



Описание колонок

1.Age : возраст пациента в годах;
2.Anaemia : болеет ли пациент анемией;
3.Creatinine_phosphokinase: уровень фермента в креатининфосфокиназа [ед/литр];
4.Diabetes : есть ли у пациента заболевание диабет;
5.Ejection_fraction : фракция выброса крови [%];
6.High_blood_pressure : страдает ли повышенным давлением пациент [да/нет];

Описание колонок

7. Platelets : уровень тромбоцитов [тыс/мкл];
8. Serum_creatinine : уровень сывороточного креатинина [мкмоль/л];
9. Serum_sodium : уровень натрия в сыворотка [Ммоль/л];
10. Sex : пол пациента;
11. Smoking : курение [да /нет];
12. Time : период наблюдения ;
13. DEATH_EVENT : наступила ли смерть за период наблюдения за пациентом.

Датасет состоит из 13 столбцов, в нем нет недостающих или ложных данных, все признаки можно разделить на числовые и категориальные.

Анализ данных

Поскольку выбранный набор данных имеет много числовых функций, начинаю анализ с просмотра агрегированных показателей.

```
df.describe()
df.describe().round(2)
```

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.00	299.00	299.00	299.00	299.00	299.00	299.00
mean	60.83	581.84	38.08	263358.03	1.39	136.63	130.26
std	11.89	970.29	11.83	97804.24	1.03	4.41	77.61
min	40.00	23.00	14.00	25100.00	0.50	113.00	4.00
25%	51.00	116.50	30.00	212500.00	0.90	134.00	73.00
50%	60.00	250.00	38.00	262000.00	1.10	137.00	115.00
75%	70.00	582.00	45.00	303500.00	1.40	140.00	203.00
max	95.00	7861.00	80.00	850000.00	9.40	148.00	285.00

Делаю вывод, что средний возраст пациентов составляет 60 л, время наблюдения от 4 до 285 дней, в среднем 130 дней.

Тромбоциты (кровяные пластинки) — элементы крови, которые отвечают за свёртывание. Они помогают остановить кровотечение и защищают от массивной кровопотери. У здорового человека их количество варьируется от 150 000 до 400 000 кило тромбоцитов / мл крови, в датасете у 75% пациентов находится в данном диапазоне

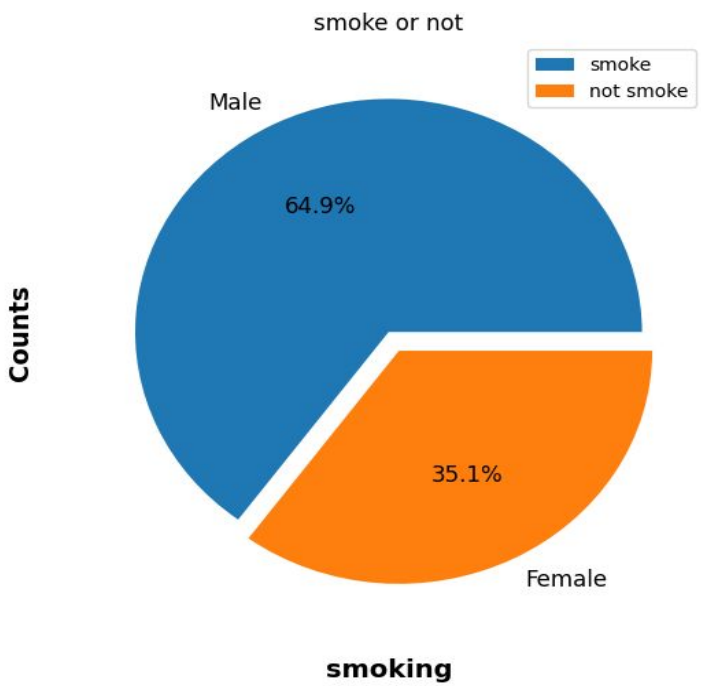
Натрий в сыворотке: уровень натрия в сыворотке крови в норме должен составлять от 135 до 145 ммоль/л. Все, что ниже — гипонатриемия , выше — гипернатриемия (типична для людей с сердечной недостаточностью). При избытке натрия человек испытывает сильную жажду — организм стремится восполнить нехватку воды и «разбавить» чрезмерно концентрированный раствор. В тяжелых случаях возможны неврологические симптомы, выражающиеся в спутанности сознания, появление судорог, больной может впасть в кому. В данном датасете среднее значение выше нормы

