Московский Авиационный Институт
(Национальный Исследовательский Университет)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу
«Машинное обучение»

Data Mining и исследование данных

Студент: Горохов М.А.
Группа: М80 – 308Б-19
Оценка: _____
Подпись: _____

Москва, 2022

1. Постановка задачи

Найти набор данных и провести исследовательский анализ. Подготовить отчет с результатами исследования.

2. Описание датасета

Имеется датасет с информацией о пациентах США. Нужно выявить пациентов с высоким риском смерти из-за болезни сердца.

Имеем следующие сведения об опрошенных людях

3. Описание признаков

- age - возраст пациента
- anaemia - анемия, Снижение эритроцитов или гемоглобина (boolean)
- creatinine_phosphokinase - Уровень фермента CPK в крови (мкг/л)
- diabetes - есть ли диабет (boolean)
- ejection_fraction - Процент крови, покидающей сердце при каждом сокращении (в процентах)
- high_blood_pressure - Гипертония (boolean)
- platelets - Тромбоциты в крови (килотромбоциты/мл)
- serum_creatinine - Уровень сывороточного креатинина в крови (мг/дл)
- serum_sodium - Уровень сывороточного натрия в крови (мэкв/л)
- sex - пол пациента
- smoking - курит ли пациент (boolean)
- time - Период наблюдения (дни)
- DEATH_EVENT - умер ли пациент в течение периода наблюдения (логическое значение)

Пояснение по поводу boolean данных:

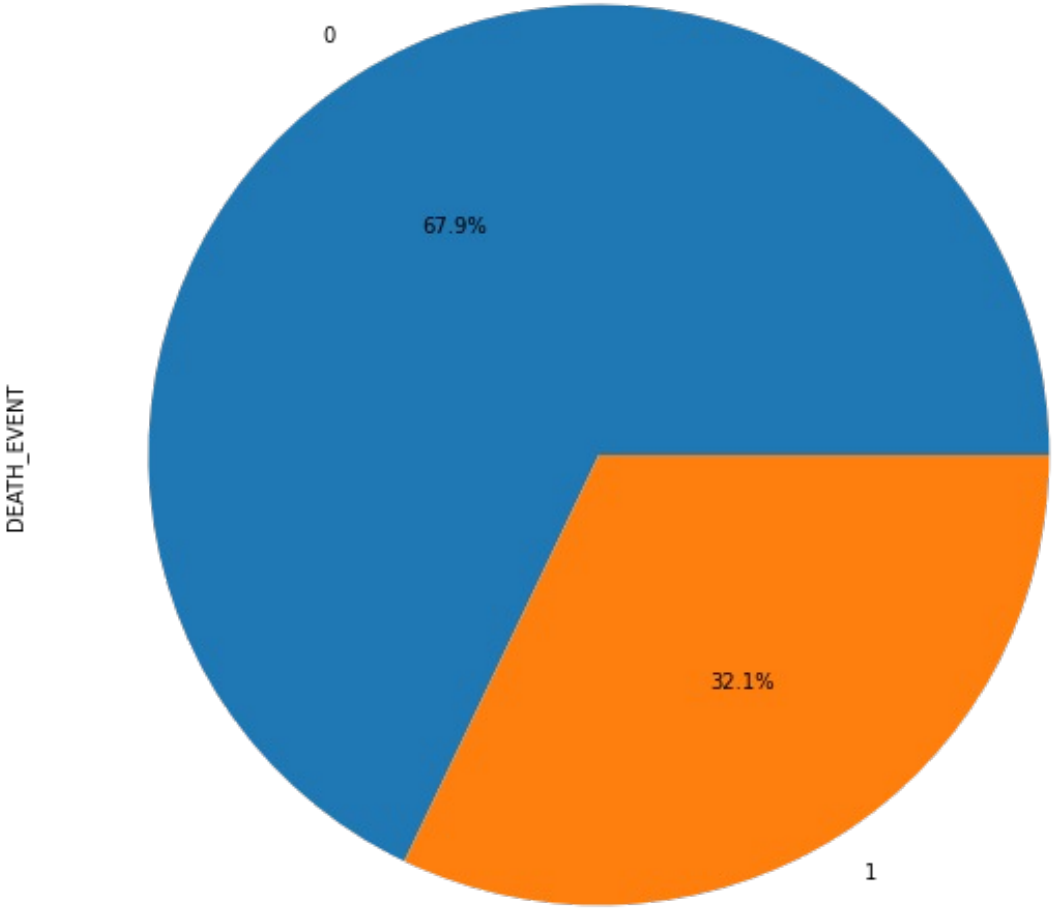
- Sex - Пол пациента Мужской = 1, Женский = 0
- Age - Возраст пациента
- Diabetes - 0 = Нет, 1 = Да
- Anaemia - 0 = Нет, 1 = Да
- high_blood_pressure — 0 = Нет, 1 = Да
- smoking - 0 = Нет, 1 = Да
- DEATH_EVENT — 0 = нет, 1 = да

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

Датасет не содержит пропусков и повторений, можно приступать к анализу.

