|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_ СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_Определение диабета по данным\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_\_ ИУ5-61Б\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ Федотов А.Н.\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_Гапанюк Ю. Е.\_\_**

(Подпись, дата) (И.О.Фамилия)

*Москва, 2023 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени**

**Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Определение диабета по данным\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_ИУ5-64Б\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Федотов Александр Никитич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

Техническое задание Исследовать использование методов машинного обучения для решения задачи определения наличия у человека диабета по определённому перечню даныых\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_18\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «07» февраля 2023 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_**Гапанюк Ю.Е.**\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_ Федотов А.Н.\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc131035556)

[1 Постановка задачи 5](#_Toc131035557)

[2 Описание данных, используемых методов и метрик 5](#_Toc131035558)

[3 Выполнение примера по варианту 8](#_Toc131035559)

[ЗАКЛЮЧЕНИЕ 17](#_Toc131035560)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 18](#_Toc131035561)

# **ВВЕДЕНИЕ**

Предсказание и определение наличия у человека диабета является крайне важной задачей, так как раннее определение по простейшим признакам может дать человеку возможность заранее пойти к врачу и вылечить данное заболевание или же избежать тяжелых последствий, к которым может привести незнание о столь опасном заболевании.

Нам предстоит использовать набор данных связанный с диабетом для создания алгоритма, который предсказывает наличие диабета по входным данным.

Цель данной работы заключается в создании алгоритма для определения диабета. Основная цель - достижение желаемого результата, учитывая ограниченность вычислительных мощностей.

# **Постановка задачи**

Диабет – это хроническое заболевание, которое возникает либо в случаях, когда поджелудочная железа не вырабатывает достаточное количество инсулина, либо когда организм не может эффективно использовать вырабатываемый инсулин. Инсулин – это гормон, регулирующий уровень глюкозы в крови. Распространенным следствием неконтролируемого диабета является гипергликемия - или повышенный уровень содержания глюкозы (сахара) в крови, - со временем приводящая к серьезному повреждению многих систем организма, особенно нервной системы и кровеносных сосудов.

В 2014 г. заболеваемость диабетом среди взрослого населения в возрасте 18 лет и старше составляла 8,5%. В 2019 г. диабет стал непосредственной причиной 1,5 миллиона случаев смерти, и 48% всех связанных с диабетом случаев смерти приходится на людей в возрасте до 70 лет. Еще 460 000 случаев смерти были вызваны заболеваниями почек, обусловленными диабетом; кроме того, повышенный уровень глюкозы в крови является причиной около 20% смертей от сердечно-сосудистых заболеваний *(1)*.

В период с 2000 по 2019 г. стандартизированный по возрасту коэффициент смертности от диабета вырос на 3%. В странах с доходом ниже среднего уровня коэффициент смертности от диабета повысился на 13%.

Напротив, вероятность смерти от любого из четырех основных неинфекционных заболеваний (сердечно-сосудистые заболевания, рак, хронические респираторные заболевания или диабет) в возрасте от 30 до 70 лет в период с 2000 по 2019 г. снизилась на 22% во всем мире.

Именно поэтому раннее определение диабета может помочь большому количеству людей во всем мире и дать возможность на ранних этапах диагностировать заболевание, а так же заранее предпринять действия.

# **Описание данных, используемых методов и метрик**

Используемый набор данных содержит 100000 записей, но в силу ограниченных вычислительных мощностей будем использовать только 5000 записей. Каждая запись содержит данные:

1. gender - пол
2. age – возраст
3. hypertension - наличие гипертонии
4. heart\_disease – сердечные заболевания
5. smoking\_history – история курения
6. bmi - давление
7. HbA1c\_level - уровень HbA1c
8. blood\_glucose\_level – уровень глюкозы в крови
9. diabetes – наличие диабета у испытуемого

Для обучения модели, которая будет определять диабет, необходимо решить следующие задачи:

1. Собрать и подготовить достаточное количество данных для построения модели машинного обучения.

2. Выбор архитектуры модели: для определения наличия диабета можно использовать различные модели. Такие как логическая регрессия, KNN, решающее дерево, случайный лес, градиентный бустинг. Их мы и будем использовать.

3. Обучение модели: после выбора архитектуры модели необходимо провести ее обучение на подготовленных данных.

4. Оценка качества модели: после обучения модели необходимо оценить ее качество на тестовых данных. Для этого используются:

* Метрика accuracy

Метрика accuracy в распознании речи показывает долю правильно распознанных людей с диабетом от общего количества распознанных людей. Эта метрика является одной из основных метрик для оценки качества модели в задачах классификации.

* Бинарная классификация

Является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение AUC, тем «лучше» модель классификации. Данный показатель часто используется для сравнительного анализа нескольких моделей классификации.

* Recall

Recall измеряет способность модели обнаруживать выборки, относящиеся к классу Positive. Чем выше recall, тем больше Positive семплов было найдено Python

* F1-мера

Для реализации работы был выбран язык программирования Python, так как данный язык имеет большое количество сторонних библиотек для разработки нейронных сетей и анализа данных. Он является самым передовым языком в области машинного обучения, при этом предоставляет простой синтаксис для работы.

