|  |  |
| --- | --- |
| Федеральное государственное автономное  образовательное учреждение высшего образования  «Пермский государственный национальный исследовательский университет» (ПГНИУ)  Региональный институт непрерывного образования (РИНО ПГНИУ)  Цифровая кафедра | |
|  | |
| Выпускная аттестационная (квалификационная) работа  по курсу профессиональной переподготовки  «Анализ данных и машинное обучение» | |
| **АНАЛИЗ ВЛИЯНИЯ ФАКТОРОВ НА ВОЗНИКНОВЕНИЕ РАКА ЛЕГКИХ** | |
|  | |
|  | Разработчики проекта:  Канюков Дмитрий Романович,  Радостев Степан Михайлович |
| Пермь 2024 | |

СОДЕРЖАНИЕ

[ПАСПОРТ ПРОЕКТА 3](#_Toc29296)

[СОДЕРЖАНИЕ ПРОЕКТА 4](#_Toc5298)

[Анализ проблемы исследования 4](#_Toc30748)

[Исходные данные 5](#_Toc11452)

[Реализация проекта 6](#_Toc31217)

[Этап 1. Загрузка и предобработка данных 6](#_Toc31236)

[Этап 2. Кодирование категориальных признаков 9](#_Toc2028)

[Этап 3. Визуализация данных 9](#_Toc9279)

[Этап 4. Разделение данных на обучающую и тестовую выборки и балансировка данных 13](#_Toc27602)

[Этап 5. Моделирование и оценка моделей 14](#_Toc13155)

[Этап 6. Сравнение моделей 19](#_Toc26112)

[Заключение 21](#_Toc17839)

[Библиографический список 22](#_Toc17857)

# ПАСПОРТ ПРОЕКТА

**Название проекта:** Анализ влияния факторов на возникновение рака легких.

**Сведения об авторах:** Канюков Дмитрий Романович, Радостев Степан Михайлович.

**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, в том числе выявление и обработку выбросов, проверку распределения данных на нормальность, 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.

**Краткое описание проекта:**

Требуется проанализировать данные о пациентах и определить, возможно ли описать зависимость возникновения рака легких от имеющихся факторных переменных методами машинного обучения. Дать интерпретацию полученным результатам и сделать выводы о влиянии различных факторов на риск заболевания.

**Конкретные ожидаемые результаты:**

Построенная регрессионная модель и рекомендации по ее использованию.

# СОДЕРЖАНИЕ ПРОЕКТА

## Анализ проблемы исследования

Рак легких — одна из самых распространенных и смертоносных форм рака в мире. По данным GLOBOCAN, смертность от рака легких составляет 18% от всех случаев смерти от рака. В Новосибирской области этот показатель еще выше — 19,7% в 2021 году.

Заболеваемость раком легкого в России немного снизилась за последние три года, но в Новосибирской области смертность остается высокой. Это может быть связано с поздней диагностикой: в 2021 году 41,5% случаев были обнаружены на последней стадии.

Рак легкого делится на два основных типа: мелкоклеточный (15% случаев) и немелкоклеточный (85% случаев). Немелкоклеточный рак включает аденокарциному (40%), плоскоклеточный рак (30%) и крупноклеточную карциному (15%). Аденокарцинома чаще встречается у некурящих, тогда как плоскоклеточный рак — у курящих [1].

Рак легкого трудно лечить, особенно на поздних стадиях. Стандартное лечение не всегда эффективно, и выживаемость остается низкой. Современные исследования открыли новые методы лечения, особенно для аденокарциномы, но часто возникает устойчивость к этим методам.

Частота мутаций, которые помогают в лечении, не превышает 30%. В плоскоклеточном раке таких мутаций пока не найдено. Частота мутаций зависит от региона, и в Новосибирской области такие исследования ранее не проводились.

Иммунная система играет важную роль в борьбе с раком. Опухоли могут избегать иммунного контроля, и преодоление этой устойчивости — важная задача. Эпигенетические механизмы, такие как изменения в ДНК, также важны в развитии рака легкого.

Курение — главный фактор риска рака легкого. Опухоли у курящих и некурящих пациентов различаются, и у курящих чаще встречается повышенная экспрессия PD-L1, что делает их подходящими для иммунотерапии.

*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, в том числе выявление и обработку выбросов, проверку распределения данных на нормальность, 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.

## Исходные данные

Этот набор данных является бесценным активом в сфере здравоохранения, обеспечивая структурированную основу для разработки моделей выявления рака. В настоящей работе анализируется эффективность системы прогнозирования рака легких, которая помогает людям узнать свой риск заболевания с минимальными затратами и принять соответствующие решения на основе этого риска. Данные собраны с веб-сайта онлайн-системы прогнозирования рака легких. Этот набор данных иллюстрирует разнообразие симптомов рака легких.

Общее количество атрибутов: 16.

Количество записей: 284.

