



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

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Предсказание наличия

у пациентов диабета

Студент ИУ5-63Б
(Группа)

Д.С. Цуприков
(Подпись, дата) (И.О.Фамилия)

Руководитель

Ю.Е. Гапанюк
(Подпись, дата) (И.О.Фамилия)

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)
В.И. Терехов
(И.О.Фамилия)
« 07 » февраля 2024 г.

ЗАДАНИЕ
на выполнение научно-исследовательской работы

по теме Предсказание наличия у пациентов диабета

Студент группы ИУ5-63Б

Цуприков Дмитрий Сергеевич
(Фамилия, имя, отчество)

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

ИССЛЕДОВАТЕЛЬСКАЯ

Источник тематики (кафедра, предприятие, НИР) КАФЕДРА

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

Техническое задание Исследовать методы машинного обучения для решения задачи классификации

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 23 листах формата А4.

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

Дата выдачи задания « 07 » февраля 2024 г.

Руководитель НИР

Ю.Е. Гапанюк
(Подпись, дата) (И.О.Фамилия)

Студент

Д.С.Цуприков
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание	
Введение	4
Постановка задачи	6
Выполнение работы	7
Заключение	22
Список использованной литературы	23

Введение

Проблема неверной постановки диагноза очень актуальна в наше время. Врачи всячески пытаются решить эту проблему, заканчивая дополнительные курсы и улучшая техническое оборудование. Однако, можно своевременно предсказать предрасположенность пациента к тому или иному заболеванию, например, сахарному диабету, и внимательно отслеживать курс его лечения.

В данной работе мы будем использовать данные, полученные из отчетов института медицины о диагностике пациентов, чтобы построить модель машинного обучения, которая сможет предсказывать вероятность заболевания раком. Мы будем использовать алгоритмы классификации для определения факторов риска, включая количество беременностей, концентрацию глюкозы в плазме крови через 2 часа при пероральном тесте на толерантность к глюкозе, диастолическое артериальное давление, толщину кожной складки на трицепсе, 2-часовую дозу сывороточного инсулина, индекс массы тела, наследственную функцию диабета, возраст пациента и параметр отслеживания вероятности наличия диабета.

Целью данной работы является разработка эффективной модели, которая может помочь работникам медицинских центров быстро и точно определить вероятность обнаружения у пациента диабета и принять меры для минимизации врачебных ошибок и ускоренного процесса лечения.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.

6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

Постановка задачи

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание наличия у пациента сахарного диабета.

В качестве набора данных используется набор, собранный в индийском Национальном институте диабета, заболеваний органов пищеварения и почек, который включает в себя информацию о таких измерениях, как количество беременностей, концентрация глюкозы в плазме крови, диастолическое артериальное давление, толщина кожной складки на трицепсе, доза сывороточного инсулина, индекс массы тела, наследственная функция диабета, возраст в годах. Каждый пациент может быть классифицирован как склонный к наличию диабета, так и не являющийся носителем данного заболевания.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания риска наличия у пациента диабета на основании его диагностических измерений. Для этого мы будем использовать различные алгоритмы классификации, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет предсказывать вероятность наличия или отсутствия у пациентов диабета, и помочь работникам медцентра оптимизировать диагностику и курс лечения для ускоренного осмотра пациента и выявления опасного заболевания.

Выполнение работы

Для решения задачи классификации был выбран набор данных, содержащий информацию о женщинах-пациентах.

В наборе данных присутствуют следующие столбцы:

- Pregnancies: количество беременностей;
- Glucose: концентрация глюкозы в плазме крови через 2 часа при пероральном тесте на толерантность к глюкозе;
- BloodPressure: диастолическое артериальное давление (мм рт.ст.);
- SkinThickness: толщина кожной складки на трицепсе (мм);
- Insulin: 2-часовая доза сывороточного инсулина (мкме/мл);
- BMI: индекс массы тела (вес в кг/(рост в м)²);
- DiabetesPedigreeFunction: наследственная функция диабета;
- Age: возраст в годах;
- Outcome: целевая переменная, где 1 означает, что у пациента был обнаружен сахарный диабет, а 0 - что он не был обнаружен.

Данный датасет использован для решения задачи классификации - предсказания наличия у пациента сахарного диабета, основываясь на определенных диагностических измерениях, включенных в набор данных.

Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 768 строк и 9 столбцов, из которых 7 типа int64 и 2 типа float64.

Наличие нуля в полях столбцов, исключая Pregnancies и Outcome, свидетельствует о пропусках в исходном наборе данных. Поэтому исключаем столбец Insulin из датасета как столбец с наибольшим количеством пропущенных значений, а пропуски в других колонках (BloodPressure, Glucose, SkinThickness, BMI) заменяем модой.

Дубликаты в наборе данных отсутствуют; кодирование категориальных признаков проводить незачем, ведь такого типа данных нет в исходном наборе.

Строим график pairplot для визуализации распределения данных попарно для множества колонок.