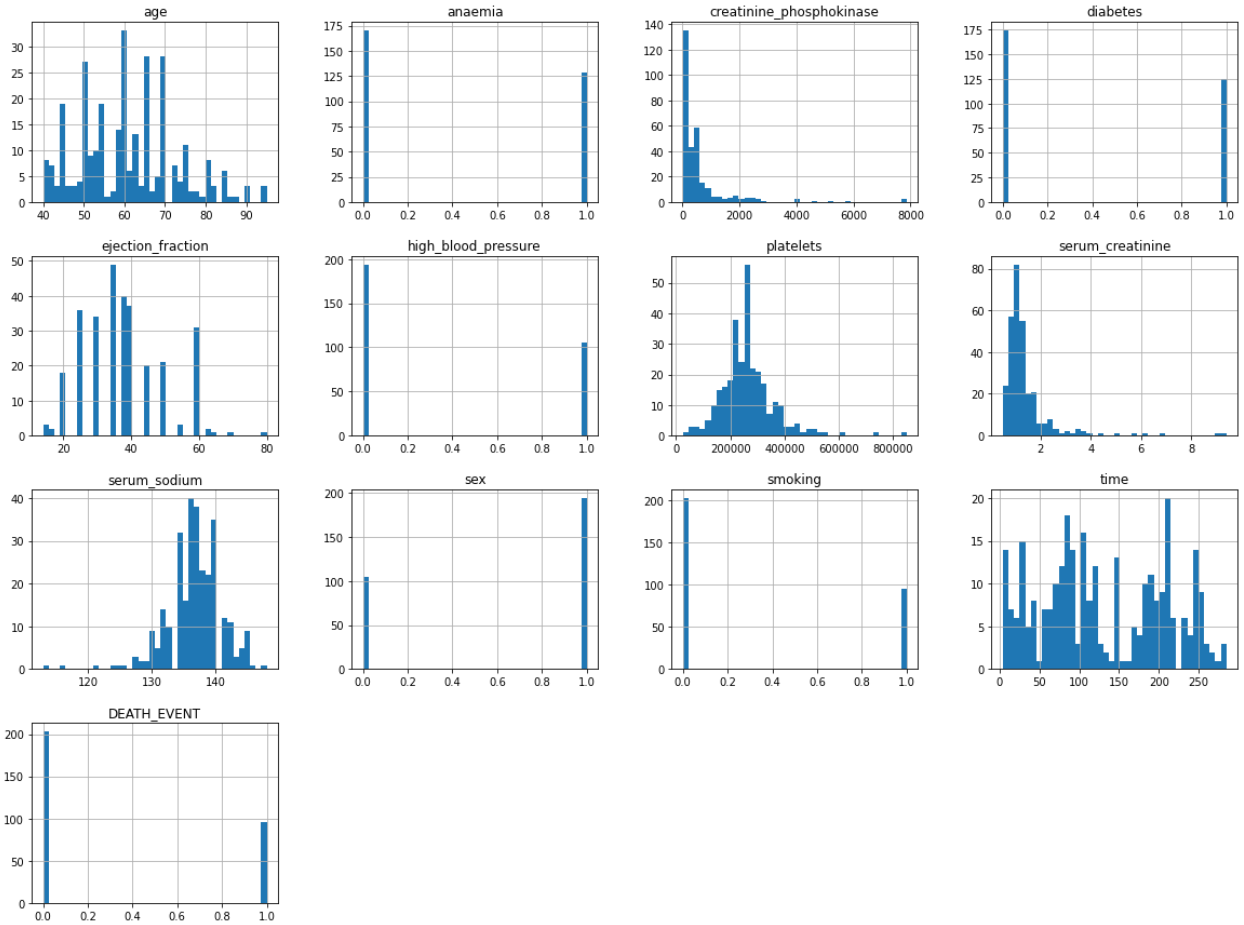
4. Признаки

Умершие пациенты: 32.11 % (96 пациентов)
Неумершие: 67.89 % (203 пациентов)



```
# Информация о данных
data.describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='RdPu')
```