* Pandas

Pandas - программная библиотека на языке Python для обработки и анализа данных. С ее помощью производилась обработка данных о звуковых файлах и производился анализ информации о наборе данных.

* NumPy

NumPy - библиотека с открытым исходным кодом для языка программирования Python.

Данная библиотека применялась для предобработки звуковых файлов, в частности для аугментации, а также был использован контейнер npy, который позволил загружать и обрабатывать данные в несколько раз быстрее, чем стандартные cvs и mp3, за счет чего было ускорено обучение модели.

* Sklearn

Это набор готовых решений для реализации поставленной цели

# **Выполнение работы**

Импортируем необходимые модули и зависимости. (рисунок 3.1)

# 

Рисунок 3.1 – импорт библиотек

Загрузим данные и выведем некоторые данные о наборе данных для понимания. (Рис. 3.2)

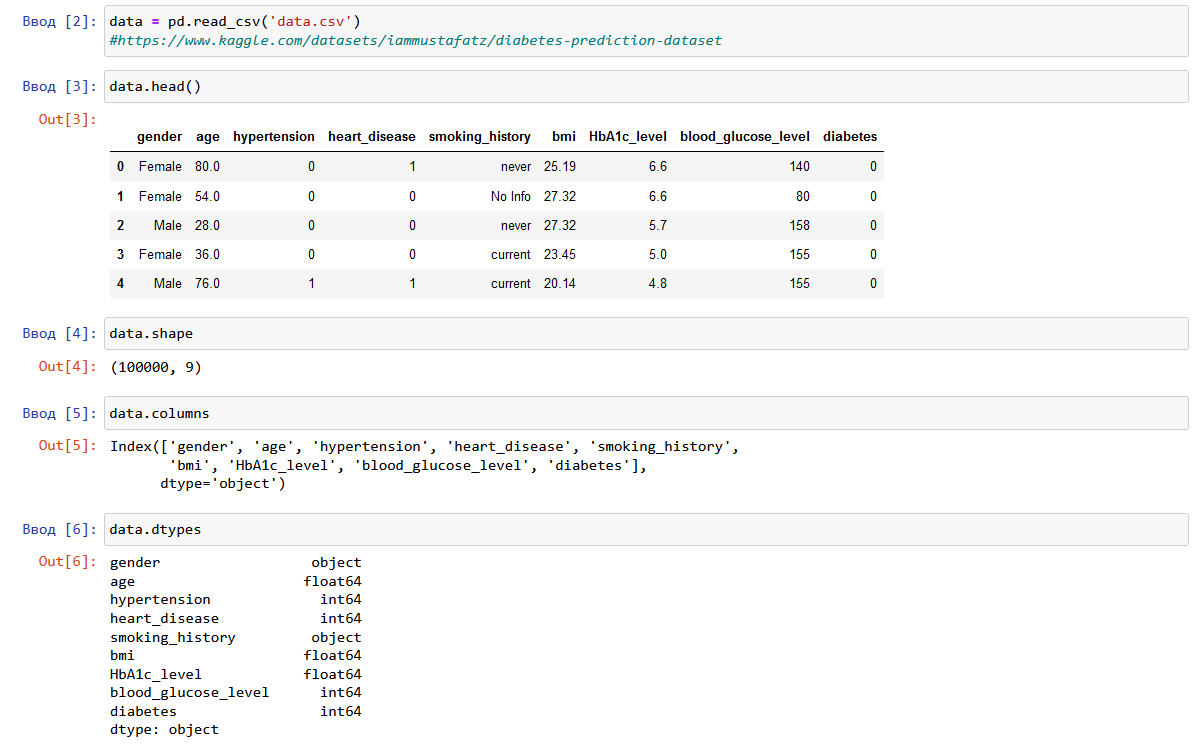


Рисунок 3.2 – выгрузка данных.

Проверим наличие пропуском и уменьшим размер выборки, а так же категорируем некоторые признаки. (Рис 3.3)



Рисунок 3.3

Начнем предварительный анализ данных. (Рис 3.4.)

