Анализ влияния факторов на возникновение рака легких

Канюков Д.Р.
Радостев С.М.

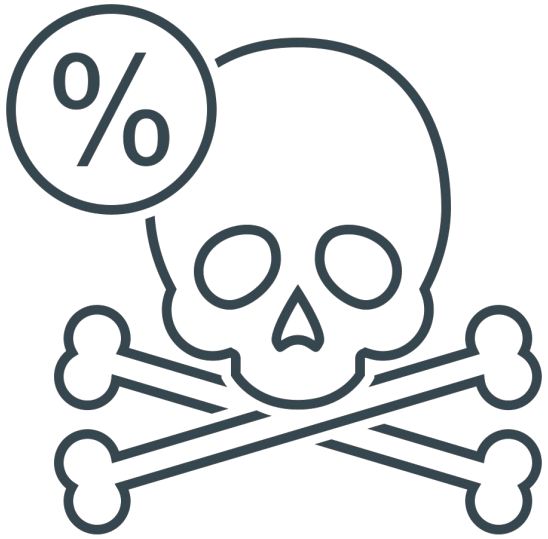
Паспорт проекта

Цель: Выполнить анализ данных о пациентах и построить модель зависимости возникновения рака легких от различных факторов с помощью машинного обучения, позволяющую делать прогнозы с высокой точностью.

Задачи:

1. Выполнить анализ проблемы, обосновать ее актуальность.
2. Осуществить загрузку данных и подготовку их к анализу количественными методами, включая устранение пропущенных значений.
3. Выполнить предварительный анализ данных, корреляционный анализ.
4. Осуществить моделирование зависимости целевого признака от факторных методами машинного обучения, в том числе подобрать наилучшую модель, оценить ее качество и выполнить прогнозирование.
5. Выполнить интерпретацию полученных результатов и сделать выводы о достижении цели.

Анализ проблемы исследования



~20%

Исходные данные

Исходный размер датасета: 16x284 | Целевая переменная: Lung_Cancer

Информация об атрибутах:

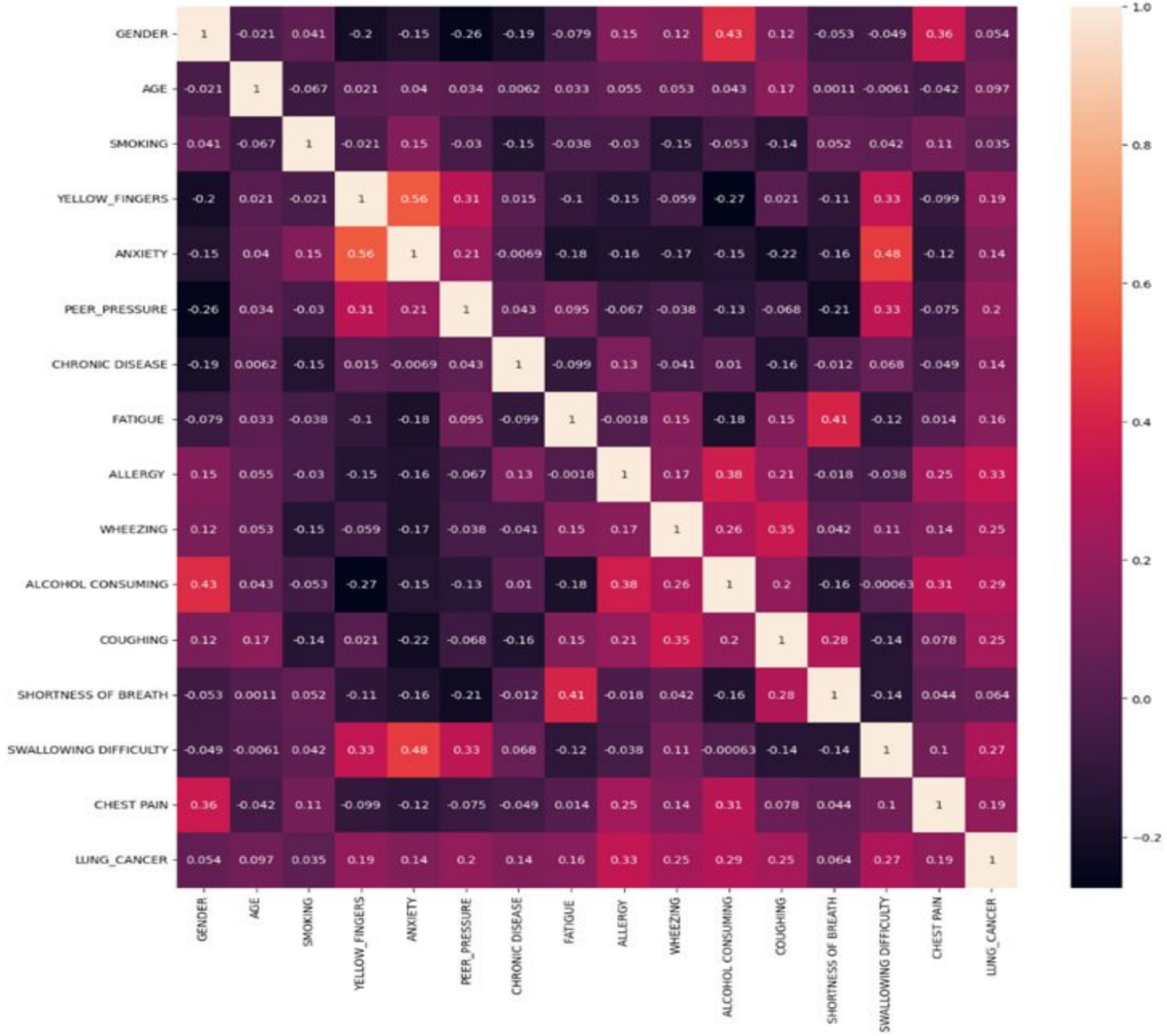
1. Gender – пол пациента (M – мужской, F – женский).
2. Age – возраст пациента.
3. Smoking – курение
4. Yellow_fingers – желтые пальцы
5. Anxiety – тревожность
6. Peer_pressure – давление со стороны окружения
7. Chronic Disease – хронические заболевания
8. Fatigue – утомляемость
9. Allergy – аллергия
10. Wheezing – хрип
11. Alcohol consuming – употребление алкоголя
12. Coughing – кашель
13. Shortness of Breath – одышка
14. Swallowing Difficulty – трудности с глотанием
15. Chest pain – боль в груди
16. Lung_Cancer – рак легких