Информация об атрибутах:

1. Gender – пол пациента (M – мужской, F – женский).
2. Age – возраст пациента.
3. Smoking – курение (ДА-2, НЕТ-1).
4. Yellow fingers – желтые пальцы (ДА-2, НЕТ-1).
5. Anxiety – тревожность (ДА-2, НЕТ-1).
6. Peer\_pressure – давление со стороны окружения (ДА-2, НЕТ-1).
7. Chronic Disease – хронические заболевания (ДА-2, НЕТ-1).
8. Fatigue – утомляемость (ДА-2, НЕТ-1).
9. Allergy – аллергия (ДА-2, НЕТ-1).
10. Wheezing – одышка (ДА-2, НЕТ-1).
11. Alcohol – употребление алкоголя (ДА-2, НЕТ-1).
12. Coughing – кашель (ДА-2, НЕТ-1).
13. Shortness of Breath – одышка (ДА-2, НЕТ-1).
14. Swallowing Difficulty – трудности с глотанием (ДА-2, НЕТ-1).
15. Chest pain – боль в груди (ДА-2, НЕТ-1).
16. Lung Cancer – рак легких (ДА-YES, НЕТ-NO).

Необходимо проанализировать данные и определить, возможно ли описать зависимость риска рака легких от имеющихся факторных переменных методом линейной регрессии. Lung Cancer - целевая переменная.

Выдвинем гипотезу исследования: риск рака легких зависит от пола, возраста, курения, желтых пальцев, тревожности, давления со стороны окружения, хронических заболеваний, утомляемости, аллергии, одышки, употребления алкоголя, кашля, одышки, трудностей с глотанием и боли в груди, и может быть предсказан с помощью этих показателей.

## Реализация проекта

### Этап 1. Загрузка и предобработка данных

На этом этапе выполняется загрузка данных из CSV-файла и проводится первичный анализ данных.

Для начала работы необходимо подключить библиотеки.



Рисунок 1 - Подключение библиотек

Загрузка данных в датафрейм и вывод первых 5 строк для просмотра содержимого.

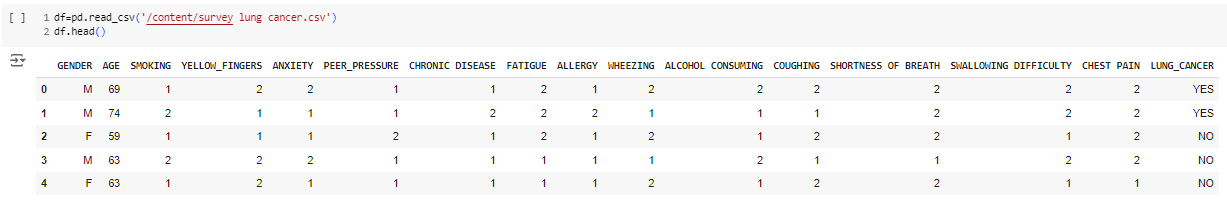


Рисунок 2 - Загрузка данных и их вывод

Вывод информации о столбцах датасета и вывод статистической информации. На рисунке 3 видно, что все столбцы имеют числовой тип данных.

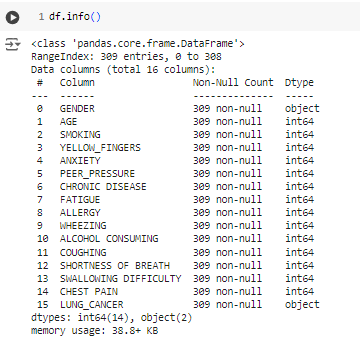


Рисунок 3 - Типы данных и наличие пропущенных значений

Получение статистических характеристик данных, таких как среднее значение, стандартное отклонение, минимальное и максимальное значения.

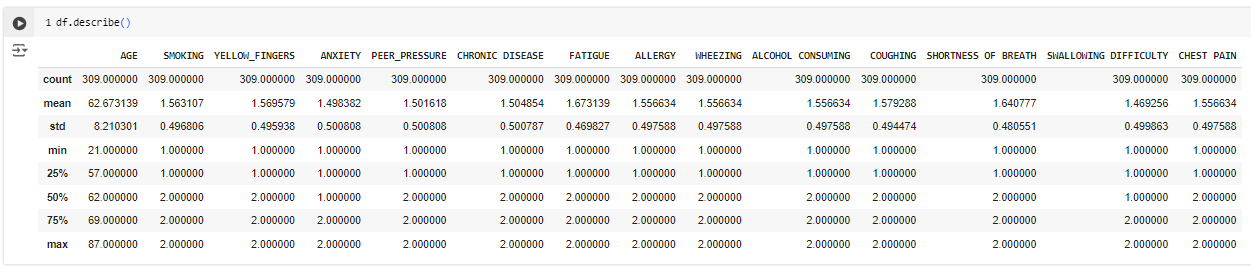


Рисунок 4 - Статистические характеристики данных

Определение размеров датасета (количество строк и столбцов).