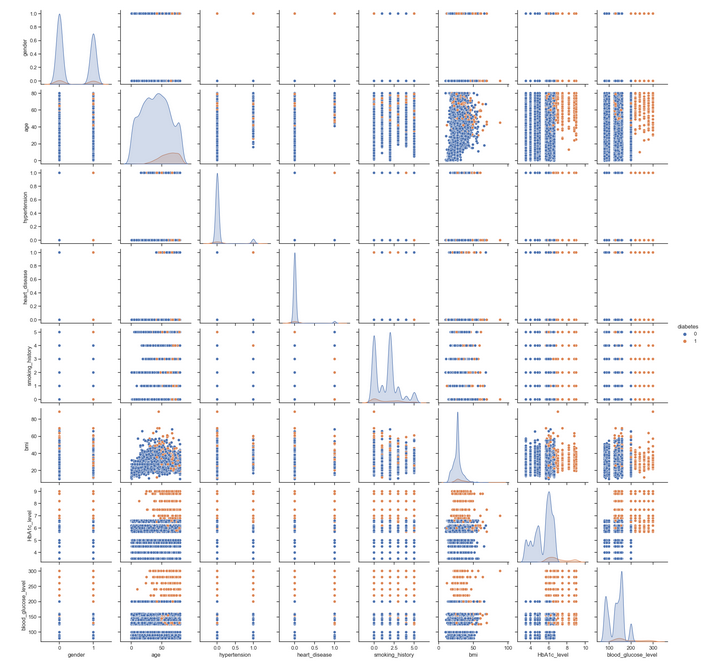


Рисунок 3.4

Дисбаланс классов (Рис 3.5) связан с тем, что число людей с диабетом не так велико. У 91 процента опрашиваемых нет диабета.

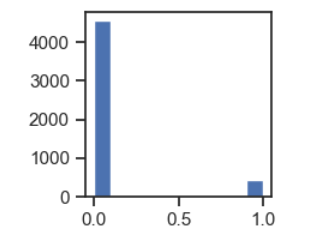


Рис 3.5

Построим скрипичные диаграммы для понимания распределения значений.(Рис 3.6, 3.7, 3.8)

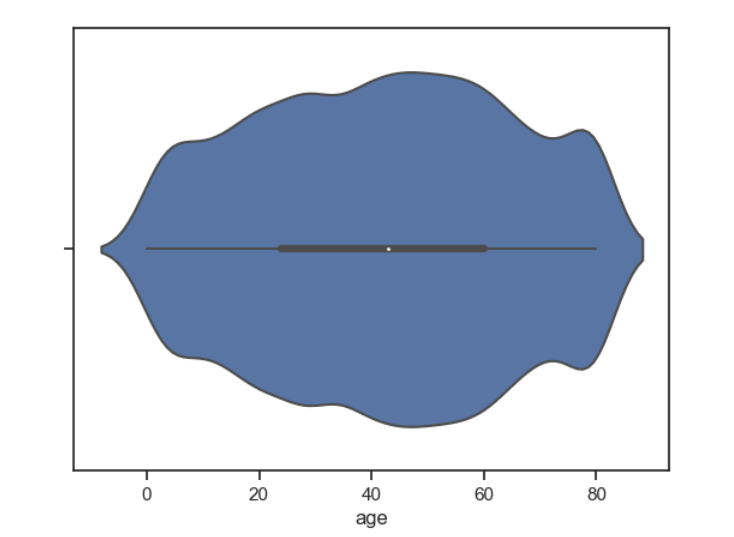


Рис 3.6

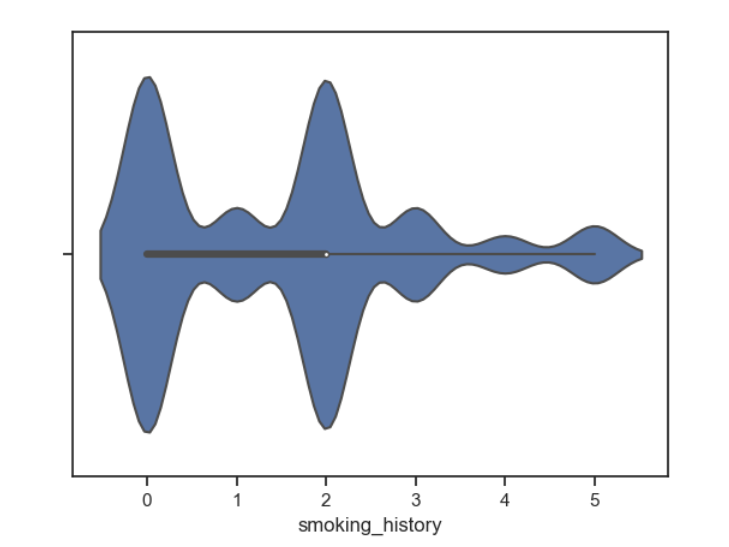


Рис 3.7

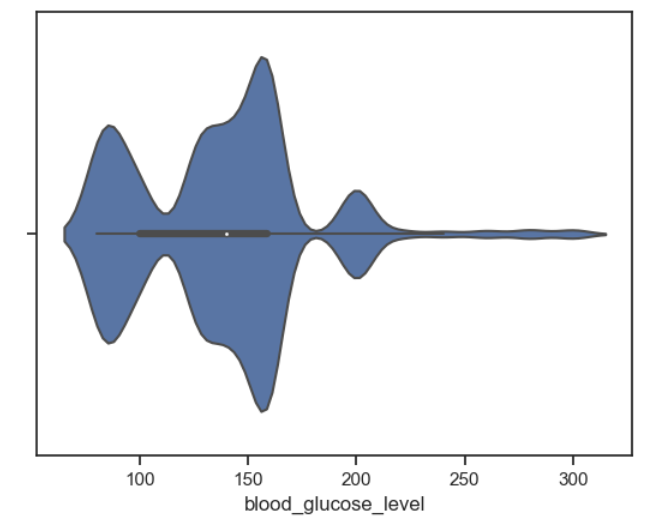


Рис 3.8

Построим корреляционную матрицу для понимания от каких признаков можно избавиться. (рис 3.9)

