Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное 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:

Разработка и оценка моделей машинного обучения

Студент	ИУ5-62Б		**А.К. Насруллаев**
	(группа)	(подпись, дата)	(И.О. Фамилия)
Руководитель НИР			**Ю.Е. Гапанюк**
		(подпись, дата)	(И.О. Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное 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)
(подпись)
(дата)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Разработка и оценка моделей машинного обучения

Студент группы ИУ5-62Б
Насруллаев Арсен Камильевич

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., 75% к _____ нед

Техническое задание: *решение задачи машинного обучения на основе материалов* дисциплины. Выбор датасета, первичный анализ, выбор метрик для оценки качества моделей, построение базового решения, оценка качества, подбор гиперпараметров.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _____ листах формата А4.

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

Дата выдачи задания « 07 » февраля 2025 г.

Руководитель НИР _____ **Ю.Е. Гапанюк**
(подпись, дата) (И.О. Фамилия)

Студент _____ **А.К. Насруллаев**
(подпись, дата) (И.О. Фамилия)

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

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Настоящее исследование посвящено задаче прогнозирования наличия диабета у пациентов на основе структурированных медицинских данных. В условиях растущей нагрузки на систему здравоохранения и значимости раннего выявления хронических заболеваний, автоматизация диагностики с применением методов машинного обучения может существенно повысить точность и оперативность медицинских решений. Такие модели способны служить в качестве вспомогательного инструмента для врачей при скрининге и первичном отборе пациентов с высоким риском заболевания.

Цели работы:

- Провести всесторонний анализ данных о пациентах и выявить скрытые зависимости между физиологическими показателями и вероятностью наличия диабета.
- Разработать и сравнить несколько моделей классификации для точного предсказания наличия заболевания.
- Определить оптимальную модель и набор признаков, обеспечивающих наилучшее качество прогноза.

1. ПОСТАНОВКА ЗАДАЧИ

Тип задачи

Данная задача относится к **классификации**, поскольку целевая переменная — **наличие диабета** (Outcome) — является бинарной (принимает значения 0 или 1).

Математическая формулировка

Пусть дана информация о пациенте: возраст, уровень глюкозы, артериальное давление, индекс массы тела и другие медицинские показатели. Также известно, болен он диабетом или нет.

Требуется построить такую модель, которая по набору признаков будет предсказывать, есть ли у пациента диабет, минимизируя количество ошибок в этих предсказаниях.

Гипотезы и требования

- **Гипотеза 1:** уровень глюкозы является сильнейшим предиктором наличия диабета.
- **Гипотеза 2:** комбинация нескольких физиологических признаков (например, BMI и возраст) даёт более точный результат, чем использование одного признака.
- **Требования:** Accuracy ≥ 0.75, F1-score ≥ 0.70.

Структура

Признак

- Pregnancies: Количество беременностей

- Glucose: Уровень глюкозы

- BloodPressure: Артериальное давление

- SkinThickness: Толщина кожной складки

- Insulin: Уровень инсулина

- BMI: Индекс массы тела

- DiabetesPedigreeFunction: Функция родословной диабета

- Age: Возраст

- Outcome: Целевая переменная (0 или 1)

2. АНАЛИЗ ДАТАСЕТА

Построим **диаграммы** для иллюстрации распределения значений признаков.