Рисунок 5 - Размер датасета

Проверяем датасет на наличие пропущенных значений.



Рисунок 6 - Пропущенные значения

Проверяем датасет на наличие дубликатов.

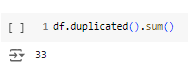


Рисунок 7 - Дубликаты

Удаляем найденные дубликаты в датасете.



Рисунок 8 - Удаление дубликатов

### Этап 2. Кодирование категориальных признаков

Категориальные признаки кодируются с помощью LabelEncoder для преобразования их в числовой формат, который может быть использован в моделях машинного обучения. Для этого импортируем библиотеку LabelEncoder. Применяем ее к каждому столбцу датасета.

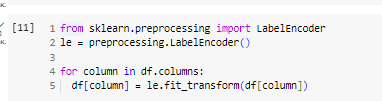


Рисунок 9 - Кодирование данных

Отображаем несколько строк датасета для проверки правильности кодирования. И видим, что данные закодировались в {0,1}, где 0 означает «нет», 1 означает «да», для пола «1» - мужской, «0» - женский.

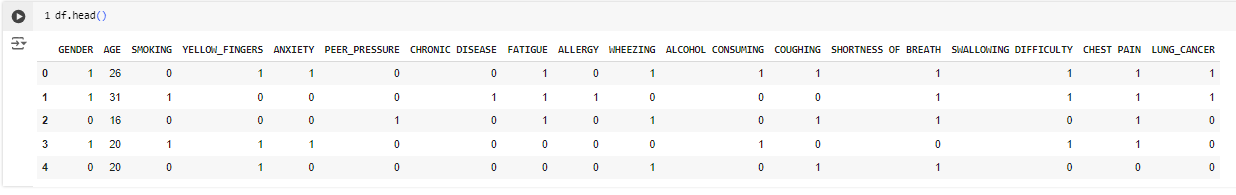


Рисунок 10 - Вывод датасета

### Этап 3. Визуализация данных

На этом этапе проводится визуализация данных для выявления зависимостей и распределений.

Корреляционная связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого [4].

Выводим тепловую карту для отображения корреляций между признаками.