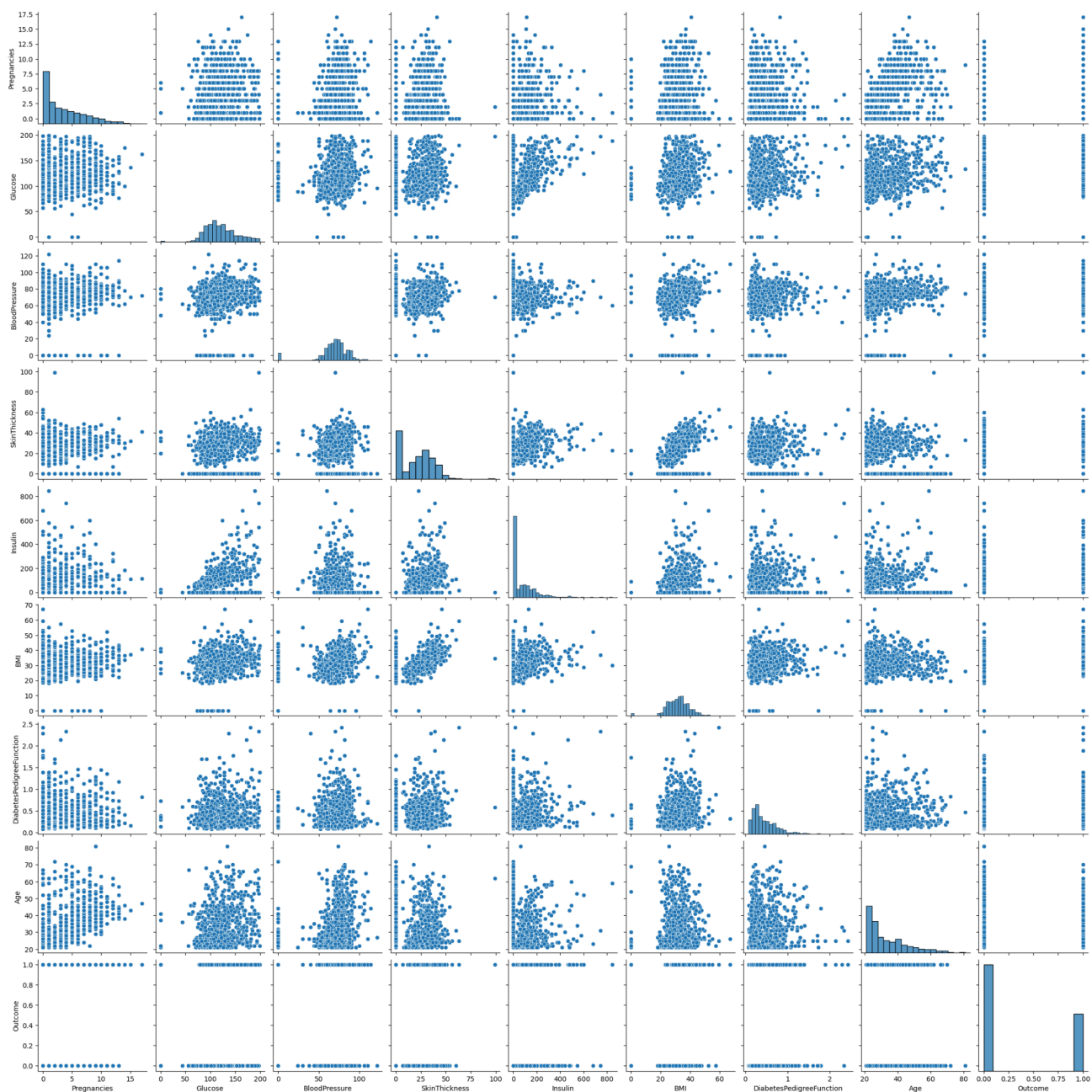


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующую гистограмму:

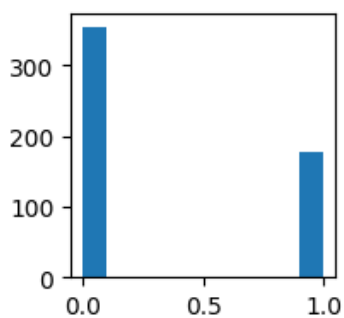


Рисунок 2 - Гистограмма классов

Видим, что классы немного не сбалансированы.

Строим таблицу средних значений с группировкой по целевому признаку и делаем следующие предположения:

- большее количество беременностей связано с повышенной вероятностью диабета;
- более высокие уровни концентрации глюкозы в плазме крови связаны с повышенной вероятностью наличия диабета;
- с нарастанием толщины кожной складки на трицепсе увеличивается предрасположенность к диабету.

Подтвердим наши предположения графиками.

Построим гистограмму зависимости количества беременностей от вероятности наличия диабета и проверим t-статистику.

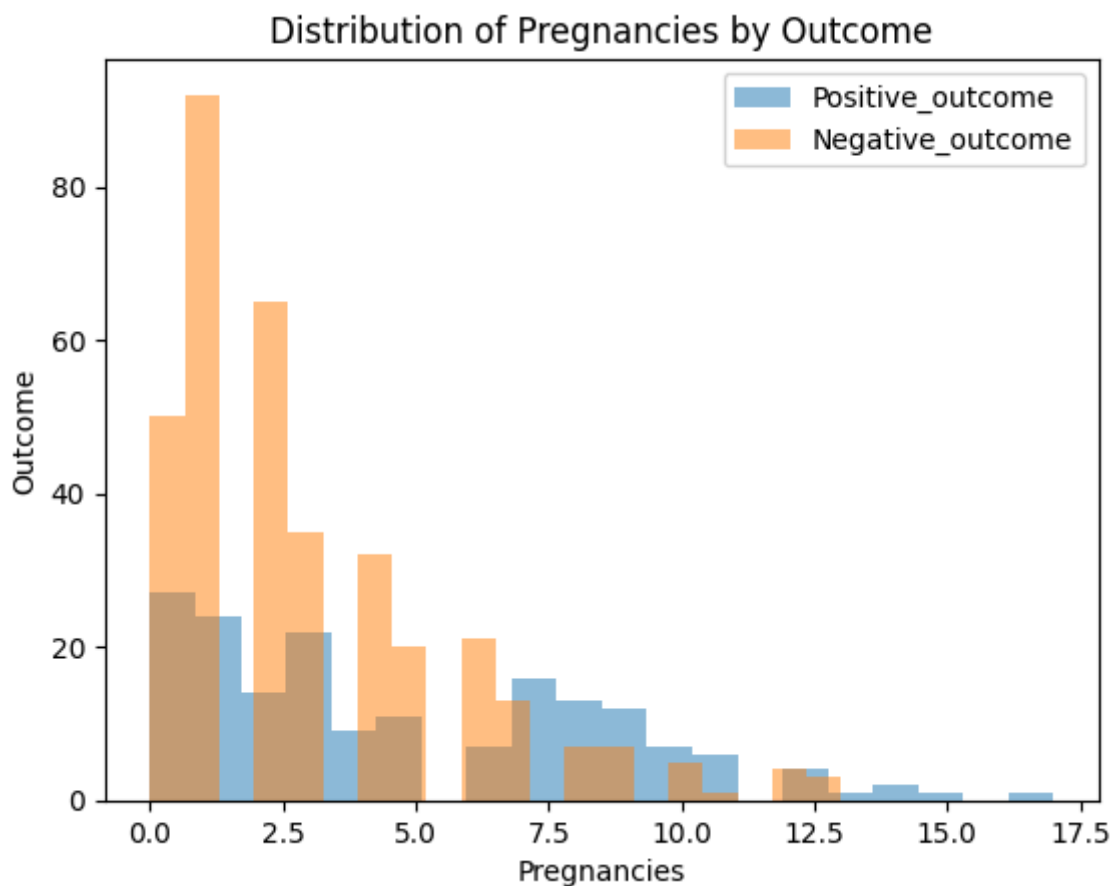


Рисунок 3 - Гистограмма зависимости количества беременностей от целевого признака

Можно заметить, что чем больше пациент перенес беременностей, тем значимее перевес положительного теста на диабет над отрицательным.

t-statistic: 6.009826316978722

p-value: 3.4569736448729003e-09

Построим гистограмму зависимости концентрации глюкозы в плазме крови от вероятности наличия диабета и проверим t-статистику.