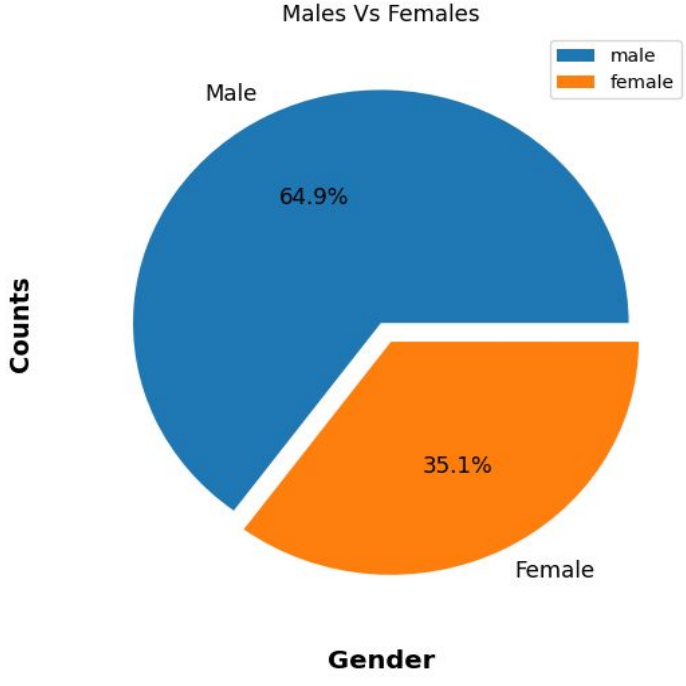
Креатининфосфокиназа: это фермент, который присутствует в крови и помогает восстанавливать поврежденные ткани. Высокий уровень КФК означает сердечную недостаточность или травму. Нормальный уровень у мужчин составляет 55–170 мкг / л, а у женщин - 30–135 мкг / л .

В представленном наборе данных, поскольку у всех пациентов была сердечная недостаточность, среднее значение (581.84) выше нормы.