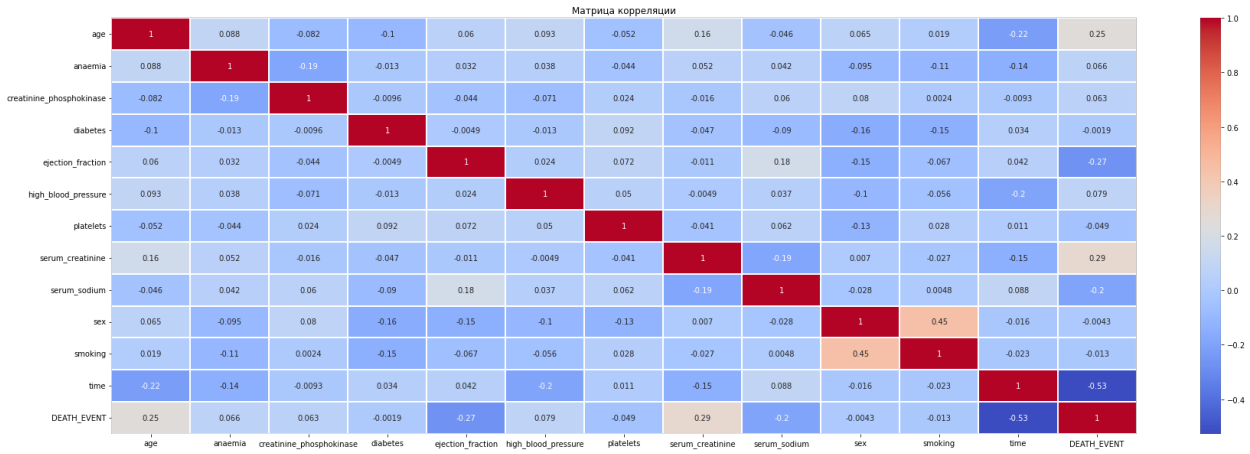
	count	mean	std	min	25%	50%	75%	max
age	299.000000	60.833893	11.894809	40.000000	51.000000	60.000000	70.000000	95.000000
anaemia	299.000000	0.431438	0.496107	0.000000	0.000000	0.000000	1.000000	1.000000
creatinine_phosphokinase	299.000000	581.839465	970.287881	23.000000	116.500000	250.000000	582.000000	7861.000000
diabetes	299.000000	0.418060	0.494067	0.000000	0.000000	0.000000	1.000000	1.000000
ejection_fraction	299.000000	38.083612	11.834841	14.000000	30.000000	38.000000	45.000000	80.000000
high_blood_pressure	299.000000	0.351171	0.478136	0.000000	0.000000	0.000000	1.000000	1.000000
platelets	299.000000	263358.029264	97804.236869	25100.000000	212500.000000	262000.000000	303500.000000	850000.000000
serum_creatinine	299.000000	1.393880	1.034510	0.500000	0.900000	1.100000	1.400000	9.400000
serum_sodium	299.000000	136.625418	4.412477	113.000000	134.000000	137.000000	140.000000	148.000000
sex	299.000000	0.648829	0.478136	0.000000	0.000000	1.000000	1.000000	1.000000
smoking	299.000000	0.321070	0.467670	0.000000	0.000000	0.000000	1.000000	1.000000
time	299.000000	130.260870	77.614208	4.000000	73.000000	115.000000	203.000000	285.000000
DEATH_EVENT	299.000000	0.321070	0.467670	0.000000	0.000000	0.000000	1.000000	1.000000



5. Таргет

Посмотрим на распределение таргета.

Посмотрим на корреляционную матрицу. Матрица с численными значениями слишком большая, чтобы вставлять ее сюда. При необходимости ее можно посмотреть в ноутбуке.



На основе полученной корреляционной матрицы можно видеть:

- Сильная корреляция между age, serum_creatinine и DEATH_EVENT
- Положительная корреляция между anaemia, creatinine_phosphokinase, high_blood_pressure и DEATH_EVENT
- Отрицательная корреляция между ejection_fraction, platelets, serum_sodium и time имеет HeartDisease
- serum_creatinine имеет наибольшую положительную корреляцию с HeartDisease

Действительно, высокий уровень креатина свидетельствует о заболеваниях почек. Есть выражение "почки топят сердце", то есть из-за плохой работы почек слабеет сердце, из-за чего возникает преждевременная смерть.

6. Вывод

В данной лабораторной работе я провел полное исследование медицинского датасета. Я изучил все имеющиеся признаки. Я практиковался в работе с таблицами и в визуализации.

Я убедился, что целевая переменная зависит от имеющихся признаков. Следовательно, у нас есть все шансы получить хорошую модель.