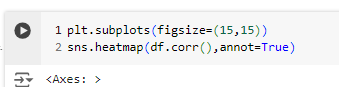


Рисунок 11 - Визуализация корреляции

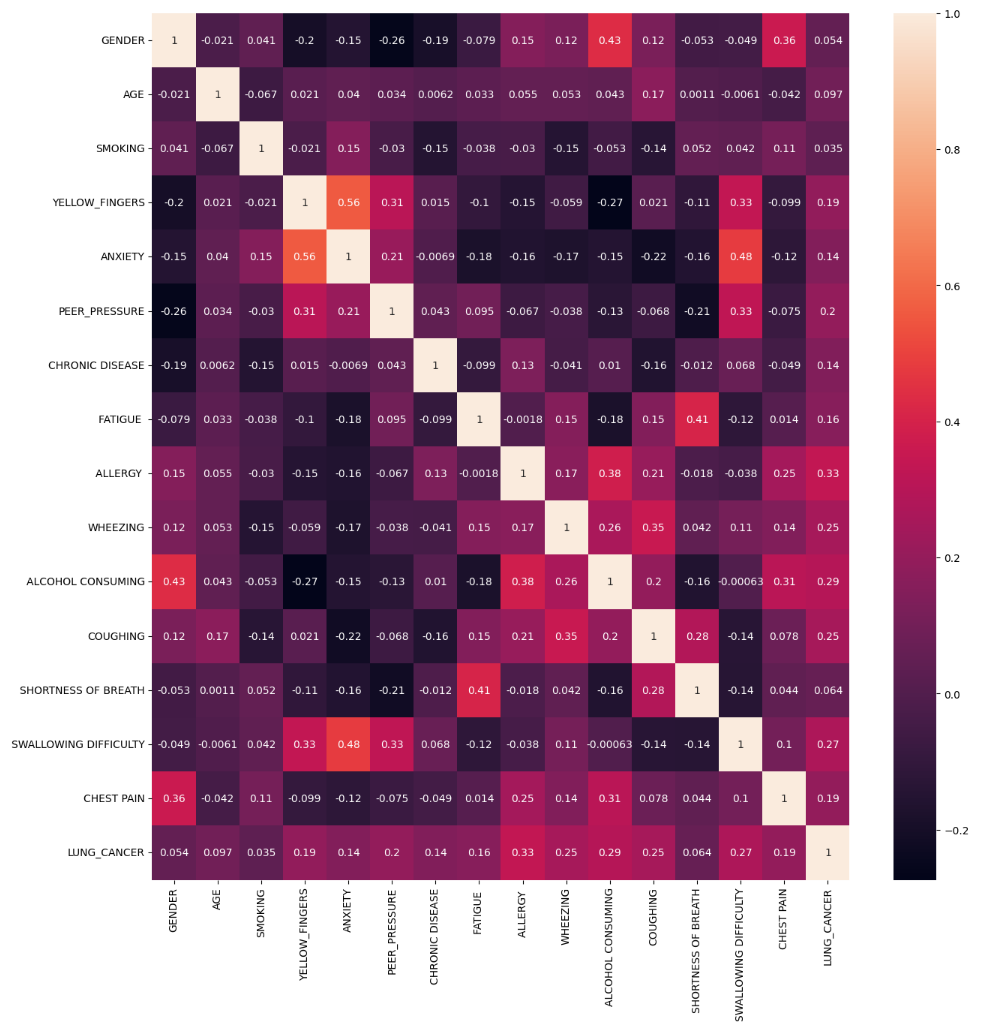


Рисунок 12 - Тепловая карта

Из рисунка 12 видно, что наибольшая корреляция между целевой переменной и признаком наблюдается у переменных «Allergy», «Wheezing», «Alcohol consuming», «Coughing», «Swallowing dificulty», «Chest pain».

Между различными признаками не так много сильных корреляций, за исключением, возможно, некоторых незначительных, таких как:

* усиление беспокойства приводит к затруднению дыхания
* повышенное беспокойство вызывает пожелтение пальцев или может быть наоборот, поскольку любые изменения в организме могут привести к усилению беспокойства
* одышка вызывает усталость
* повышенное беспокойство затрудняет глотание

Однако между симптомами и целевой переменной нет большой корреляции, и это может быть связано с типом данных или обследователей.

Используем sns.countplot для отображения распределения целевого признака (наличие или отсутствие рака легких).

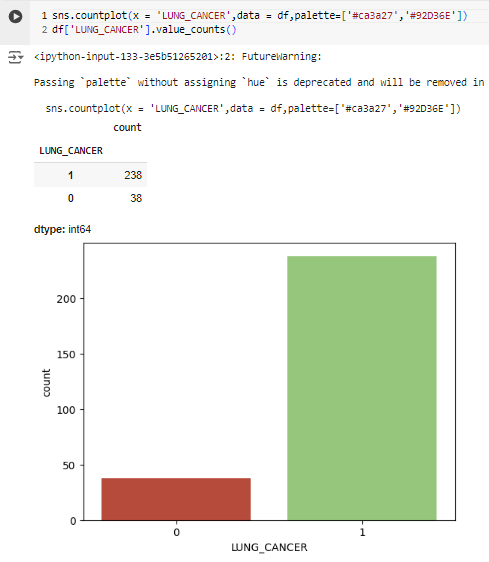


Рисунок 13 - Визуализация распределения целевого признака

Из рисунка 13 видно, что данные не сбалансированы и пациентов с заболеванием гораздо больше.

Для большего понимания количества пациентов с определенным признаком, воспользуемся циклом, в котором изобразим все симптомы и посчитаем с раком легких и без него. Код изображен на рисунке 14.

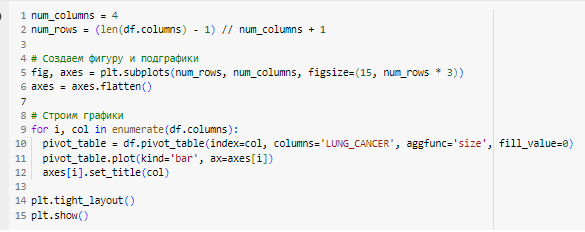


Рисунок 14 – Цикл отображения графиков

Отображение всех гистограмм с распределением пациентов.