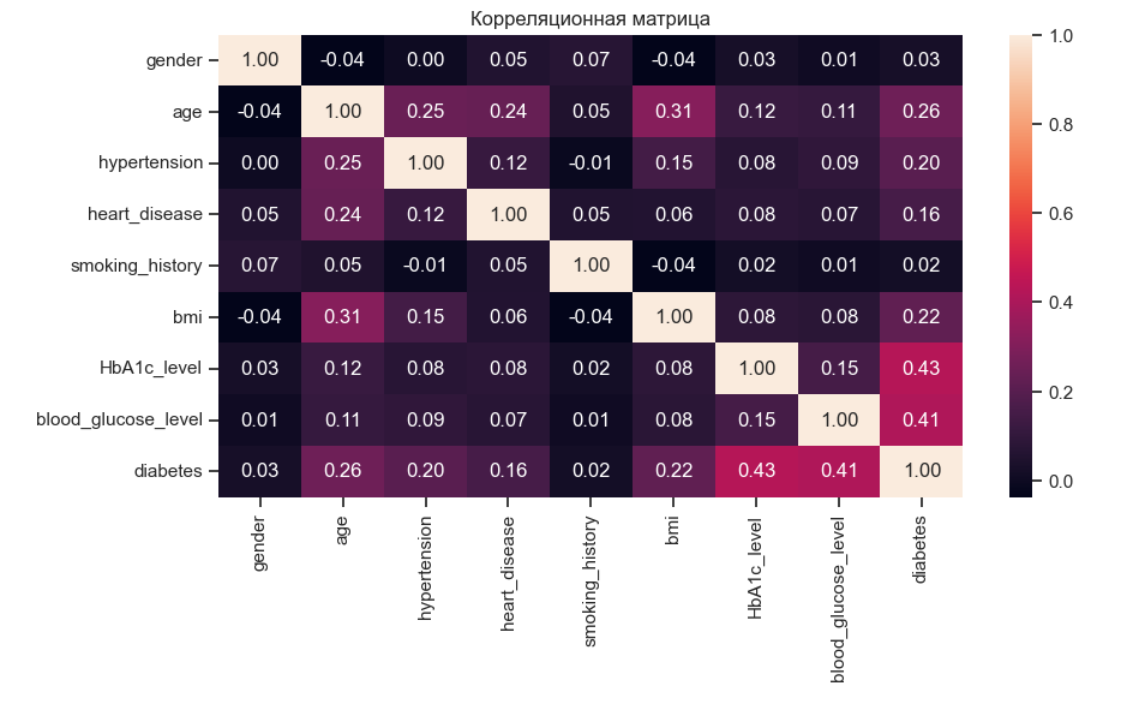


Рис 3.9

Исходя из данных корреляционной матрицы избавимся от истории курения и гендера, так как они не оказывают значимого влияния на наличие диабета

Разделим дата сет на выборки для обучения и тестирования моделей. (Рис 3.10)

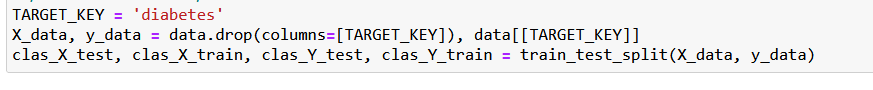


Рис 3.10

Создадим, обучим и протестируем модели. (рис 3.11, 3.12, 3.13, 3.14,3.15)