Кодирование данных

```
1 df.head()
```

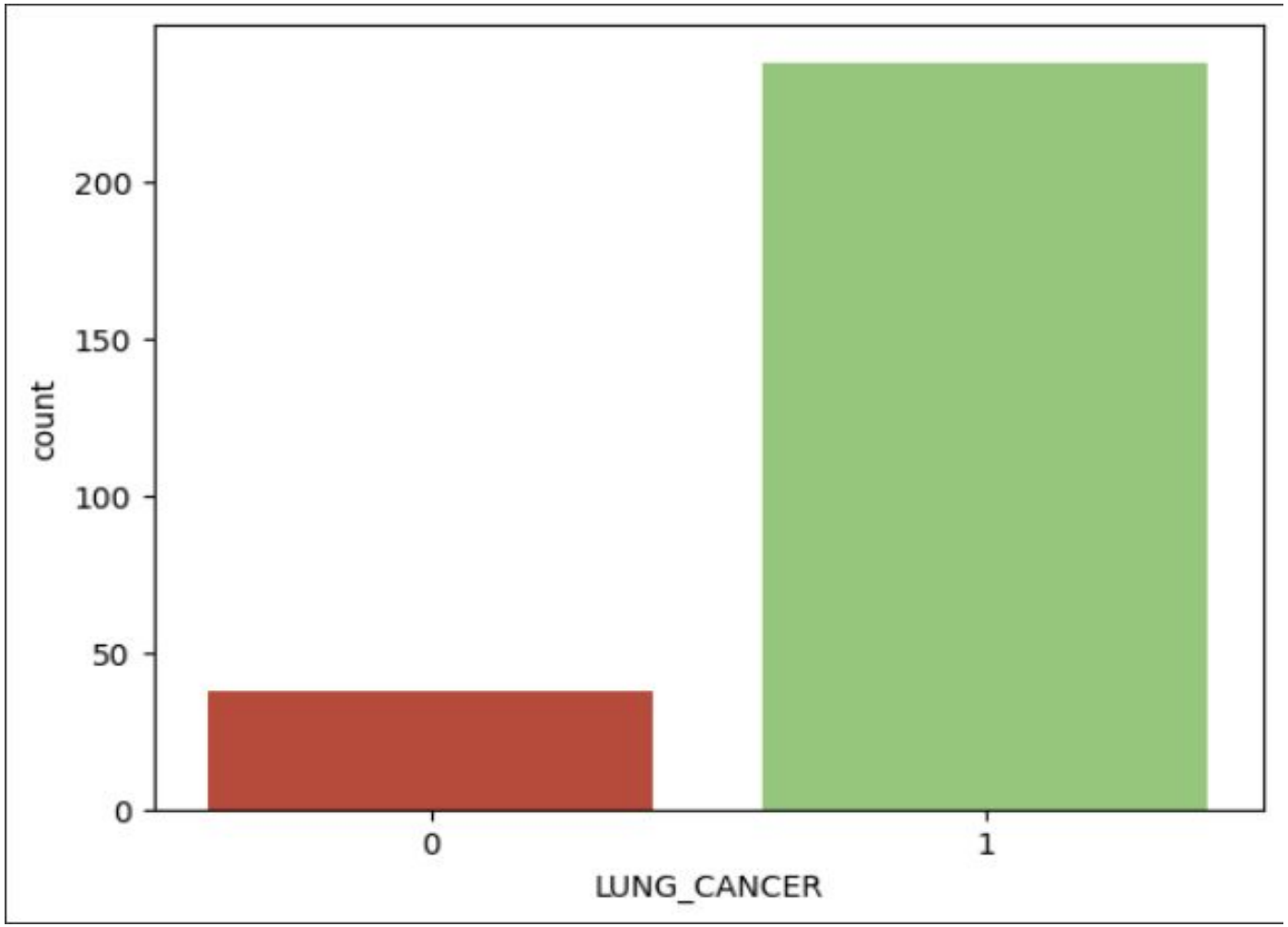
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	1	26	0	1	1	0	0	1	0	1	1	1	1	1	1	1
1	1	31	1	0	0	0	1	1	1	0	0	0	1	1	1	1
2	0	16	0	0	0	1	0	1	0	1	0	1	1	0	1	0
3	1	20	1	1	1	0	0	0	0	0	1	0	0	1	1	0
4	0	20	0	1	0	0	0	0	0	1	0	1	1	0	0	0