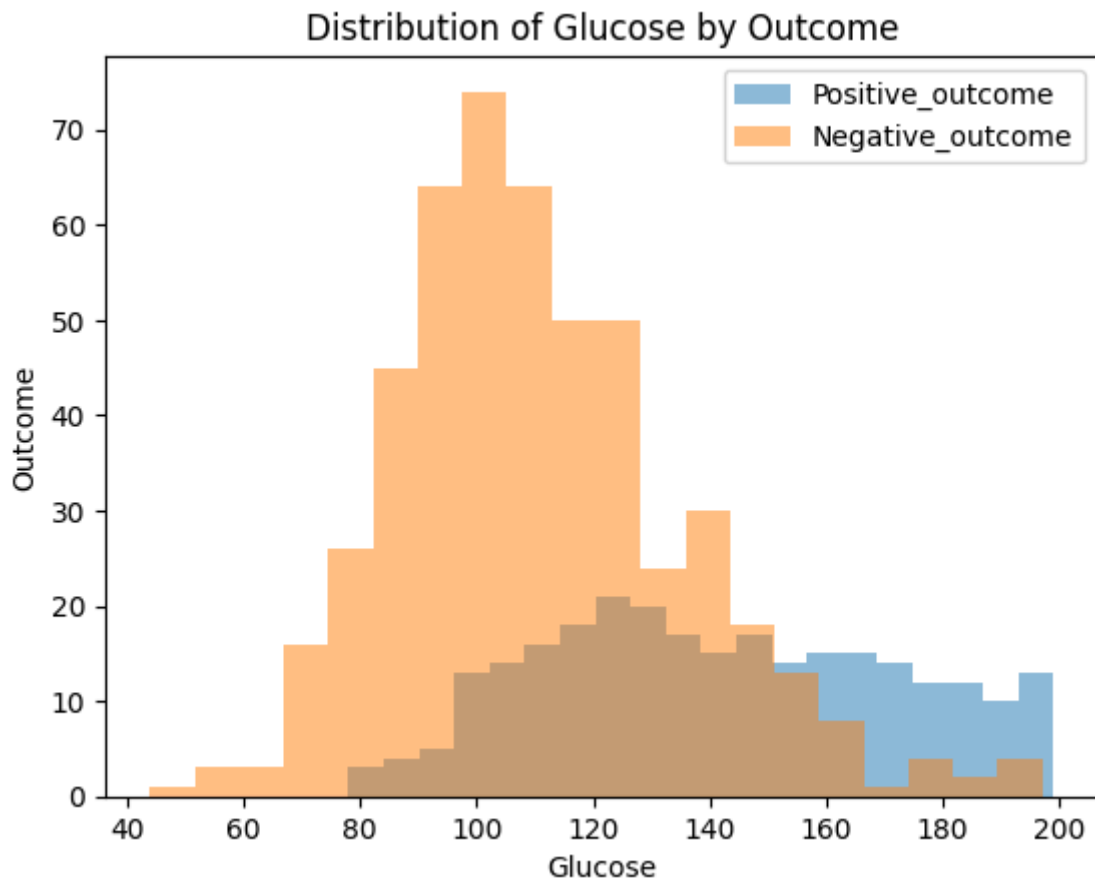


Рисунок 4 - Гистограмма зависимости концентрации глюкозы от целевого признака

Можно заметить, что чем выше концентрация глюкозы в плазме крови, тем вероятнее, что у пациента обнаружат диабет.

t-statistic: 13.420160648973402

p-value: 1.4738341999538704e-35

Построим гистограмму зависимости толщины кожной складки на трицепсе от вероятности наличия диабета и проверим t-статистику.