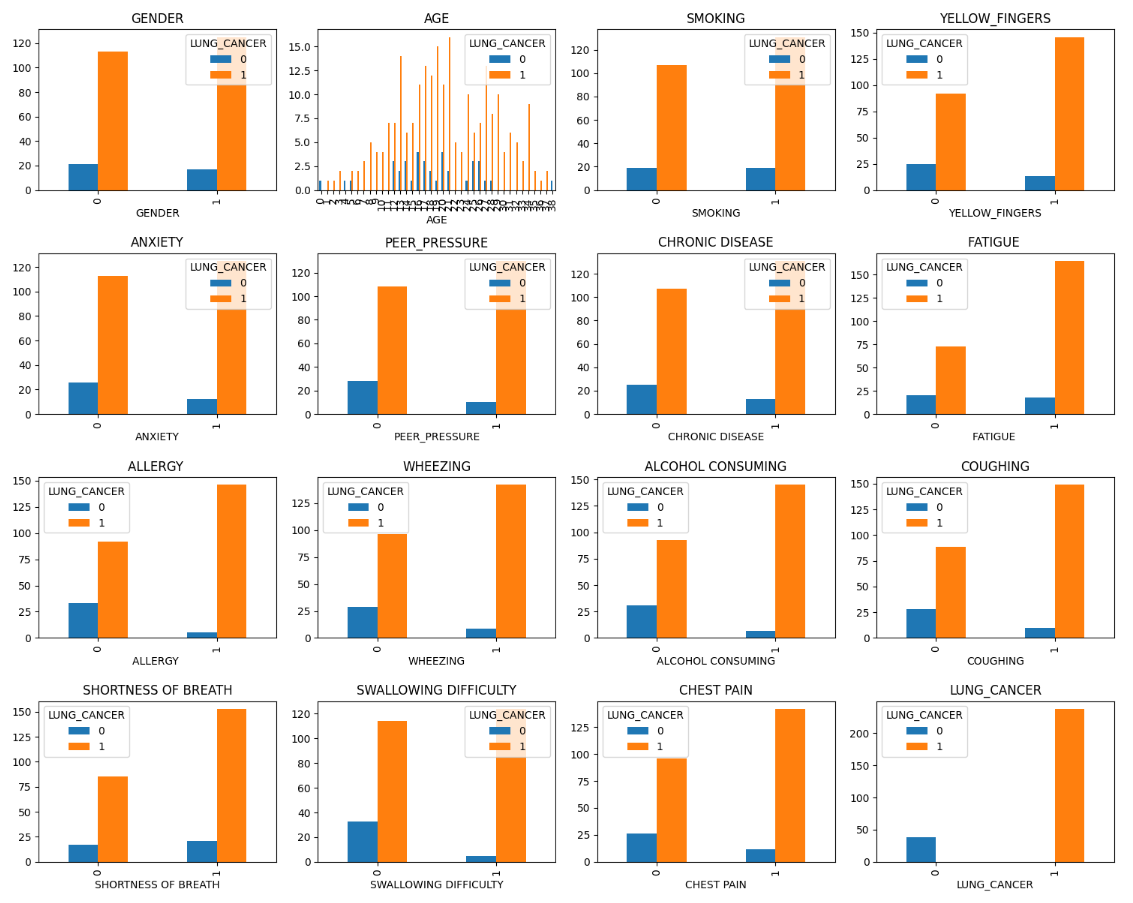


Рисунок 15 - Визуализация распределения признаков

Используем гистограммы для отображения распределения возраста пациентов с раком легких в зависимости от пола.

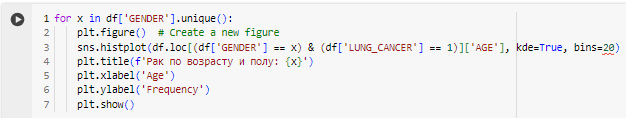


Рисунок 16 - Визуализация распределения возраста и пола

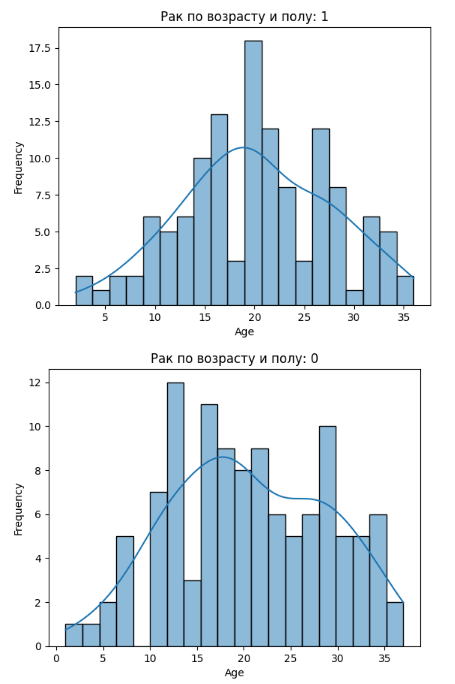


Рисунок 17 - Визуализация распределения возраста и пола

Гистограммы на рисунке 17 отображают, что мужчин, больных раком, больше, чем женщин.

### Этап 4. Разделение данных на обучающую и тестовую выборки и балансировка данных

Исходя из рисунка 13, поставим цель избавления от несбалансированности данных.

Несбалансированные данные – это одна из наиболее распространенных проблем, с которой сталкиваются в машинном обучении. Они возникают, когда классы, которые мы пытаемся предсказать, представлены неравномерно. Другими словами, один класс имеет гораздо меньше примеров, чем другой, что создает дисбаланс в распределении данных.

Для устранения дисбаланса классов применяется метод RandomOverSampler [3]. Но перед этим необходимо разделить данные на целевую переменную и остальные признаки.



Рисунок 18 - Разделение данных на признаки и целевой признак

Увеличение выборки заключается в добавлении дополнительных примеров класса-меньшинства, чтобы сделать его более представительным. Это можно сделать различными способами, например, путем дублирования существующих примеров, генерации синтетических данных или комбинирования этих методов.

Имортируем библиотеку RandomOverSampler для балансировки данных. Применяем этот метод к данным для увеличения количества примеров меньшего класса.