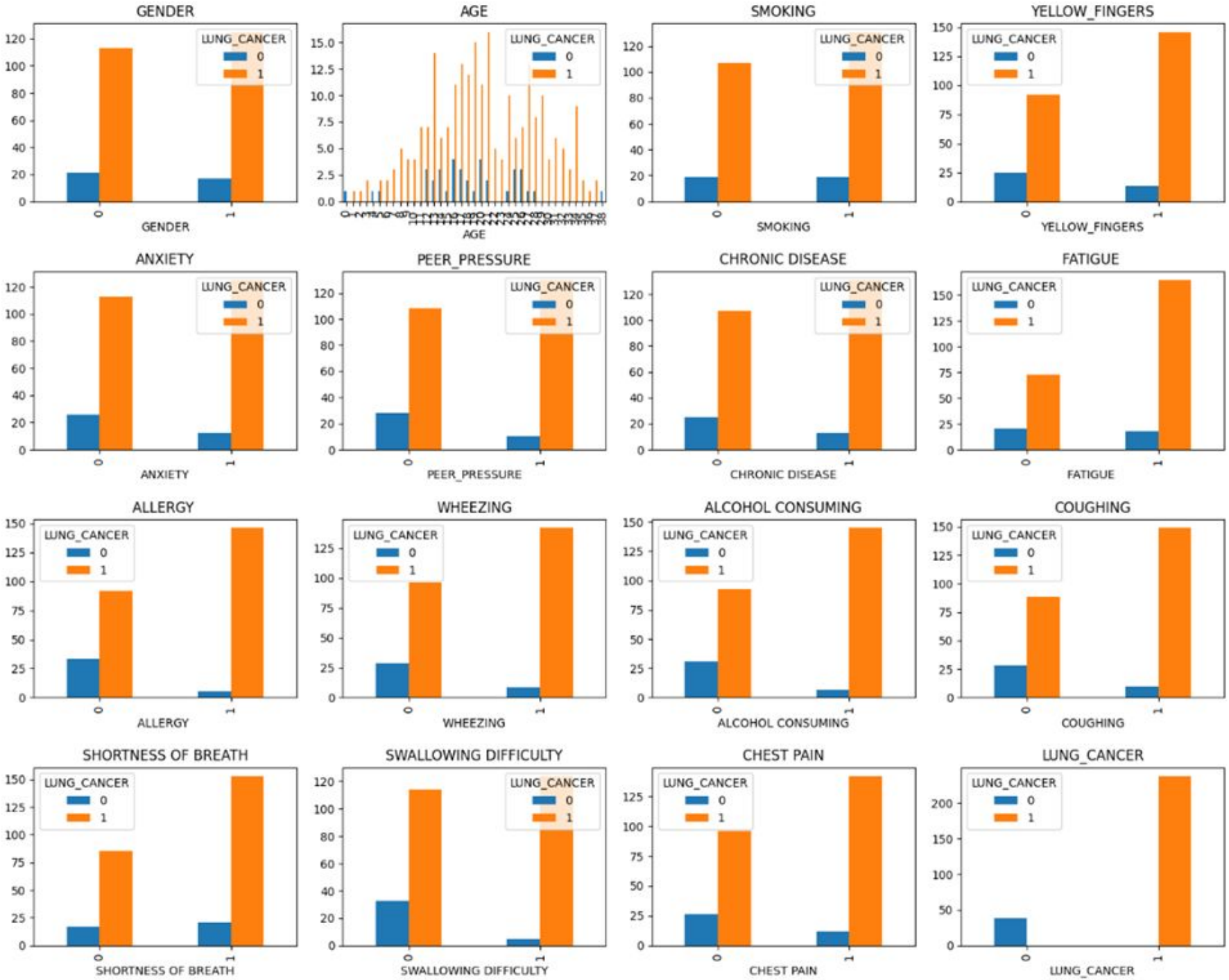
Тепловая карта



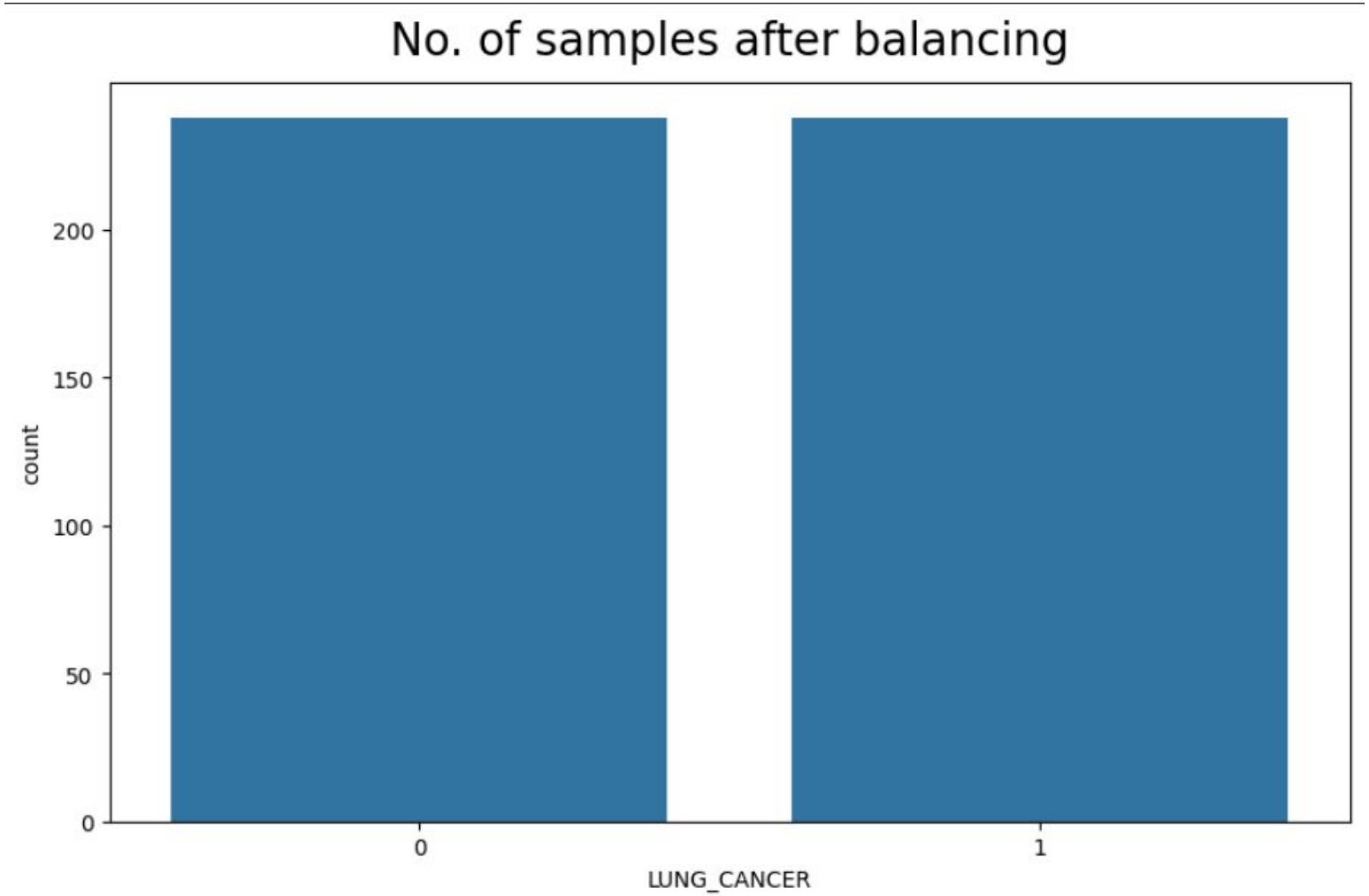

LUNG_CANCER
GENDER
AGE
SMOKING
YELLOW_FINGERS
ANXIETY
PEER_PRESSURE
CHRONIC DISEASE
FATIGUE
ALLERGY
WHEEZING
ALCOHOL CONSUMING
COUGHING
SHORTNESS OF BREATH
SWALLOWING DIFFICULTY
CHEST PAIN
LUNG_CANCER
0.054
0.097
0.035
0.19
0.14
0.2
0.14
0.16
0.33
0.25
0.29
0.25
0.064
0.27
0.19
1

Несбалансированность




GENDER
LUNG_CANCER
0
1
GENDER
AGE
LUNG_CANCER
0
1
AGE
SMOKING
LUNG_CANCER
0
1
SMOKING
YELLOW_FINGERS
LUNG_CANCER
0
1
YELLOW_FINGERS
ANXIETY
LUNG_CANCER
0
1
ANXIETY
PEER_PRESSURE
LUNG_CANCER
0
1
PEER_PRESSURE
CHRONIC DISEASE
LUNG_CANCER
0
1
CHRONIC DISEASE
FATIGUE
LUNG_CANCER
0
1
FATIGUE
ALLERGY
LUNG_CANCER
0
1
ALLERGY
WHEEZING
LUNG_CANCER
0
1
WHEEZING
ALCOHOL CONSUMING
LUNG_CANCER
0
1
ALCOHOL CONSUMING
COUGHING
LUNG_CANCER
0
1
COUGHING
SHORTNESS OF BREATH
LUNG_CANCER
0
1
SHORTNESS OF BREATH
SWALLOWING DIFFICULTY
LUNG_CANCER
0
1
SWALLOWING DIFFICULTY
CHEST PAIN
LUNG_CANCER
0
1
CHEST PAIN
LUNG_CANCER
LUNG_CANCER
0
1
LUNG_CANCER