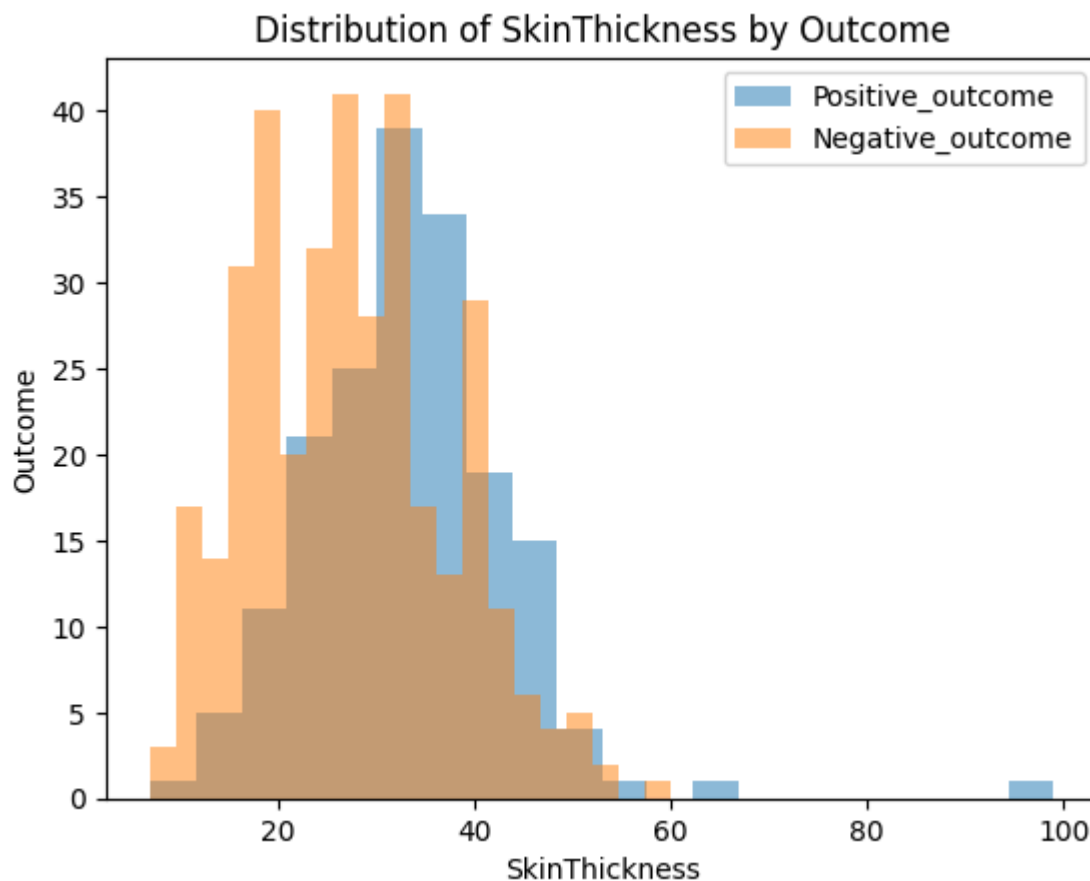


Рисунок 5 - Гистограмма зависимости толщины кожной складки от целевого признака

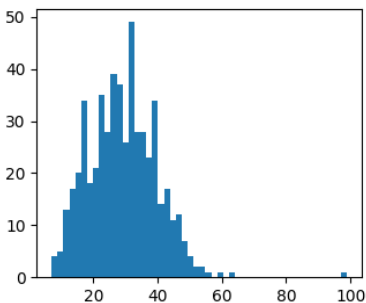
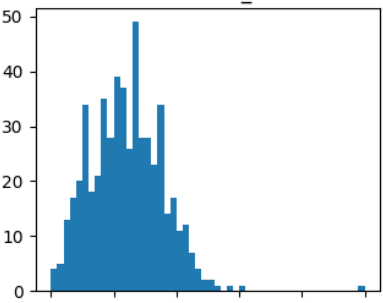
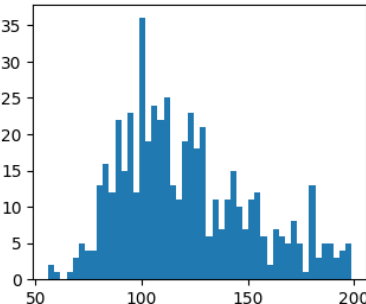
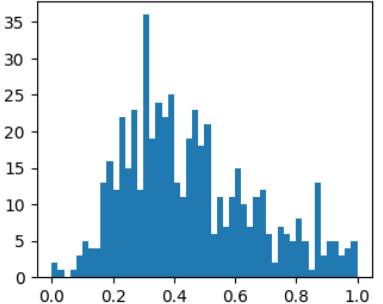
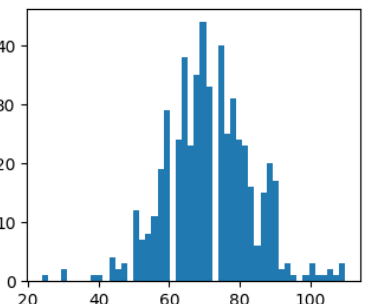
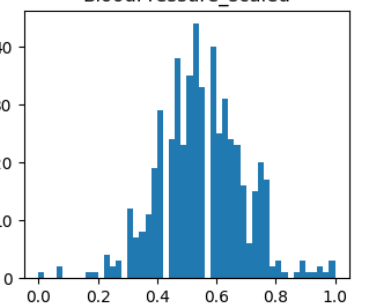
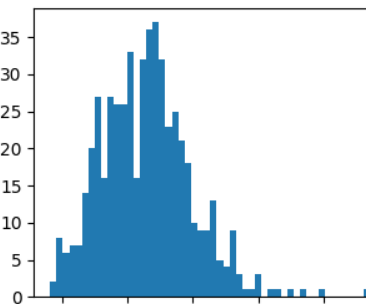
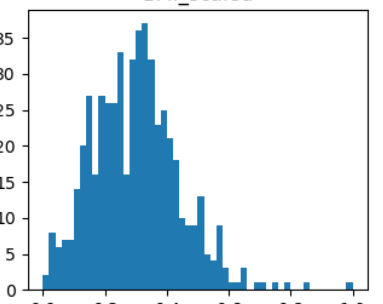
Можно заметить, что люди с положительным тестом на диабет (Positive_outcome) имеют тенденцию к большему диапазону значений толщины кожи, с большей долей в области 30-40 единиц.

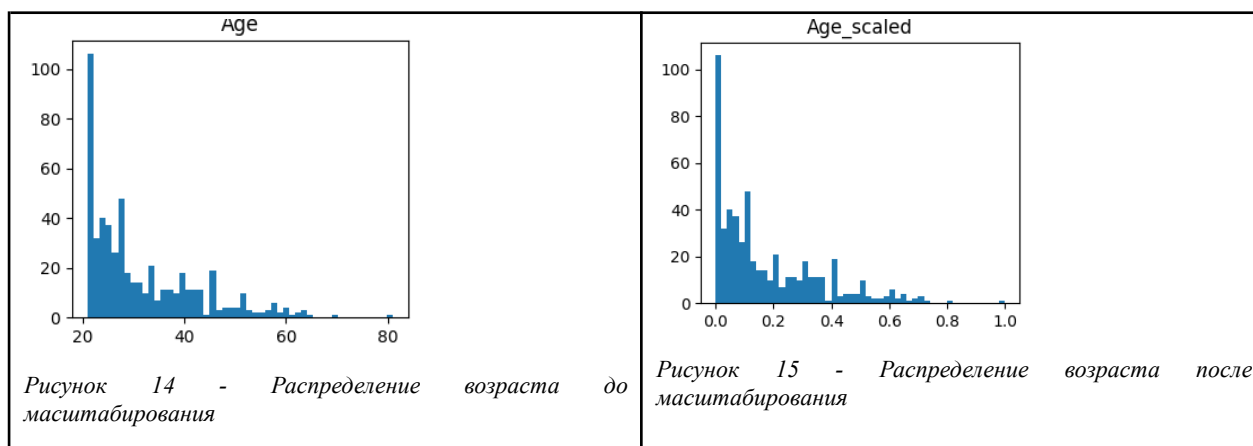
t-statistic: 6.0680342448810745

p-value: 2.466891060091481e-09

Далее приведем данные к нужному формату. Масштабируем численные признаки методом MinMax Scaler (этот метод оценивания масштабирует и преобразует каждый признак индивидуально таким образом, чтобы он находился в заданном диапазоне на обучающем наборе, например, между нулем и единицей). Посмотрим на распределение колонок до и после масштабирования.