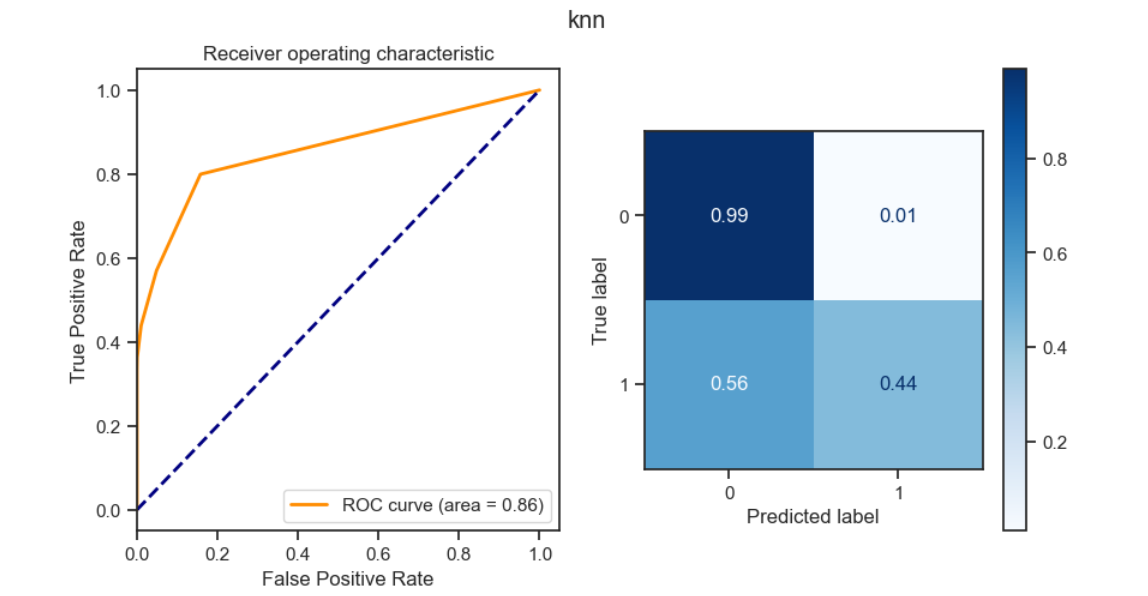


Рис 3.11

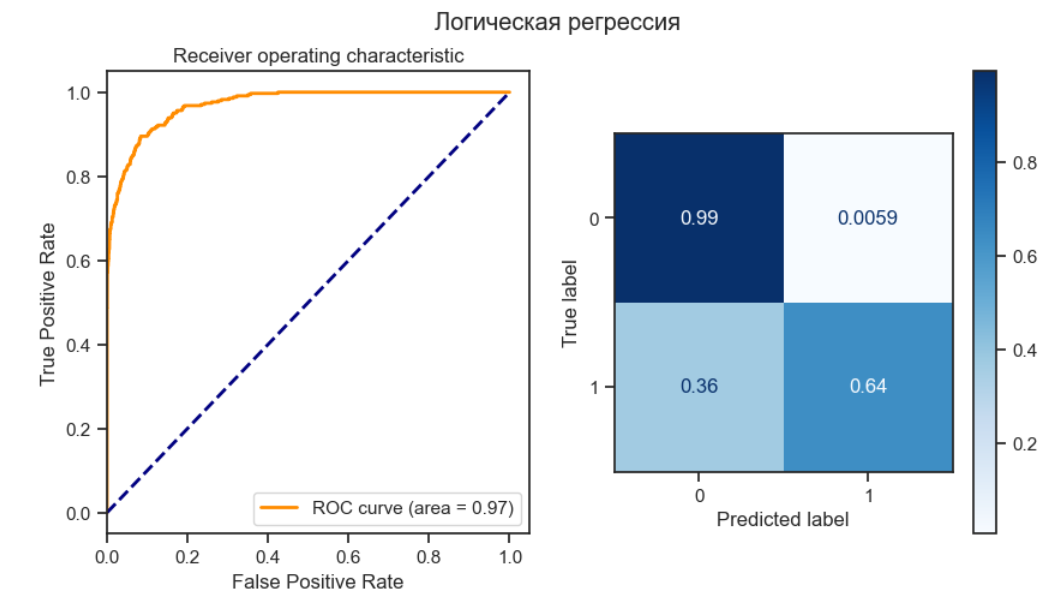


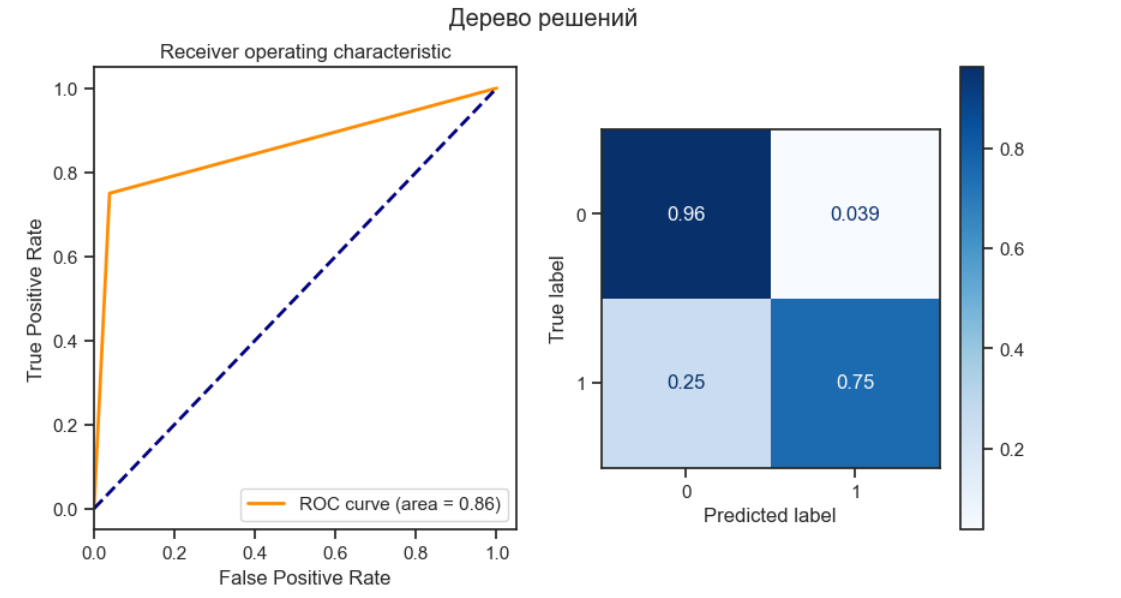
Рис 3.12

Рис 3.13