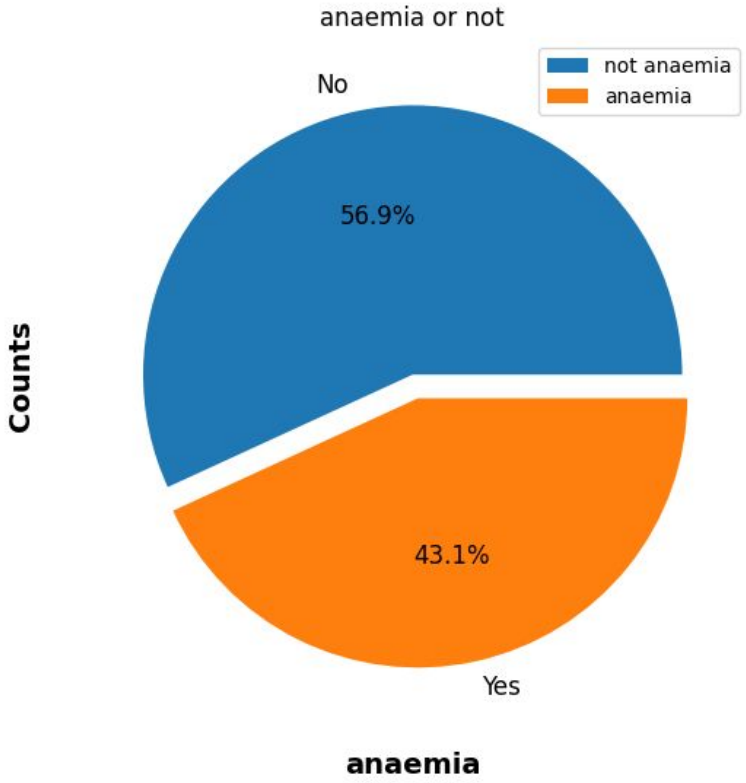
Анализ категориальных признаков

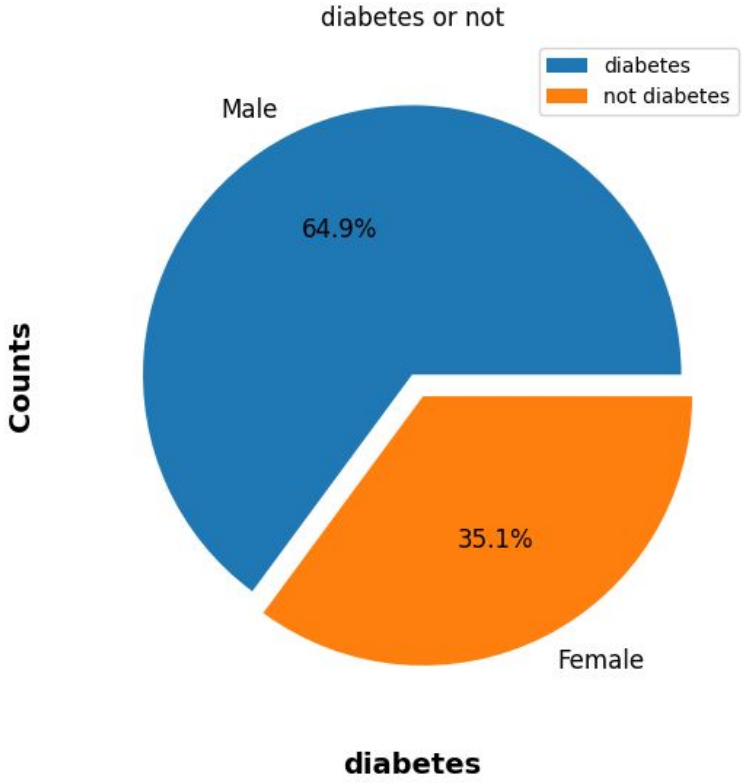
В ходе исследования, можно наглядно представить, в каком отношении идет разделение внутри категориальных признаков:





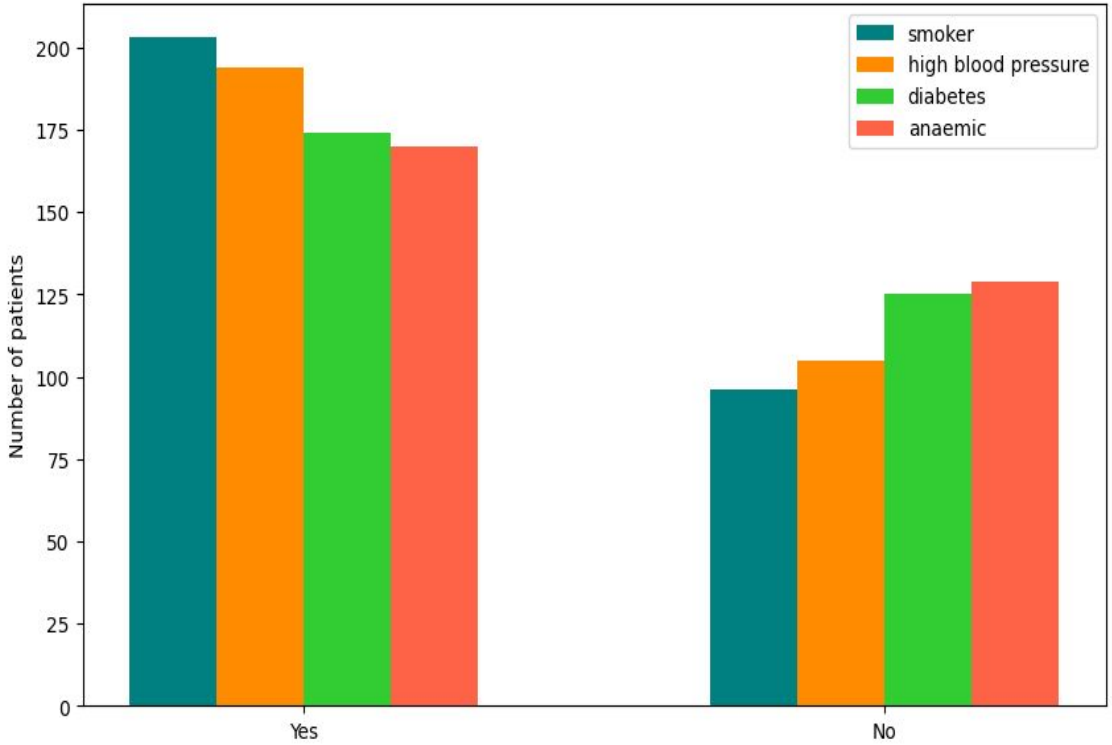
Анализ категориальных признаков



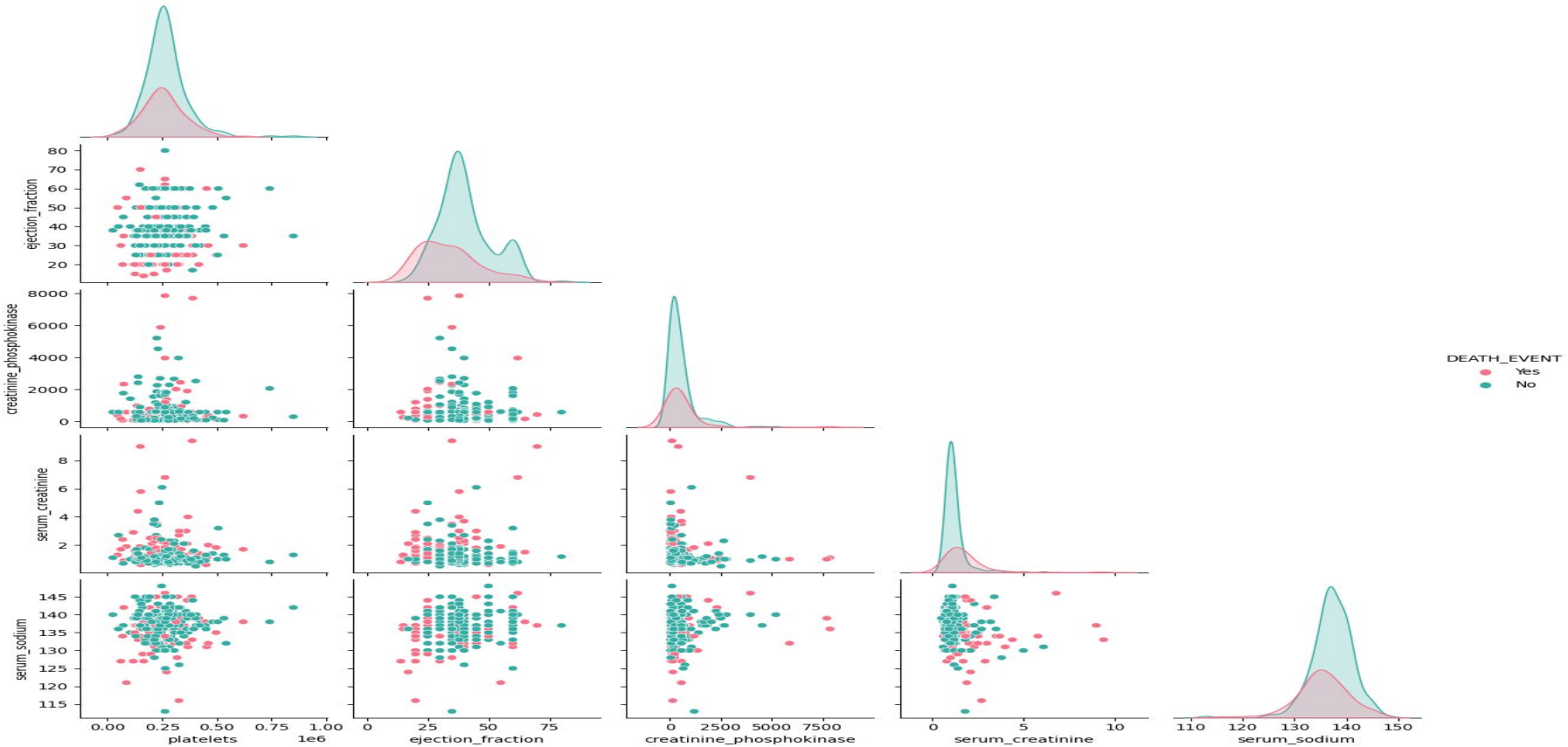


Анализ категориальных признаков

Количество пациентов, принадлежащих к тому или иному категориальному признаку можно представить таким образом:



Анализ числовых признаков



Анализ числовых признаков

Можно заметить несколько особенностей:

1.у многих пациентов, умерших от сердечной недостаточности, фракция выброса ниже, чем у выживших. Они также имеют немного более высокие уровни креатинина сыворотки и креатинфосфокиназы.

2.Между функциями нет очевидно сильной корреляции, ниже я это проверяю, рассчитав коэффициент корреляции Спирмена R. Данный коэффициент корреляции может принимать значения от минус единицы до единицы, причем при R=1 имеет место строго прямая связь, а при R= -1 – строго обратная связь. Если коэффициент корреляции отрицательный, то имеет место обратная связь, если положительный, то – прямая связь. Если коэффициент корреляции равен нулю, то связь между величинами практически отсутствует. Чем ближе модуль коэффициента корреляции к единице, тем более сильной является связь между измеряемыми величинами.

Анализ числовых признаков

	platelets	ejection_fraction	creatinine_phosphokinase	serum_creatinine	serum_sodium
platelets	1.000000	0.053696	0.060033	-0.051028	0.049453
ejection_fraction	0.053696	1.000000	-0.067839	-0.177986	0.161695
creatinine_phosphokinase	0.060033	-0.067839	1.000000	-0.049936	0.016886
serum_creatinine	-0.051028	-0.177986	-0.049936	1.000000	-0.300413
serum_sodium	0.049453	0.161695	0.016886	-0.300413	1.000000

Между параметрами нет сильной корреляции, можно заметить положительную взаимосвязь у креатинина сыворотки и содержанием натрия в сыворотке, в медицинской литературе указано, что с возрастом содержание креатинина в сыворотке увеличивается , это и объясняет их слегка положительную взаимосвязь.