Таблица 1 - Распределение числовых колонок до и после масштабирования

До масштабирования	После масштабирования
<p>SkinThickness</p>  <p>Рисунок 6 - Распределение толщины кожной складки до масштабирования</p>	<p>SkinThickness_scaled</p>  <p>Рисунок 7 - Распределение толщины кожной складки после масштабирования</p>
<p>Glucose</p>  <p>Рисунок 8 - Распределение концентрации глюкозы до масштабирования</p>	<p>Glucose_scaled</p>  <p>Рисунок 9 - Распределение концентрации глюкозы после масштабирования</p>
<p>BloodPressure</p>  <p>Рисунок 10 - Распределение артериального давления до масштабирования</p>	<p>BloodPressure_scaled</p>  <p>Рисунок 11 - Распределение артериального давления после масштабирования</p>
<p>BMI</p>  <p>Рисунок 12 - Распределение индекса массы тела до масштабирования</p>	<p>BMI_scaled</p>  <p>Рисунок 13 - Распределение индекса массы тела после масштабирования</p>



Масштабирование не повлияло на распределение данных.

Проводим корреляционный анализ данных, строя матрицу ошибок.

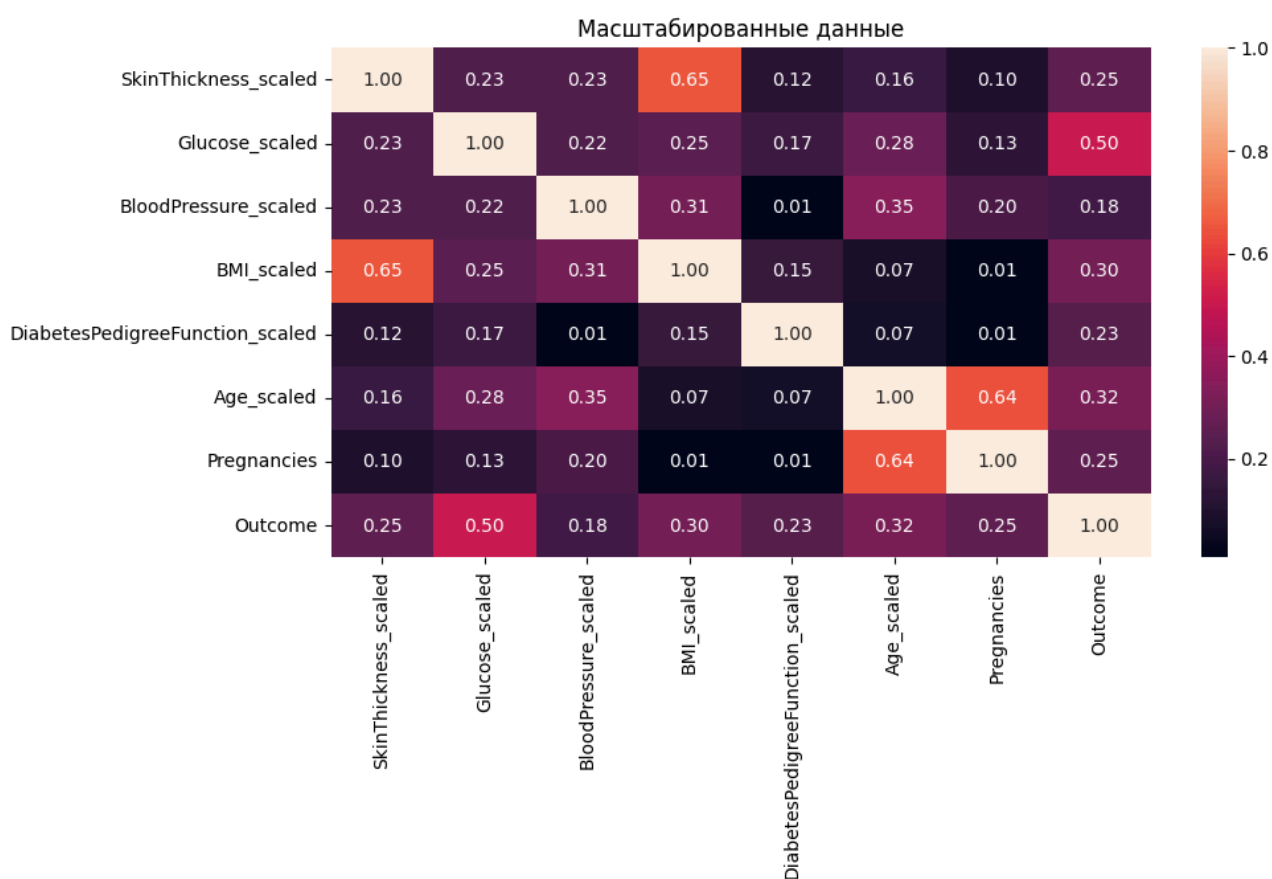


Рисунок 16 - Тепловая карта корреляций

Выводы по тепловой карте корреляций:

- корреляционные матрицы для исходных и масштабированных данных совпадают;
- целевой признак классификации "Outcome" наиболее сильно коррелирует с уровнем глюкозы Glucose (0.5) и умеренно коррелирует с признаками Age (0.32) и BMI (0.3). Эти признаки обязательно следует оставить в модели классификации;

- признаки Age и Pregnancies (0.64), а также BMI и SkinThickness (0.65) умеренно коррелируют между собой, но так как корреляция не является сильной или очень сильной не будем исключать какой-либо признак пары из набора для построения модели.;
- на основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения..

Выберем метрики для оценки качества модели:

- $Precision = \frac{TP}{TP+FP}$ - показывает, какую долю объектов, которые модель предсказала как положительные, действительно являются положительными.
- $Recall = \frac{TP}{TP+FN}$ - показывает, какую долю положительных объектов модель способна обнаружить.
- $F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ - среднее гармоническое precision и recall.

Другими словами, это средневзвешенное значение точности и отзыва. [2]

- $ROC AUC$ - основана на вычислении следующих характеристик:

$TPR = \frac{TP}{TP+FN}$ - True Positive Rate, откладывается по оси ординат.

Совпадает с recall. $FPR = \frac{FP}{FP+TN}$ - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно. Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика. Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации. [3]

Выберем модели для решения задачи классификации:

- KNN;
- SVC;
- Дерево решений;
- Случайный лес;

- Градиентный бустинг.

Формируем обучающую и тестовую выборку в соотношении 8:2.

Строим базовые решения (baseline), выводим значения метрик и ROC-кривую.