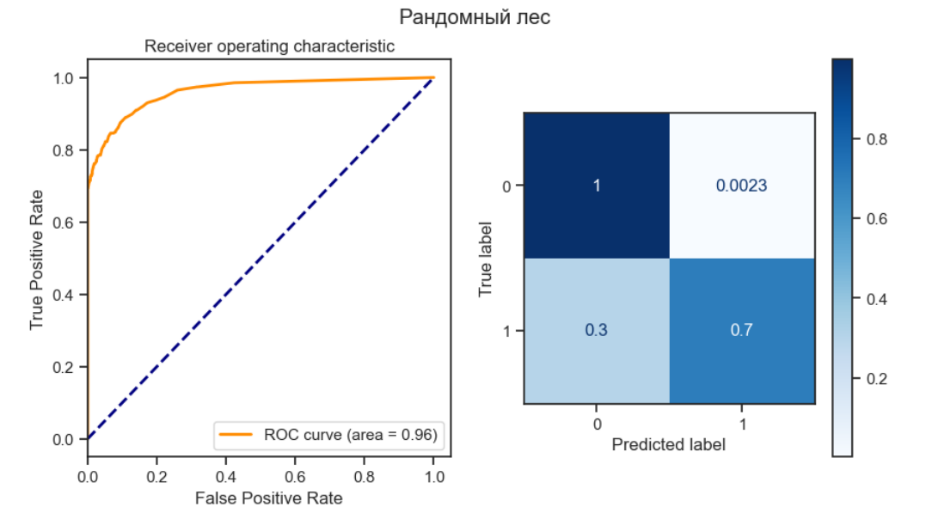


Рис 3.14

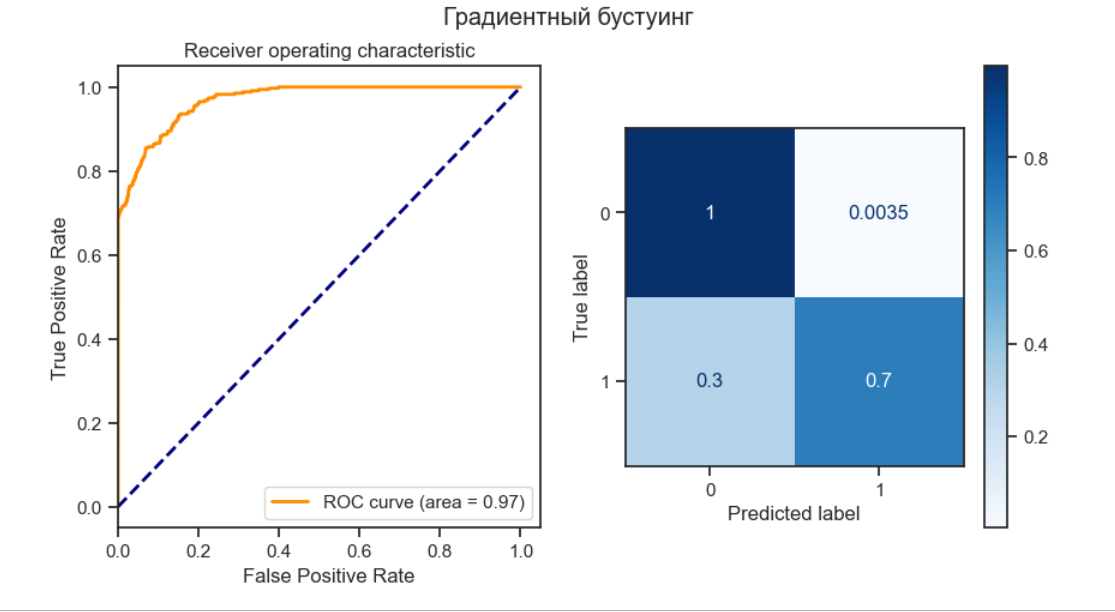


Рис 3.15

Подберем значение гиперпараметра для KNN и протестируем модель. (рис 3.16, 3.17)

