

**Московский авиационный институт  
(национальный исследовательский университет)**

**Институт №8 «Информационные технологии и прикладная  
математика»**

**Кафедра 806 «Вычислительная математика и  
программирование»**

**Лабораторная работа №0 по курсу «Искусственный интеллект»**

Студент: С. Л. Бугреев  
Преподаватели: Д. В. Сошников  
С. Х. Ахмед  
Группа: М8О-401Б-19  
Дата:  
Оценка:  
Подпись:

**Москва, 2022**

## Лабораторная работа №0

**Задача:** В данной лабораторной работе вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте. И главное, день промедления и вас опередит ваш конкурент, да и спланированная работа отразится на репутации. По сути в данной лабораторной работе вы выполняете часть работы VI системы.

# 1 Ход работы

Я выбрал набор данных Predicting Heart Failure [1] для выполнения лабораторной работы. В описании датасета предлагается предсказать, будет ли в ближайшее время у человека сердечный приступ или нет. Признаки в наборе данных:

1. age — возраст человека, который находится под наблюдением
2. anaemia — показатель есть ли у наблюдаемого пациента анемия.
3. high blood pressure — показатель, повышенное ли у человека давление.
4. creatinine phosphokinase — уровень СРК в крови человека, числовой признак.
5. diabetes — есть ли у пациента диабет.
6. ejection fraction — процент крови, покидающее сердце после каждого удара, выражается в процентах, числовой признак.
7. platelets — количество тромбоцитов в крови, числовой признак.
8. sex — пол пациента.
9. serum creatinine — количество креатинина в сыворотке крови человека.
10. serum sodium — количество натрия в сыворотке крови человека.
11. smoking — курящий человек или нет.
12. time — количество дней под наблюдением.
13. DEATH\_EVENT — целевая переменная, указывает, произошёл ли сердечный приступ в итоге или нет.

Перед выявлением зависимостей между признаками следует проверять целостность набора данных:

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 299 entries, 0 to 298
```

```
Data columns (total 13 columns):
```

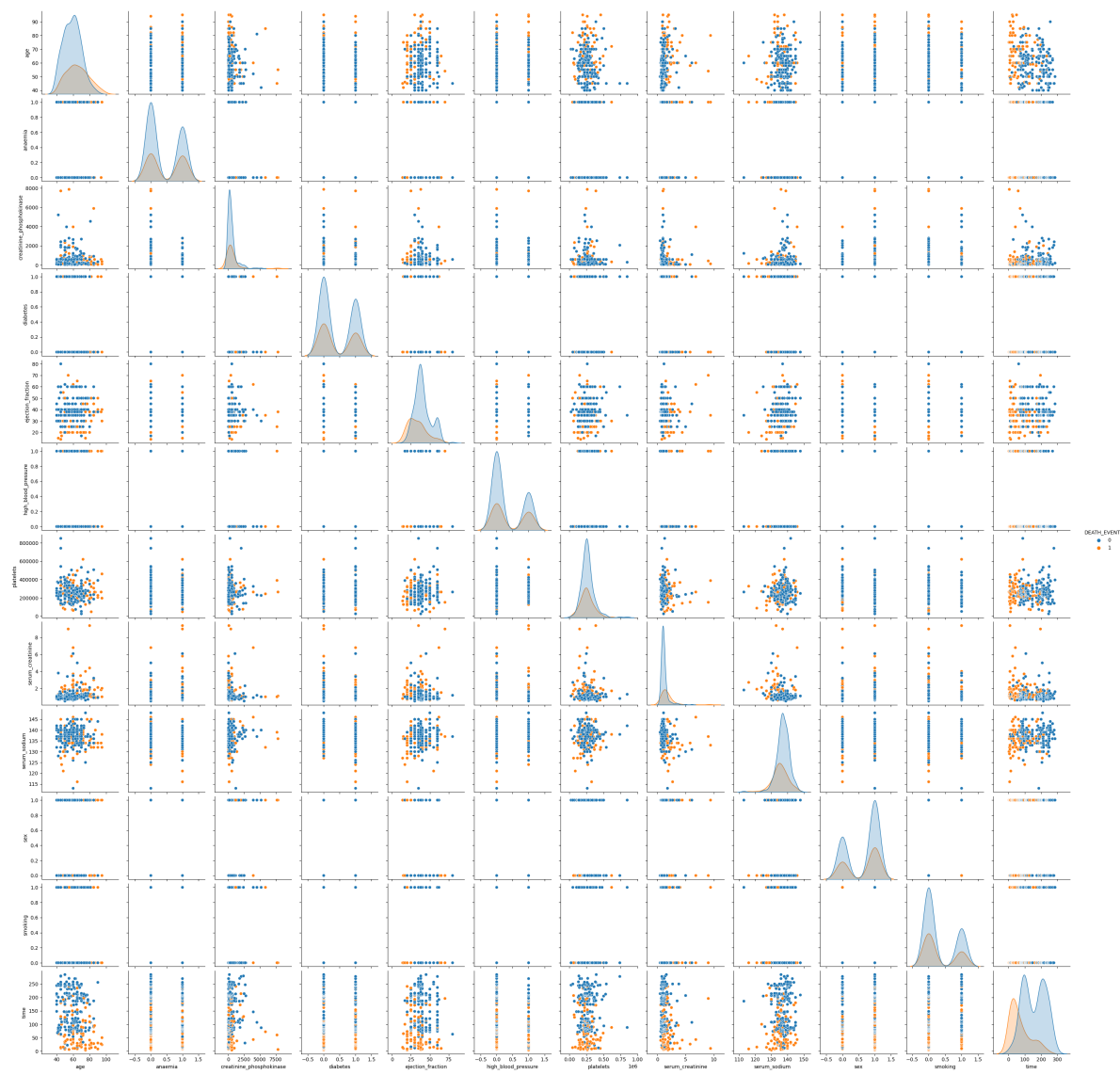
#	Column	Non-Null Count	Dtype
0	age	299 non-null	float64
1	anaemia	299 non-null	int64
2	creatinine_phosphokinase	299 non-null	float64

3	diabetes	299 non-null	int64
4	ejection_fraction	299 non-null	float64
5	high_blood_pressure	299 non-null	int64
6	platelets	299 non-null	float64
7	serum_creatinine	299 non-null	float64
8	serum_sodium	299 non-null	float64
9	sex	299 non-null	int64
10	smoking	299 non-null	int64
11	time	299 non-null	float64
12	DEATH_EVENT	299 non-null	int64

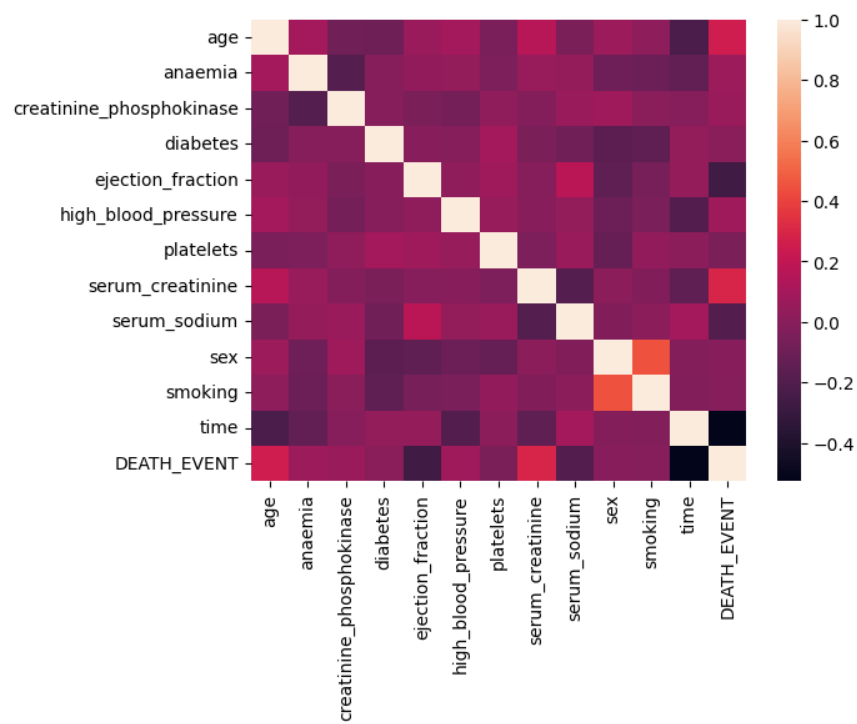
dtypes: float64(7),int64(6)  
memory usage: 30.5 KB

В наборе нет неполных данных, а все признаки - числовые.

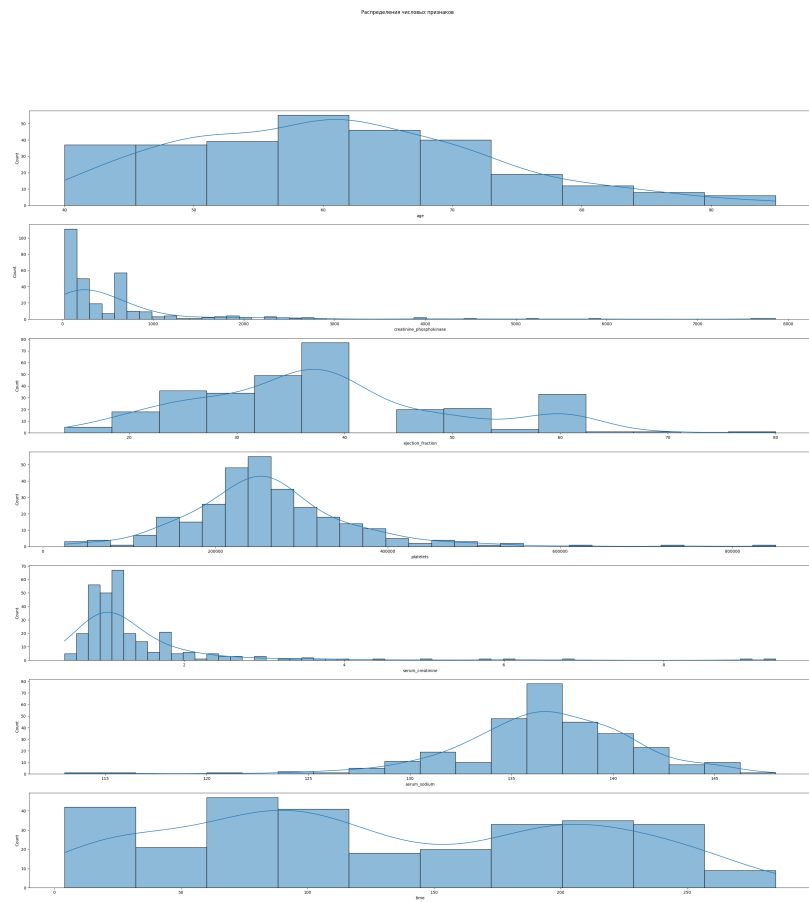
Построю графики для каждой пары признаков. Синим отмечен успех, оранжевым - неуспех:



Построю корреляционную матрицу для признаков:

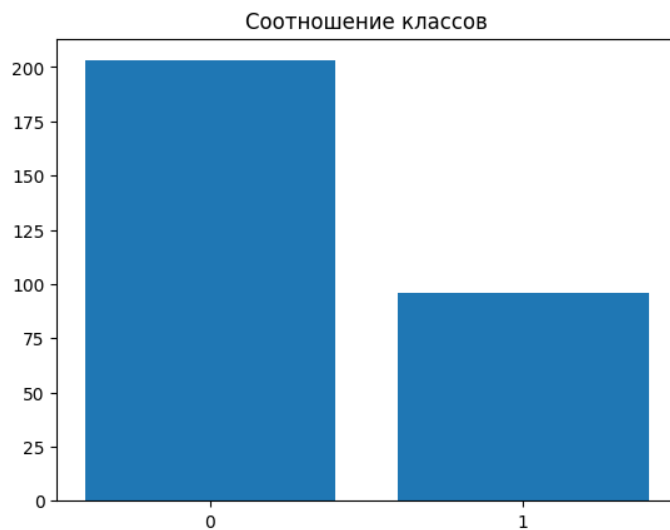


Так же построю гистограммы для числовых признаков:



Выбросов не было обнаружено, так как датасет довольно маленький.

Соотношение классов объектов:



Количество объектов разных классов заметно различается, сделаем RandomOversampling при помощи imbalanced-learn, это позволит не прибегать к неточным синтетическим данным, при этом наши данные будут лучше подходить для обработки алгоритмами классического МЛ. Данные готовы к обучению.



## 2 Выводы

В ходе выполнения лабораторной работы я освежил в памяти курс математической статистики: гистограмму, корреляцию и корреляционную матрицу для наборов данных. Так же я изучил библиотеку Pandas, она оказалась очень удобной для анализа данных.

Во время анализа датасета я заметил, что данных мало, то есть их не хватает, судя по парным графикам, данные не так хорошо разделимы линейными моделями.

В итоге, результаты которого получились неоднозначные: данные возможно будут нормально обрабатываться классическими алгоритмами МЛ, а, возможно, и нет.

## Список литературы

- [1] *Beginner's Classification Dataset*

URL: <https://www.kaggle.com/datasets/whenamancodes/heart-failure-clinical-records>  
(дата обращения: 30.09.2022).

- [2] *Exploratory data analysis with Pandas — mlcourse.ai*

URL: [https://mlcourse.ai/book/topic01/topic01\\_pandas\\_data\\_analysis.html](https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html)  
(дата обращения: 30.09.2022).