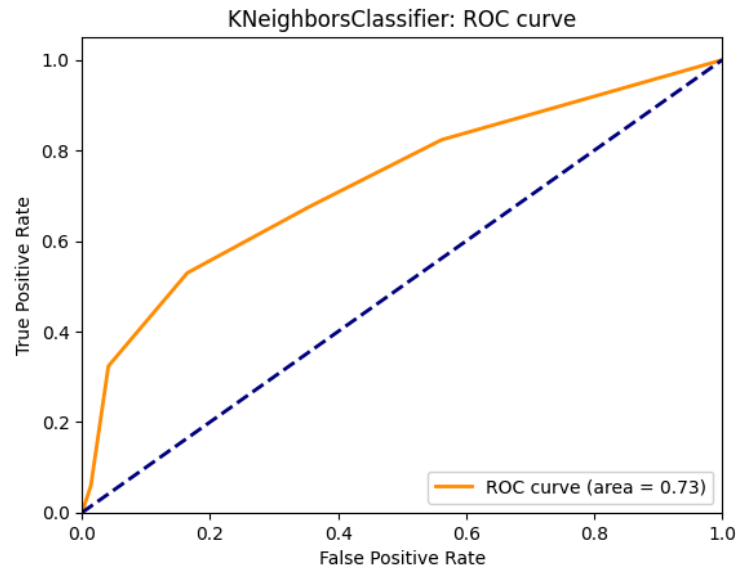


Рисунок 17 - ROC-кривая базовой модели KNN

KNeighborsClassifier:

Precision: 0.6

Recall: 0.53

F1-score: 0.56

ROC AUC score: 0.7276390008058018

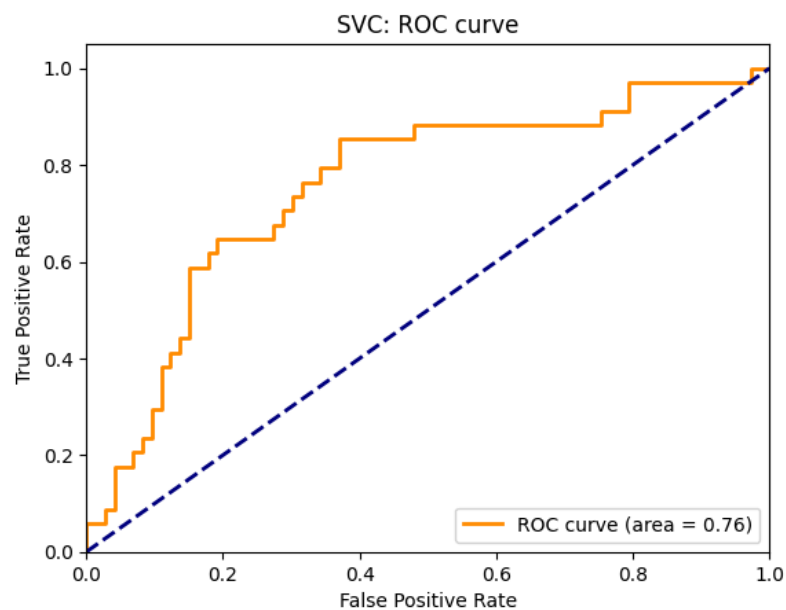


Рисунок 18- ROC-кривая базовой модели SVC

SVC:

Precision: 0.54

Recall: 0.21

F1-score: 0.3

ROC AUC score: 0.7570507655116842

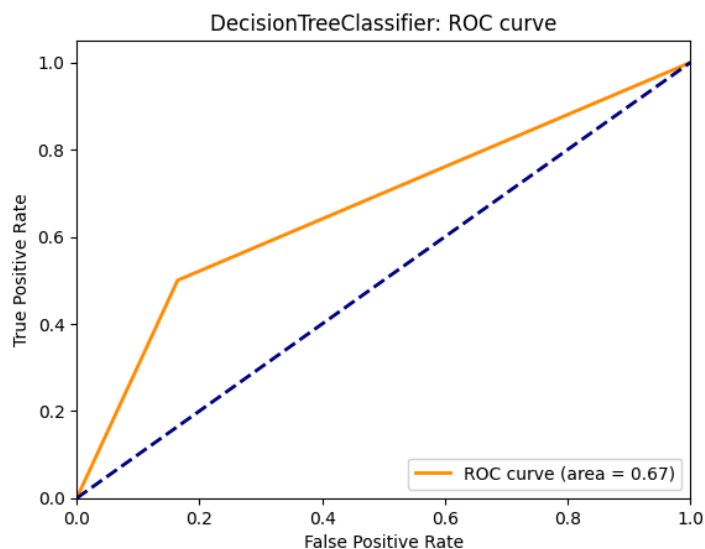


Рисунок 19 - ROC-кривая базовой модели Decision Tree

DecisionTreeClassifier:

Precision: 0.59

Recall: 0.5

F1-score: 0.54

ROC AUC score: 0.6678082191780821

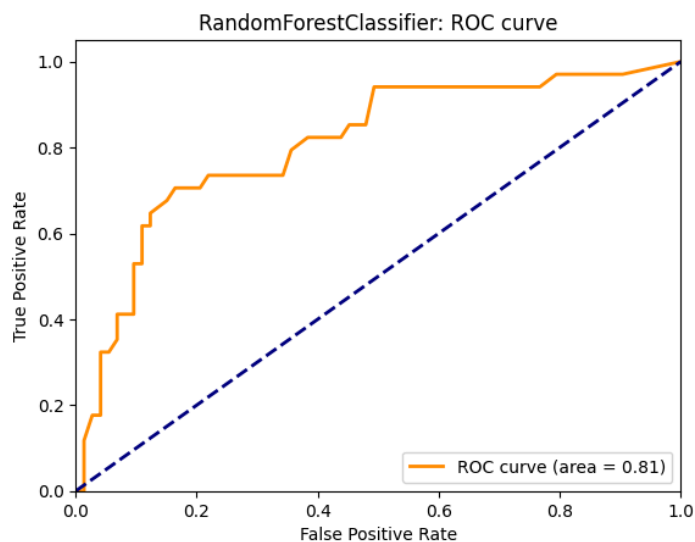


Рисунок 20 - ROC-кривая базовой модели Random Forest

RandomForestClassifier:

Precision: 0.72

Recall: 0.62

F1-score: 0.67

ROC AUC score: 0.8074133763094279

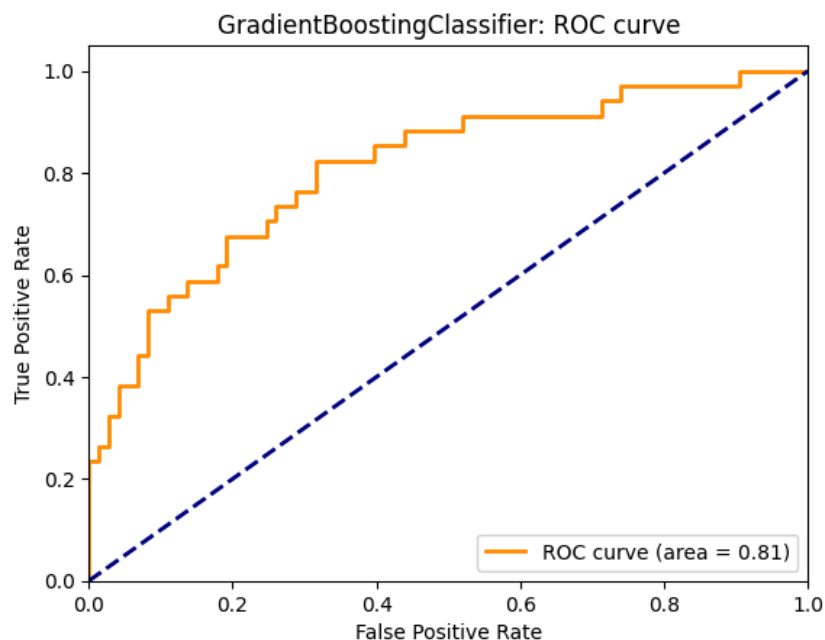


Рисунок 21 - ROC-кривая базовой модели Gradient Boosting

GradientBoostingClassifier:

Precision: 0.67

Recall: 0.59

F1-score: 0.62

ROC AUC score: 0.8094278807413376

Используем GridSearch (алгоритм поиска по сетке) для нахождения оптимальных гиперпараметров для каждой модели.

KNN:

Best hyperparameters: {'algorithm': 'auto', 'n_neighbors': 8, 'weights': 'uniform'}

Best score: 0.7027056277056276

SVC:

Best hyperparameters: {'C': 1, 'degree': 4, 'gamma': 'scale', 'kernel': 'linear'}

Best score: 0.7741317883661306

Дерево решений:

Best hyperparameters: {'criterion': 'gini', 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 10}

Best score: 0.7905882352941177

Случайный лес:

Best hyperparameters: {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}

Best score: 0.7905882352941177

Градиентный бустинг:

Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2}

Best score: 0.7976470588235294

Построение с найденными оптимальными значениями гиперпараметров

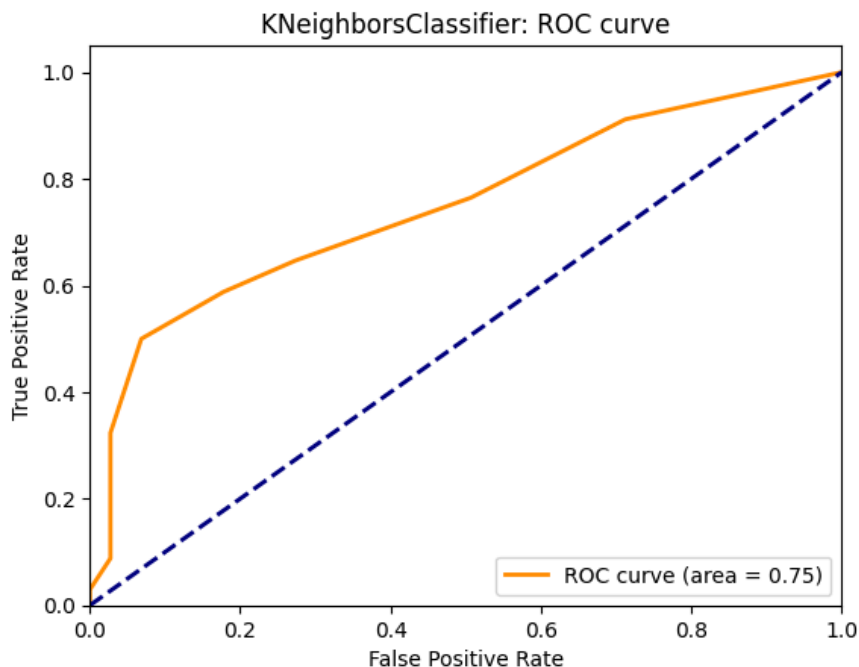


Рисунок 22 - ROC-кривая модели KNN после поиска гиперпараметров

KNeighborsClassifier:

Precision: 0.77

Recall: 0.5

F1-score: 0.61

ROC AUC score: 0.7489927477840451

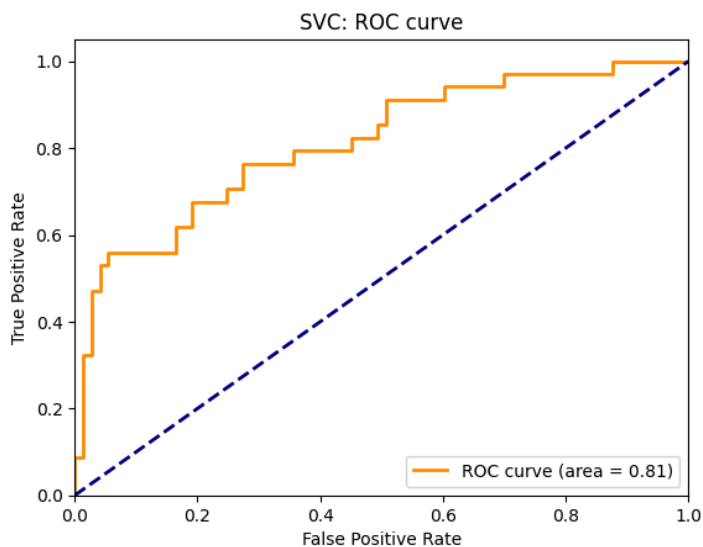


Рисунок 23 - ROC-кривая модели SVC после поиска гиперпараметров

SVC:

Precision: 0.73

Recall: 0.56

F1-score: 0.63

ROC AUC score: 0.8122481869460113

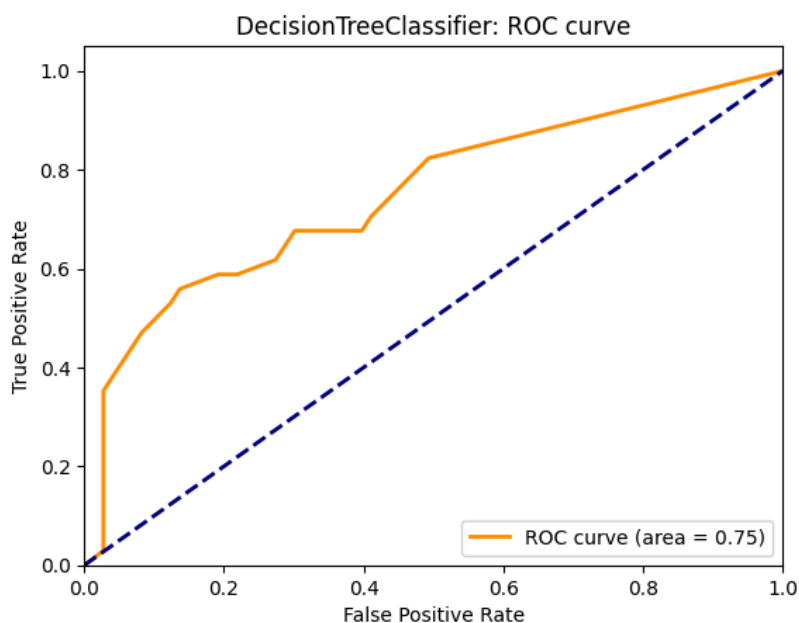


Рисунок 24 - ROC-кривая модели Decision Tree после поиска гиперпараметров

DecisionTreeClassifier:

Precision: 0.59

Recall: 0.59

F1-score: 0.59

ROC AUC score: 0.7485898468976632

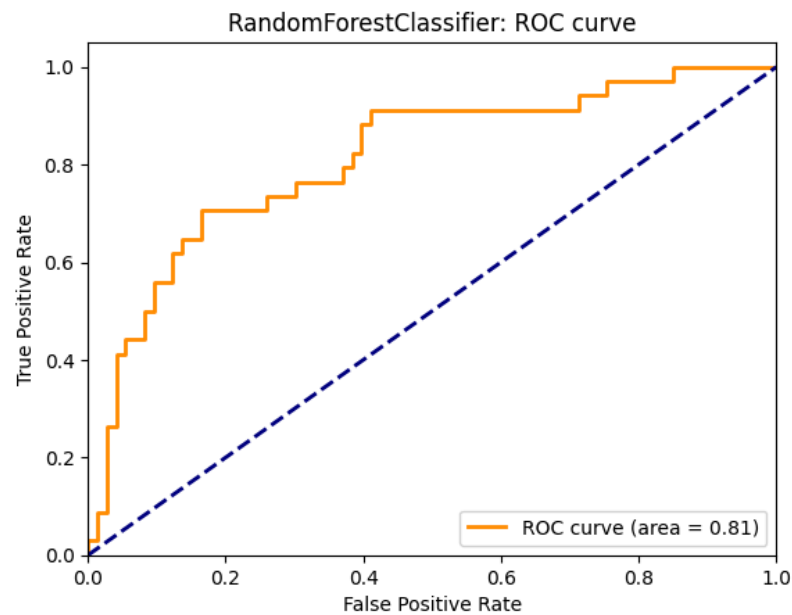


Рисунок 25 - ROC-кривая модели Random Forest после поиска гиперпараметров

RandomForestClassifier:

Precision: 0.68

Recall: 0.62

F1-score: 0.65

ROC AUC score: 0.8130539887187752

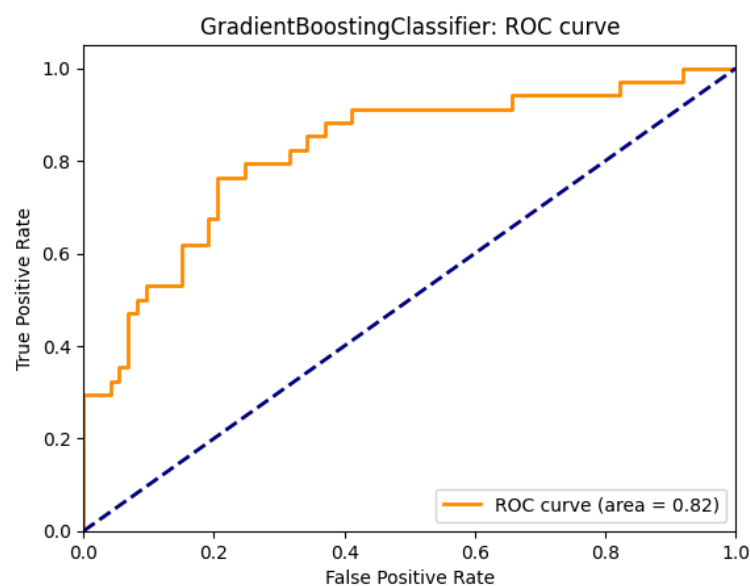


Рисунок 26 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров

GradientBoostingClassifier:

Precision: 0.64

Recall: 0.53

F1-score: 0.58

ROC AUC score: 0.8211120064464141

Модели-лидеры в метрках:

- Precision - KNeighborsClassifier
- Recall - RandomForestClassifier;
- F1 score - RandomForestClassifier;
- ROC AUC score - GradientBoostingClassifier.

На основании трех метрик из четырех используемых, лучшей оказалась модель **случайного леса**.

Заключение

Классификация параметра, отвечающего за показатель вероятности наличия/отсутствия у пациента сахарного диабета, с помощью методов машинного обучения является актуальной и перспективной задачей в области медицины. Анализ и обработка данных с помощью алгоритмов машинного обучения могут помочь своевременно предсказать заболевание до достижения им терминальной стадии и более внимательно отнестись к курсу лечения и диагностики пациента, предрасположенного к нему.

В рамках НИР была разработана эффективная модель, которая может помочь работникам медицинских центров быстро и точно определить вероятность обнаружения у пациента диабета и принять меры для минимизации врачебных ошибок и ускоренного процесса лечения.

Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самой точной на основании двух метрик из четырех оказалась модель случайного леса.

Список использованной литературы

1. T-test на Python для проверки и получения t-статистики // Помощник Python URL: <https://pythonpip.ru/osnovy/t-test-na-python>
2. Machine Learning Metrics in simple terms // Medium URL: <https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6>
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://nbviewer.org/github/ugapanyuk/courses_current/blob/main/notebooks/ml_project_example/project_classification_regression.ipynb
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses_current/wiki/COURSE_TMO_SPRING_2024/