Балансировка



	precision	recall	f1-score	support
0	0.91	0.94	0.92	51
1	0.93	0.89	0.91	45
accuracy			0.92	96
macro avg	0.92	0.92	0.92	96
weighted avg	0.92	0.92	0.92	96

Логистическая регрессия

	precision	recall	f1-score	support
0	0.93	1.00	0.96	51
1	1.00	0.91	0.95	45
accuracy			0.96	96
macro avg	0.96	0.96	0.96	96
weighted avg	0.96	0.96	0.96	96

Случайный лес

	precision	recall	f1-score	support
0	0.88	1.00	0.94	51
1	1.00	0.84	0.92	45
accuracy			0.93	96
macro avg	0.94	0.92	0.93	96
weighted avg	0.94	0.93	0.93	96

Дерево решений

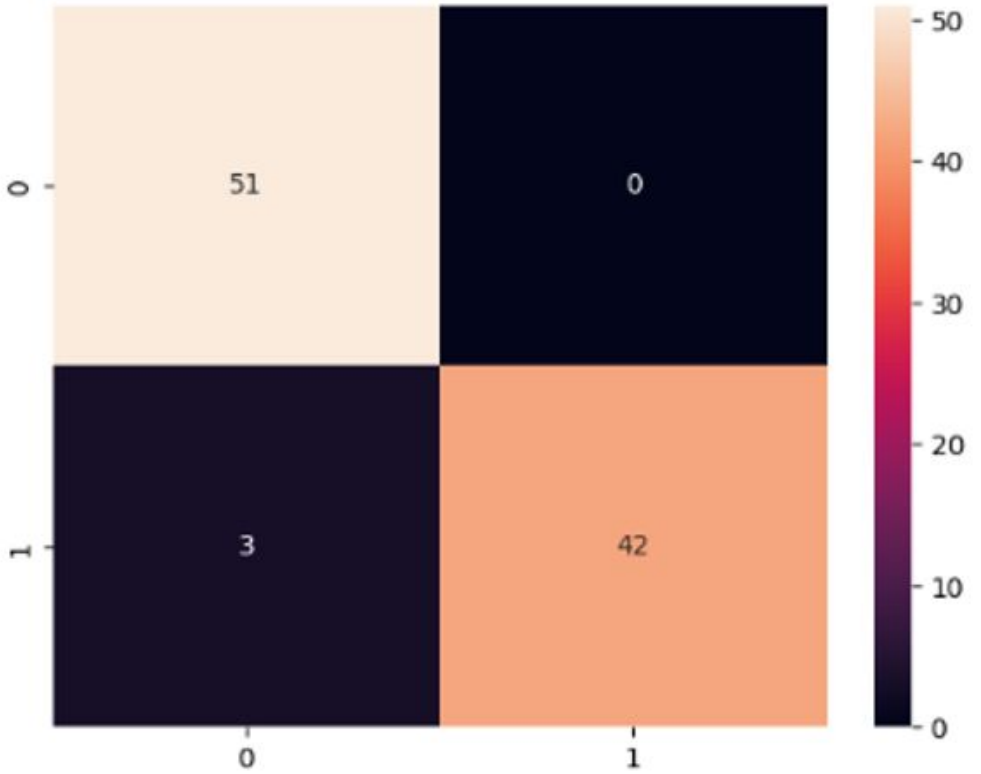
	precision	recall	f1-score	support
0	0.85	1.00	0.92	51
1	1.00	0.80	0.89	45
accuracy			0.91	96
macro avg	0.93	0.90	0.90	96
weighted avg	0.92	0.91	0.90	96