Создание модели

Сердечно-сосудистые заболевания (ССЗ) являются причиной смерти номер 1 в мире, унося, по оценкам, 17,9 миллиона жизней ежегодно, что составляет 31% всех смертей в мире.

Большинство сердечно-сосудистых заболеваний можно предотвратить, воздействуя на поведенческие факторы риска, такие как употребление табака, нездоровое питание и ожирение, гиподинамия и вредное употребление алкоголя, с использованием общегосударственных стратегий.

Цель исследования — создать модель, которая может успешно предсказать наступление смерти(положительный случай (1) в датасете) с вероятностью выше, чем случайное угадывание (0,5).

Для создания модели делю данные на обучающие и тестовые, а также стандартизирую числовые переменные, чтобы все переменные вносили одинаковый вклад в создание модели.

Для исследования буду создавать модели по алгоритму логистической регрессии и случайного леса.

Логистическая регрессия (Logistic Regression)

```
[5]  #Логистическая регрессия
     lg=LogisticRegression(C=5)
     lg.fit(x_train,y_train)
```

```
LogisticRegression(C=5)
```

```
[6]  lg.score(x_train,y_train)
```

```
0.7782426778242678
```

```
[7]  y_pred=lg.predict(x_test)
```

```
[8]  #точность модели(общее количество правильных прогнозов модели)
     acc = accuracy_score(y_test, y_pred)
     print("Точность модели логистической регрессии составляет (в %):", acc*100)
```

```
Точность модели логистической регрессии составляет (в %): 65.0
```

	precision	recall	f1-score	support
0	0.64	0.91	0.75	35
1	0.70	0.28	0.40	25
accuracy			0.65	60
macro avg	0.67	0.60	0.58	60
weighted avg	0.67	0.65	0.61	60

Логистическая регрессия (Logistic Regression)

Целью модели было предсказать, умрет ли кто-то из-за сердечной недостаточности или нет, в данном случае важно успешно прогнозировать положительные случаи (1) по сравнению с фактическими положительными случаями.

Судя по отчету о классификации можно сделать вывод, что из всех случаев со смертельным исходом по прогнозу модели, только в 70% на самом деле так и было Из всех случаев со смертельным исходом, которые действительно были, модель правильно предсказала так на 28% случаев

Значение F1 - средневзвешенное гармоническое значение точности и полноты, чем ближе к 1, тем лучше модель. В данном случае значение 0.40, что не очень близко к 1, это говорит о том, что модель плохо справляется с прогнозом

Случайный лес (RandomForestClassifier)

```
[10] #случайный лес (RandomForestClassifier)

     Rf=RandomForestClassifier()
     Rf.fit(x_train,y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier()
```

```
[11] Rf.score(x_train,y_train)

     1.0
```

```
[12] y_pred=Rf.predict(x_test)
```

```
[13] #точность модели(общее количество правильных прогнозов модели)
     acc = accuracy_score(y_test, y_pred)
     print("Точность модели RandomForestClassifier составляет (в %):", acc*100)

     Точность модели RandomForestClassifier составляет (в %): 70.0
```

```
[14] print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.70	0.86	0.77	35
1	0.71	0.48	0.57	25
accuracy			0.70	60
macro avg	0.70	0.67	0.67	60
weighted avg	0.70	0.70	0.69	60

Случайный лес (RandomForestClassifier)

Судя по отчету о классификации в этой модели делаю вывод, что из всех случаев со смертельным исходом по прогнозу модели, в 71% на самом деле так и было, что на 1% выше, чем у предыдущей модели Из всех случаев со смертельным исходом, которые действительно были, модель правильно предсказала так на 48% случаев, что также больше, чем у модели логистической регрессии на 20%.

Значение F1 - средневзвешенное гармоническое значение точности и полноты, чем ближе к 1, тем лучше модель. В данном случае значение 0.57, что гораздо ближе к 1, это говорит о том, что модель лучше справляется с прогнозом.

Рассмотрев 2 модели, можно увидеть, что оценка отзыва положительного случая (1) в 1 модели составляет 0,28, что очень мало, во 2 модели - 0.48, что, безусловно, гораздо более высокая оценка отзыва, но меньше, чем вероятность случайного угадывания, это говорит о том, что данные модели использовать нецелесообразно