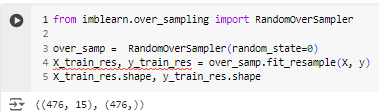


Рисунок 19 - Применение RandomOverSampler

Отображаем размеров данных после балансировки и визуализация распределения классов с помощью sns.countplot.

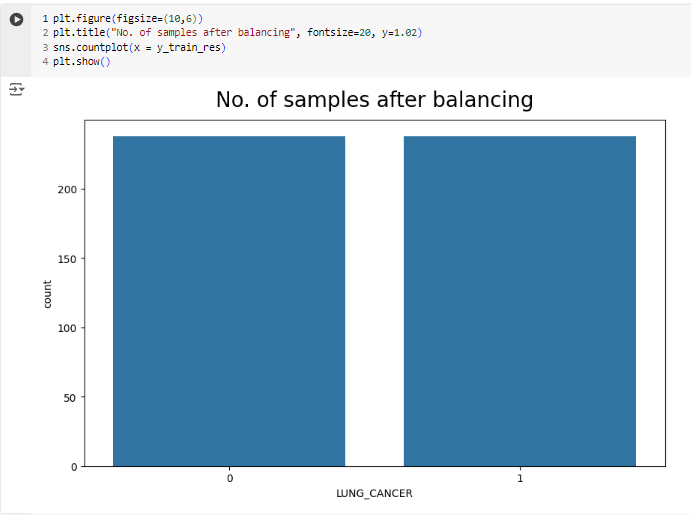


Рисунок 20 - Визуализация распределения классов

Для обучения моделей и проверки качества обучения необходимо разделить данные. Используем train\_test\_split для разделения данных на обучающую (80%) и тестовую (20%) выборки.



Рисунок 21 - Разделение данных на обучающую и тестовую выборки

### Этап 5. Моделирование и оценка моделей

На этом этапе разрабатываются и обучаются различные модели машинного обучения.

Алгоритмы, на которых обучались модели:

1. Логистическая регрессия.
2. Линейная регрессия.
3. Случайный лес.
4. K-ближайших соседей.
5. Дерево решений.
6. Градиентный бустинг.

Создаем и обучаем модели различными алгоритмами. Прогнозируем и оцениваем точность модели.

Оценивать модели будем через метод classification \_report. В данном методе имеются различные оценки качества модели.

Accuracy (точность) – отражает долю правильно спрогнозированных классов среди всех образцов.

Precision – характеризует долю правильно предсказанных положительных классов среди всех образцов, которые модель спрогнозировала как положительный класс.

Recall – отражает долю правильно предсказанных положительных классов среди всех реальных положительных образцов.

F1-score – представляет собой гармоническое среднее между Precision и Recall, обеспечивая между ними баланс, что особенно полезно при неравномерном распределении классов

Логистическая регрессия — это статистический метод, который позволяет прогнозировать вероятность наступления события и выдавать результат в виде числа от 0 до 1. Этот метод применяется, когда зависимая переменная имеет два возможных значения.

Логистическая регрессия используется для оценки вероятности наступления события. В отличие от обычной регрессии, она не предсказывает конкретное числовое значение на основе исходных данных, а определяет вероятность того, что исходное значение принадлежит к определённому классу.

Обучение модели Logistic Regression изображено на рисунке 22.



Рисунок 22 - Логистическая регрессия

Линейная регрессия — это один из наиболее распространённых методов машинного обучения, который позволяет прогнозировать значения зависимой переменной на основе независимых переменных.

Обучение модели Linear Regression изображено на рисунке 23.

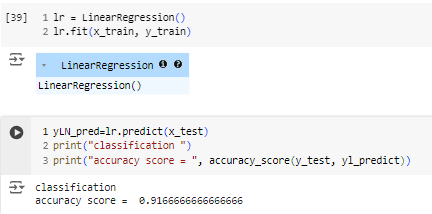


Рисунок 23 - Линейная регрессия

Метод classification\_report вывел хорошие показатели обучения у моделей Linear Regression и Logistic Regression.

Метод случайного леса — это мощный инструмент в области машинного обучения, который используется для задач классификации и регрессии. Он основан на объединении множества деревьев решений, что позволяет повысить точность и стабильность модели.

Этот метод стал одним из самых популярных благодаря своей простоте и эффективности. Он использует алгоритм случайного леса, который строится на ансамбле деревьев решений. Каждое дерево решений создаётся случайным образом, а затем результаты всех деревьев объединяются для получения более точного прогноза.

Такой подход позволяет создать модель, которая устойчива к переобучению и эффективна для работы с различными типами данных.

Обучение модели RandomForestClassifier.

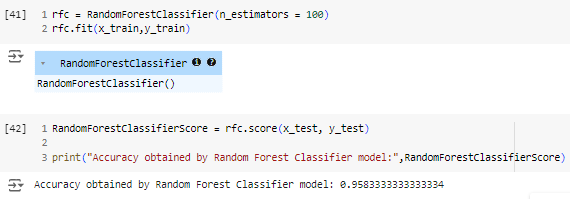


Рисунок 24 - Случайный лес

Для следующих моделей дополнительно выведем матрицу ошибок, чтобы посмотреть, как они справилась с обучением.