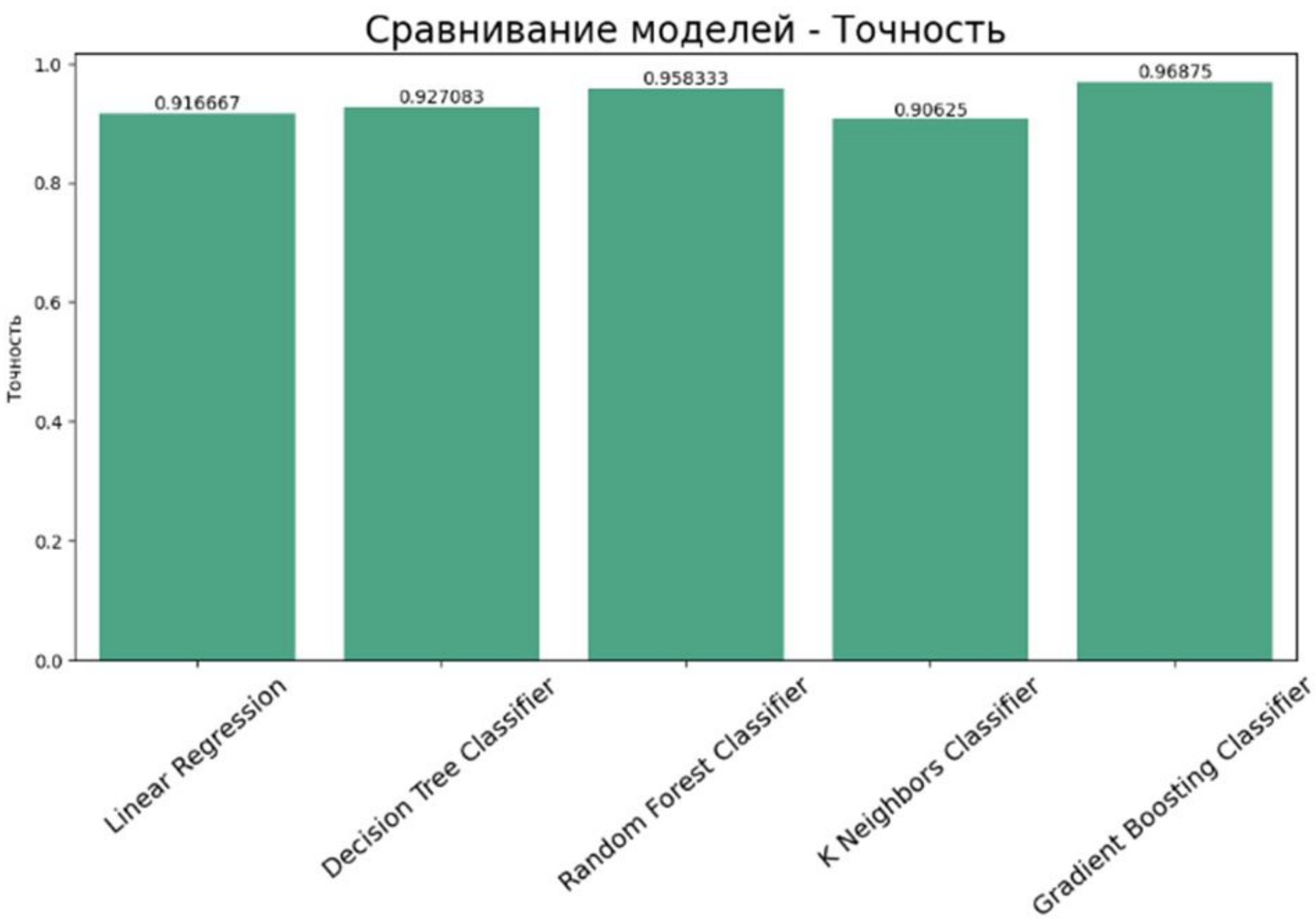
К ближайших соседей

Градиентный бустинг

	precision	recall	f1-score	support
0	0.94	1.00	0.97	51
1	1.00	0.93	0.97	45
accuracy			0.97	96
macro avg	0.97	0.97	0.97	96
weighted avg	0.97	0.97	0.97	96

Матрица ошибок




Сравнивание моделей - Точность
Точность
1.0
0.8
0.6
0.4
0.2
0.0
0.916667
0.927083
0.958333
0.90625
0.96875
Linear Regression
Decision Tree Classifier
Random Forest Classifier
K Neighbors Classifier
Gradient Boosting Classifier

Спасибо за внимание