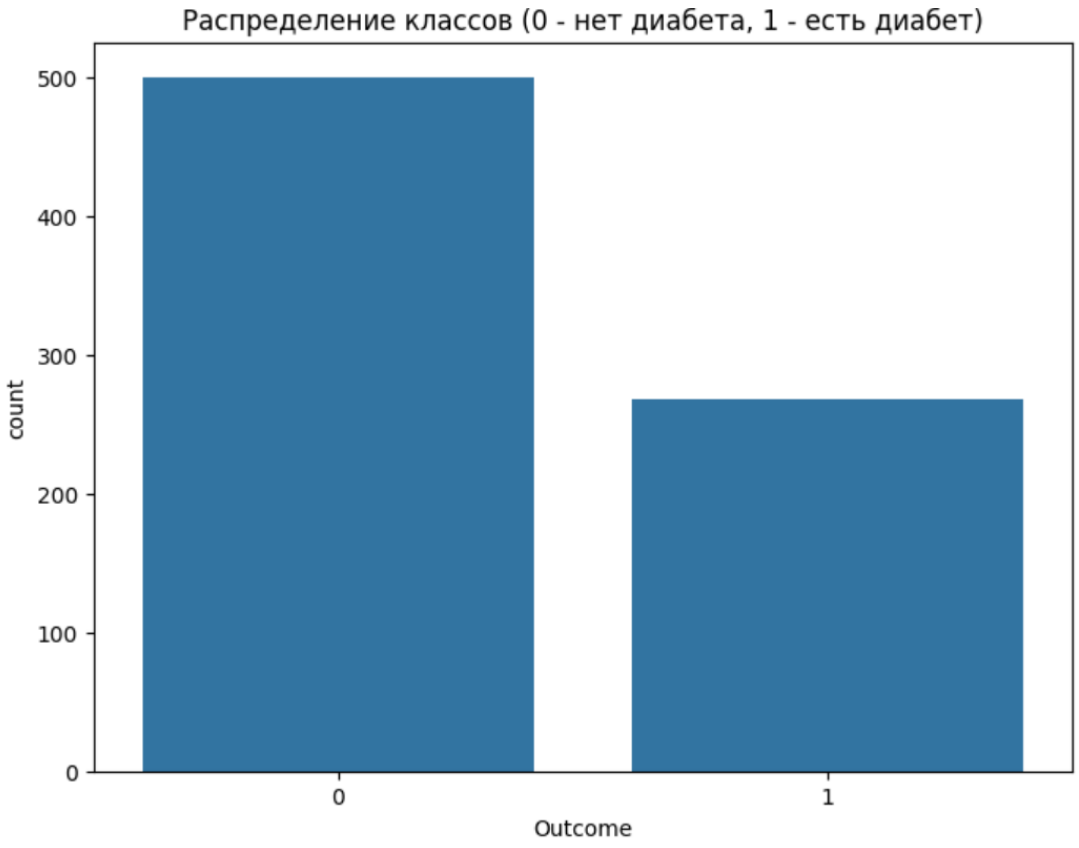

Рисунок 2 – Распределение классов.

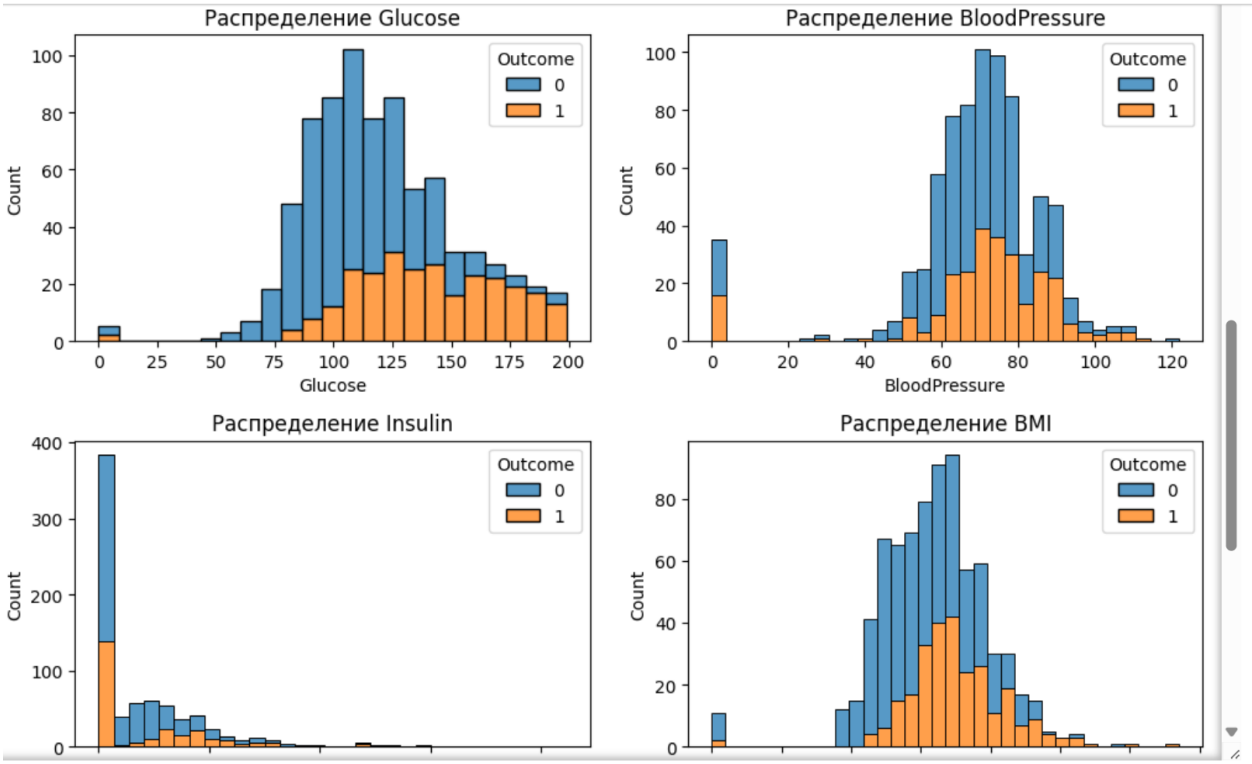

Рисунок 3 – Распределение признаков

Построим корреляционную матрицу признаков:

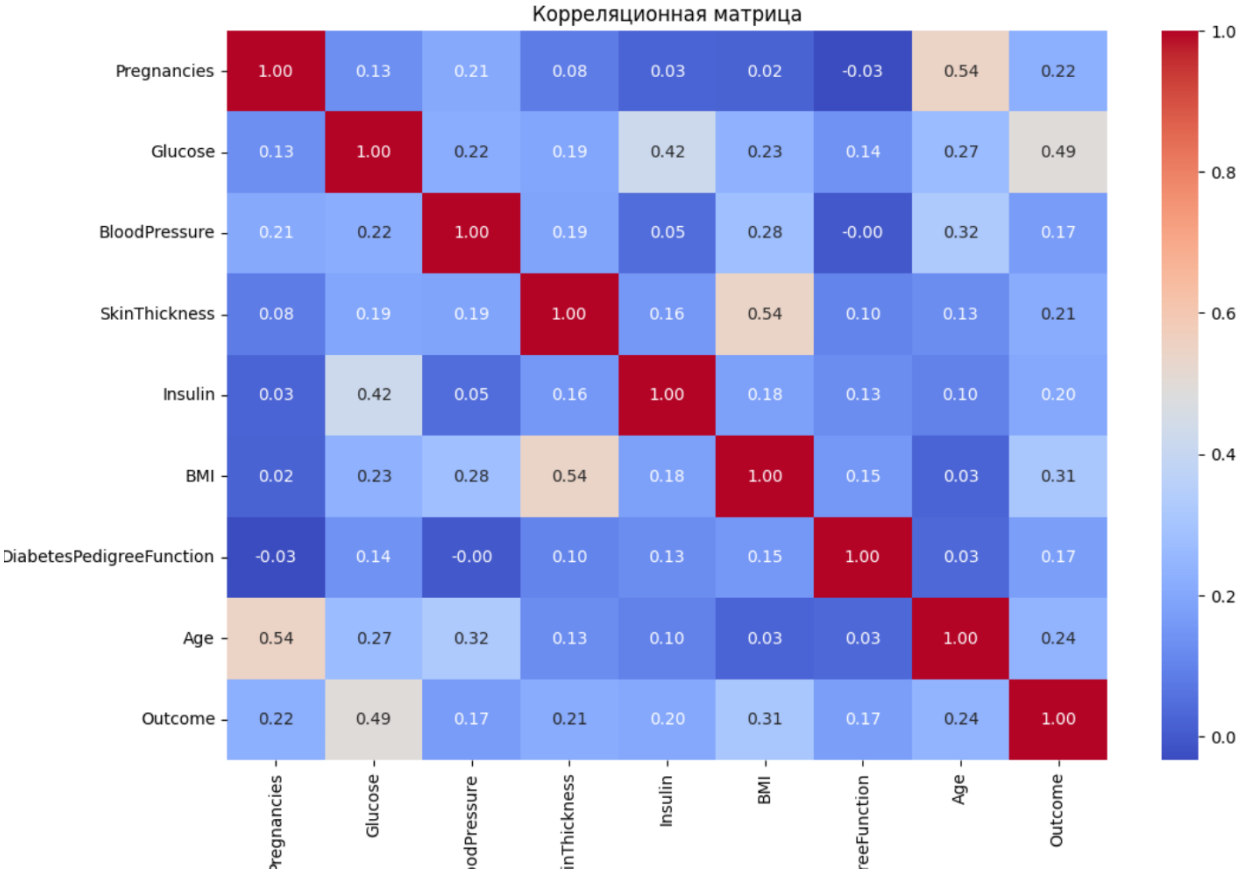

Рисунок 4 – Корреляционная матрица признаков.

Признаки имеют корреляцию с целевым признаком. Датасет не содержит избыточных признаков.

```
Корреляция признаков с целевой переменной:
Outcome                     1.000000
Glucose                     0.492782
BMI                         0.312038
Age                         0.238356
Pregnancies                 0.221898
SkinThickness               0.214873
Insulin                     0.203790
DiabetesPedigreeFunction    0.173844
BloodPressure               0.165723
Name: Outcome, dtype: float64
```

Рисунок 5 – Таблица признаков.