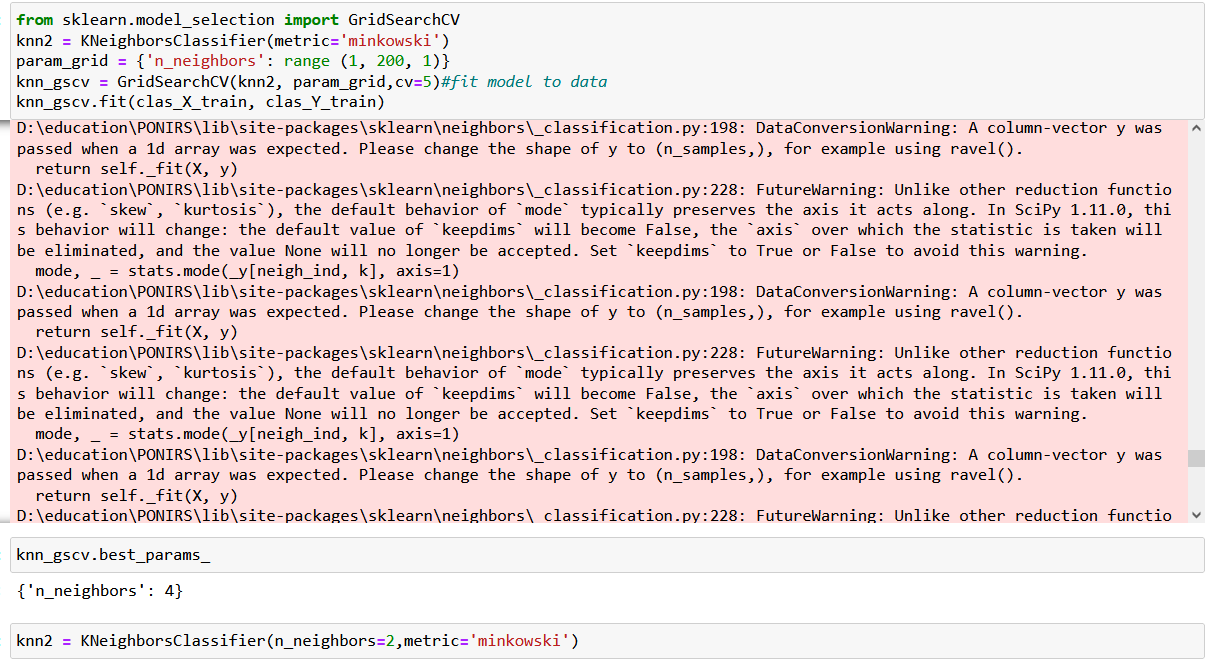


Рис 3.16

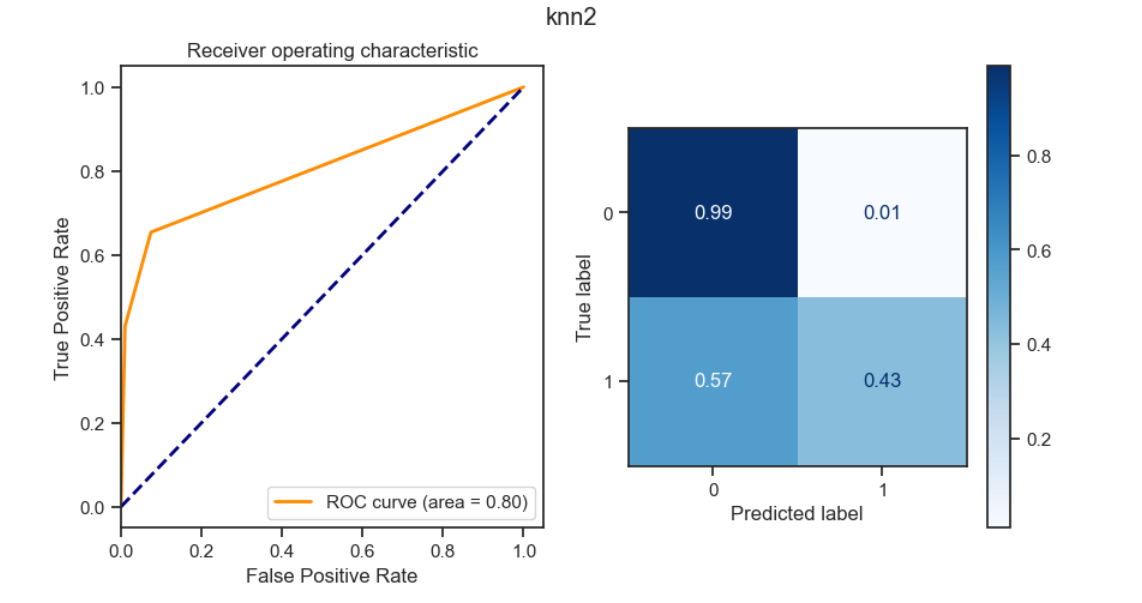


Рис 3.17

Посчитаем точность каждой модели (рис 3.18)