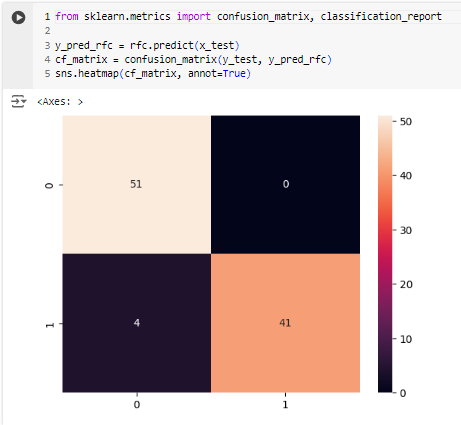


Рисунок 25 - Матрица ошибок, Случайный лес

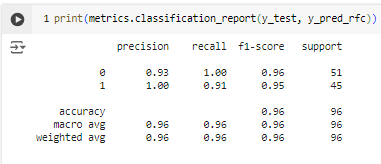


Рисунок 26 - Метрики, Случайный лес

Метод K ближайших соседей применяется для решения задач классификации и регрессии, но чаще всего используется для классификации. Суть метода в том, что объект относится к тому классу, к которому принадлежат его ближайшие соседи. Метод не требует обучения модели, что делает его привлекательным для начинающих специалистов.

KNN относится к алгоритмам, не требующим предварительной подготовки модели. Он сохраняет все данные, полученные в процессе обучения, и использует их для классификации новых объектов.

Обучение KNN модели изображено на рисунке 27.

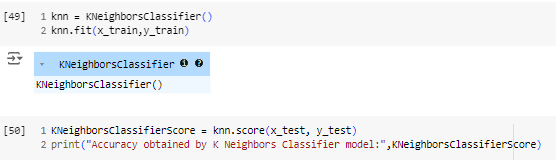


Рисунок 27 - k-ближайших соседей

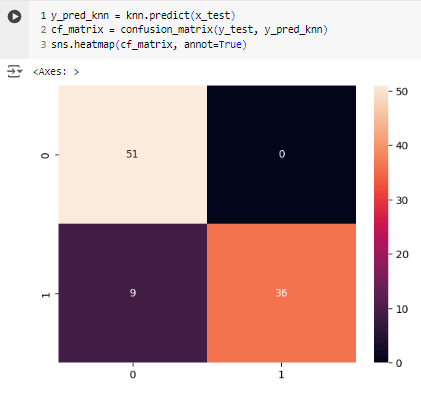


Рисунок 28 - Матрица ошибок, k-ближайших соседей

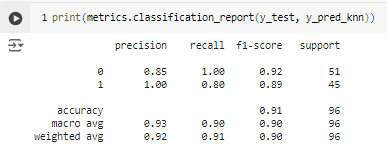


Рисунок 29 - Метрики, k-ближайших соседей

В основе дерева решений лежат узлы и ветви. Каждый узел представляет собой проверку условия, а ветви — это пути от узла в зависимости от результата проверки. В конечных узлах, которые также называют листьями, содержатся предсказания модели. Узлы могут быть внутренними или конечными, и каждый из них играет свою роль в процессе принятия решений. Внутренние узлы разделяют данные на подгруппы, а конечные узлы предоставляют окончательные предсказания.

Обучение модели дерева решений изображено на рисунке 30.

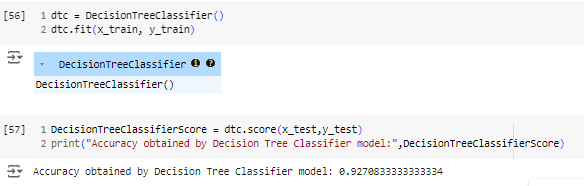


Рисунок 30 - Дерево решений

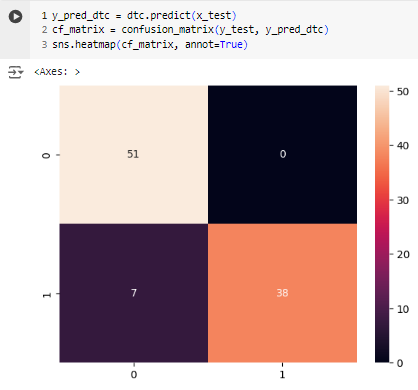


Рисунок 31 - Матрица ошибок, Дерево решений

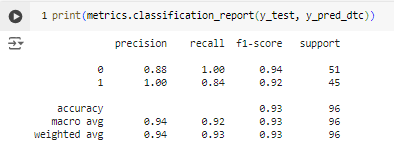


Рисунок 32 - Метрики, Дерево решений

Градиентный бустинг – это продвинутый алгоритм машинного обучения для решения задач классификации и регрессии. Он строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений. Из нескольких слабых моделей в итоге мы собираем одну, но уже эффективную. Общая идея алгоритма – последовательное применение предиктора (предсказателя) таким образом, что каждая последующая модель сводит ошибку предыдущей к минимуму.