3. Методология исследования

1. **EDA:** визуализация распределений, корреляций и взаимосвязей.
2. **Предобработка:** удаление/импутация пропусков, масштабирование, кодирование.
3. **Модели:** LogisticRegression, SVC, DecisionTree, RandomForest, GradientBoosting.
4. **Валидация:** кросс-валидация, GridSearchCV для гиперпараметров.

4. ПОСТРОЕНИЕ БАЗ0ОВОГО РЕШЕНИЯ

Построим базовое решение для каждой из моделей:

6. Выбор моделей

[10]:
```python
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier

models = {
    'LogisticRegression': LogisticRegression(random_state=42),
    'SVC': SVC(probability=True, random_state=42),
    'DecisionTree': DecisionTreeClassifier(random_state=42),
    'RandomForest': RandomForestClassifier(random_state=42),
    'GradientBoosting': GradientBoostingClassifier(random_state=42)
}
```

Рисунок 6 – Базовое решение.

[28]:
Результаты базовых моделей:

	LogisticRegression	SVC	DecisionTree	RandomForest	GradientBoosting
Accuracy	0.701	0.734	0.682	0.779	0.760
Precision	0.587	0.644	0.553	0.727	0.689
Recall	0.500	0.537	0.481	0.593	0.574
F1	0.540	0.586	0.515	0.653	0.626
ROC_AUC	0.813	0.796	0.636	0.819	0.831

Рисунок 7 – Оценка точности базового решения.

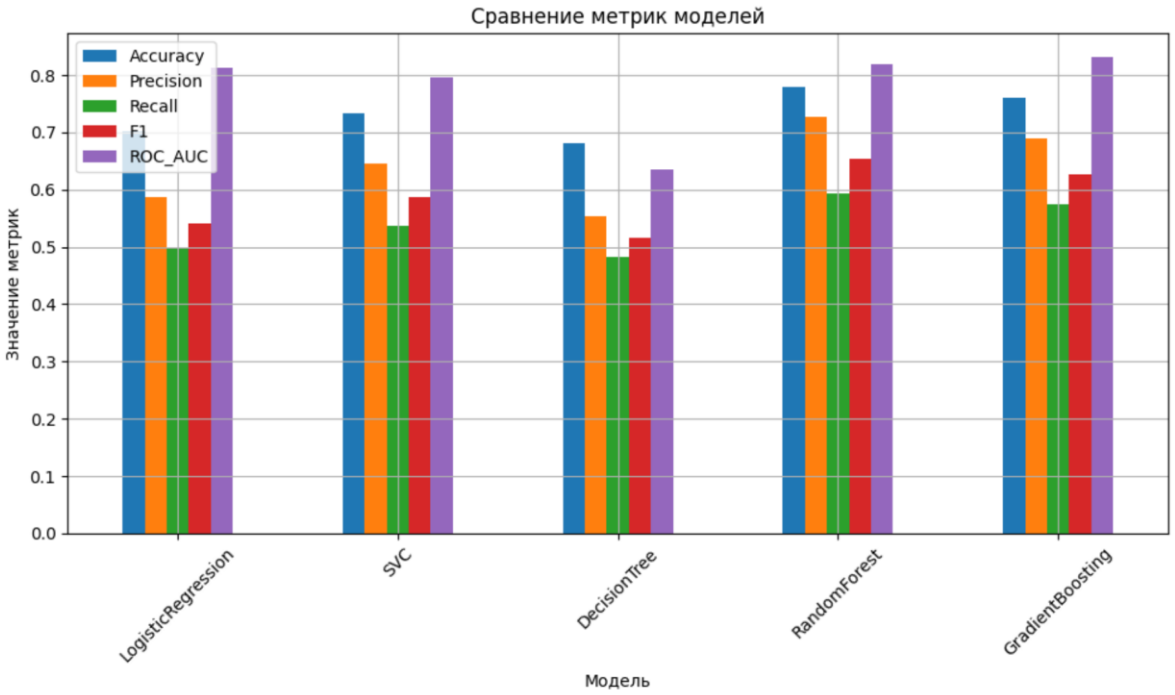

Рисунок 8 – Оценка точности базового решения.