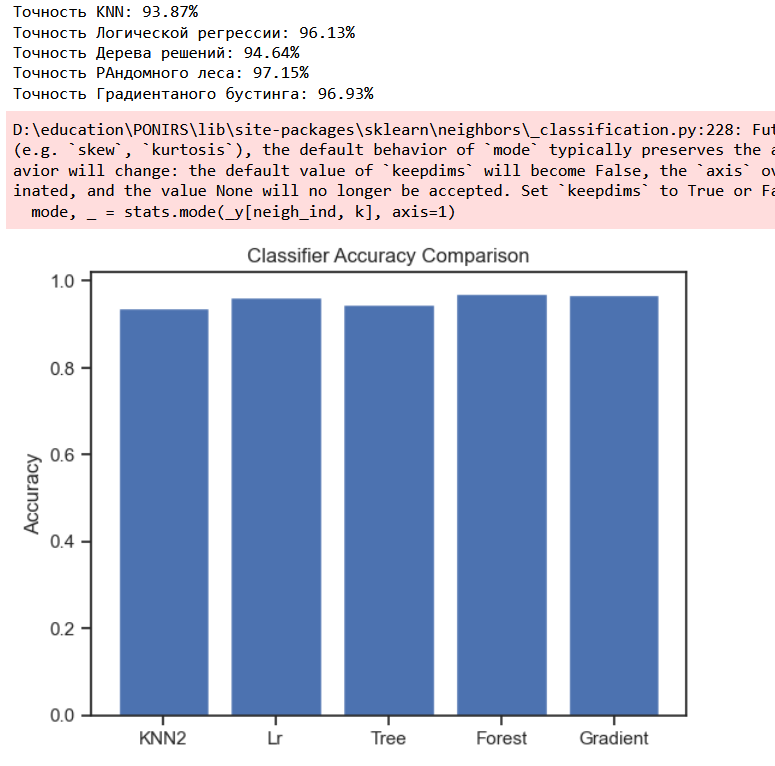


Рис 3.18

Определим значимость признаков. (рис 3.19)

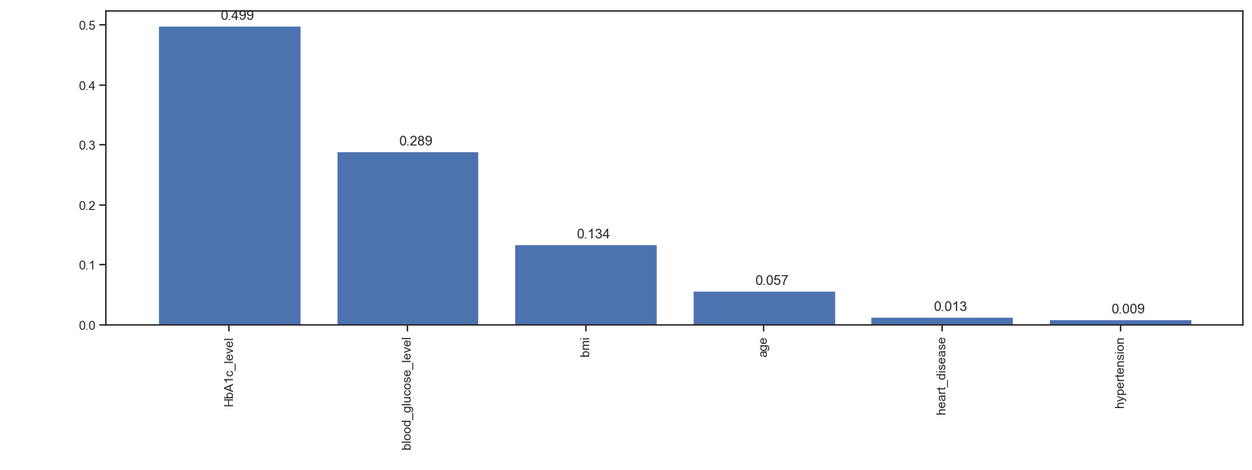


Рис 3.19

Выведем дерево решений. (рис 3.20)

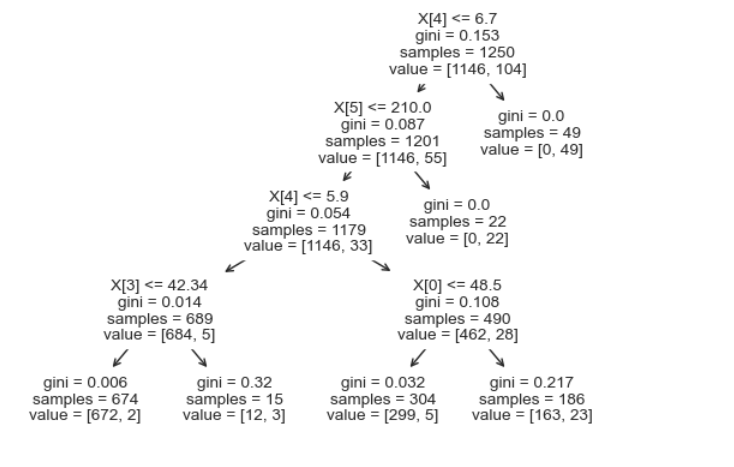


Рис 3.20

# **ЗАКЛЮЧЕНИЕ**

В результате выполнения научно-исследовательской работы были выполнена следующая задача:

1. C использованием методов машинного обучения была обучена модель, которая с точность 97,15% распознает наличие диабета.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. https://spec-zone.ru/scikit\_learn
2. Курс ИУ5
3. Статьи <https://www.bazhenov.me/>
4. https://neerc.ifmo.ru