Обучение данной модели приведено на рисунке 33.

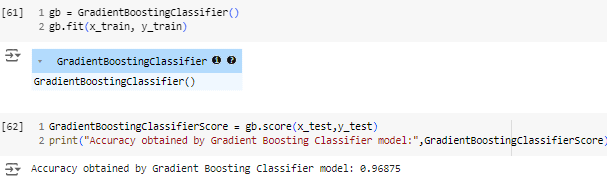


Рисунок 33 - Градиентный бустинг

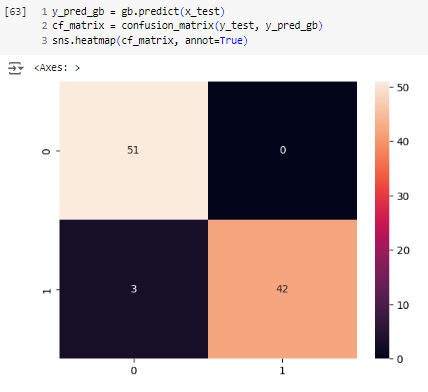


Рисунок 34 - Матрица ошибок, Градиентный бустинг

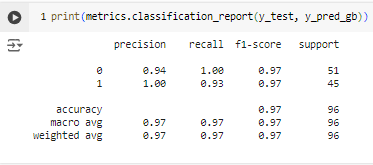


Рисунок 35 - Метрики, Градиентный бустинг

### Этап 6. Сравнение моделей

На этом этапе проводится сравнение точности различных моделей и визуализируются результаты для выбора наилучшей модели.

Создаем списки с названиями моделей и их точностями.



Рисунок 36 - Создание списка моделей и их точностей

Создаем столбчатые диаграммы, отображающие точность каждой модели.

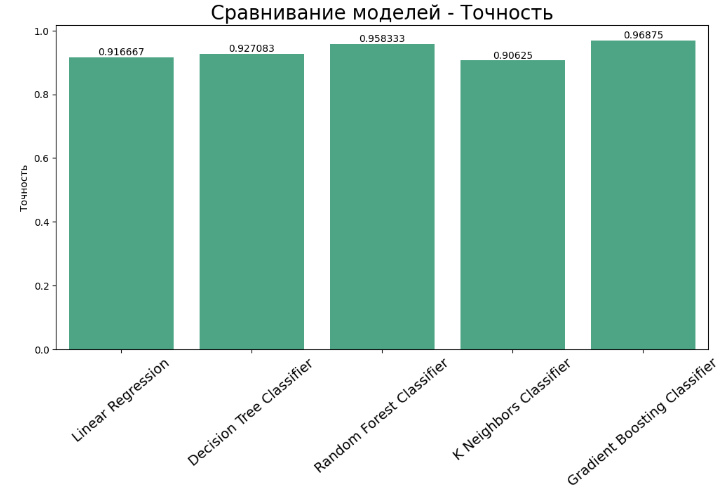


Рисунок 37 - Визуализация сравнения моделей

Как видно из диаграммы, наибольшую точность показала модель Gradient Boosting Classifier

# Заключение

На основе проведенного регрессионного анализа можно сделать вывод, что не все факторы влияют на возникновение рака легких. После построения корреляционной матрицы, наибольшее влияние на возникновение заболевания оказали такие факторы, как аллергия, одышка, употребление алкоголя, кашель, затрудненное глотание, боль в груди.

Таким образом, поставленная цель была достигнута. Был выполнен анализ факторов и построены различные модели, которые смогут предсказать заболевание пациентов. Самой лучшей моделью оказалась Gradient Boosting Classifier, она показала наилучшие метрики, среди остальных моделей.

Рекомендуется продолжить исследование с использованием более обширных и разнообразных данных для улучшения точности модели.

Включение дополнительных факторов, таких как генетические данные и результаты медицинских обследований, может повысить точность прогнозов.

# Библиографический список

1. Козлов В. В., Клинические и молекулярно-генетические особенности немелкоклеточного рака легкого и их значение для диагностики и лечения курящих и некурящих пациентов / Диссертация – Режим доступа: <https://frcftm.ru/wp-content/uploads/2023/06/dissertacija_kozlov-vv.pdf>
2. Борьба с несбалансированными данными [Электронный ресурс]. URL: <https://habr.com/ru/companies/otus/articles/769242/> (дата обращения: 18.12.2024)
3. Теоретический материал, корреляционный анализ [Электронный ресурс] URL: <https://e.vyatsu.ru/pluginfile.php/462616/mod_resource/content/3/Теоретический%20материал_корреляционный%20анализ.pdf> (дата обращения 18.12.2024)