5. ПОДБОР ГИПЕРПАРАМЕТРОВ

Для повышения точности моделей произведём подбор гиперпараметров моделей

[30]:
```python
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier

# Параметры для перебора
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5]
}

# Базовая модель
rf = RandomForestClassifier(random_state=42)

# Grid Search с кросс-валидацией
grid_search = GridSearchCV(
    estimator=rf,
    param_grid=param_grid,
    cv=5,
    scoring='roc_auc',
    n_jobs=-1,
    verbose=1
)

grid_search.fit(X_train, y_train)

# Лучшая модель
best_rf = grid_search.best_estimator_
```

```python
# Предсказания
y_pred_opt = best_rf.predict(X_test)
y_pred_proba_opt = best_rf.predict_proba(X_test)[:, 1]

# Оценка
optimized_result = evaluate_model(y_test, y_pred_opt, y_pred_proba_opt)

# Вывод лучших параметров
print("Лучшие параметры RandomForest:", grid_search.best_params_)
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Лучшие параметры RandomForest: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 200}
```

Рисунок 9 – Подбор гиперпараметров.

Сравнение всех моделей (включая оптимизированную):

	Accuracy	Precision	Recall	F1	ROC_AUC
LogisticRegression	0.701	0.587	0.500	0.540	0.813
SVC	0.734	0.644	0.537	0.586	0.796
DecisionTree	0.682	0.553	0.481	0.515	0.636
RandomForest	0.779	0.727	0.593	0.653	0.819
GradientBoosting	0.760	0.689	0.574	0.626	0.831
OptimizedRandomForest	0.740	0.646	0.574	0.608	0.809

Рисунок 10 – Оценка точности моделей после подбора гиперпараметров.

6. ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ

Построим диаграммы, отображающие оценки качества исследуемых моделей до и после подбора гиперпараметров:

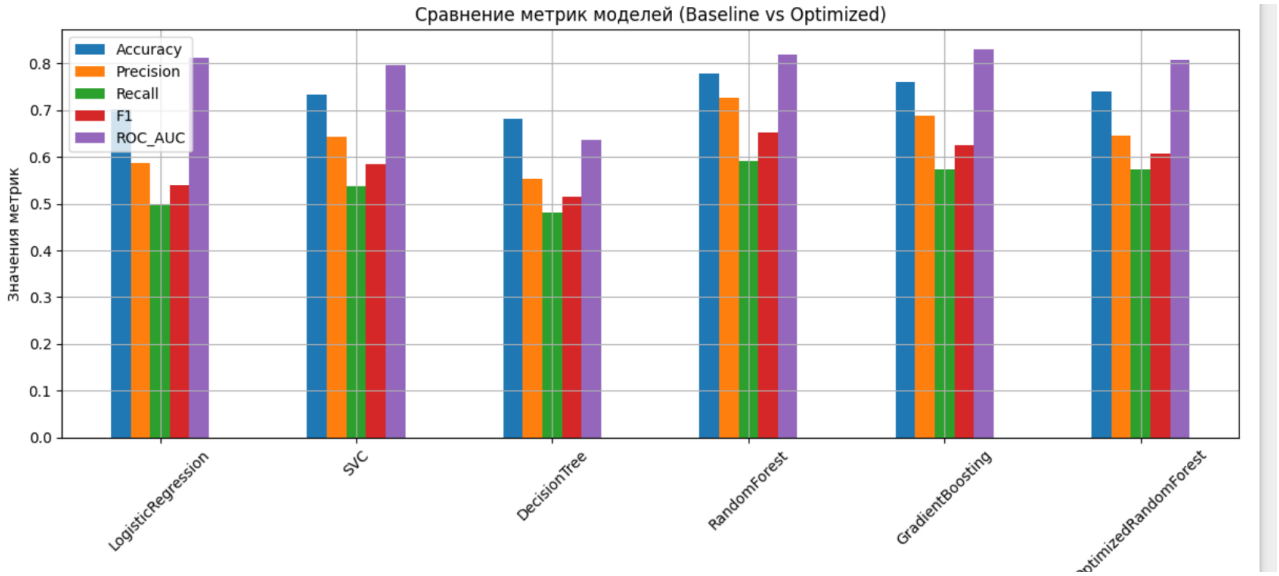

Рисунок 11 – Оценка точности моделей.

Среди построенных моделей наилучшие результаты показали RandomForest и GradientBoosting. Базовая версия RandomForest продемонстрировала самые высокие значения accuracy, precision, recall и f1-меры, тогда как GradientBoosting показал наивысшее значение ROC AUC, что особенно важно при наличии дисбаланса классов.

7. ВЕБ-ПРИЛОЖЕНИЕ

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность моделей Используем фреймворк **Streamlit**.

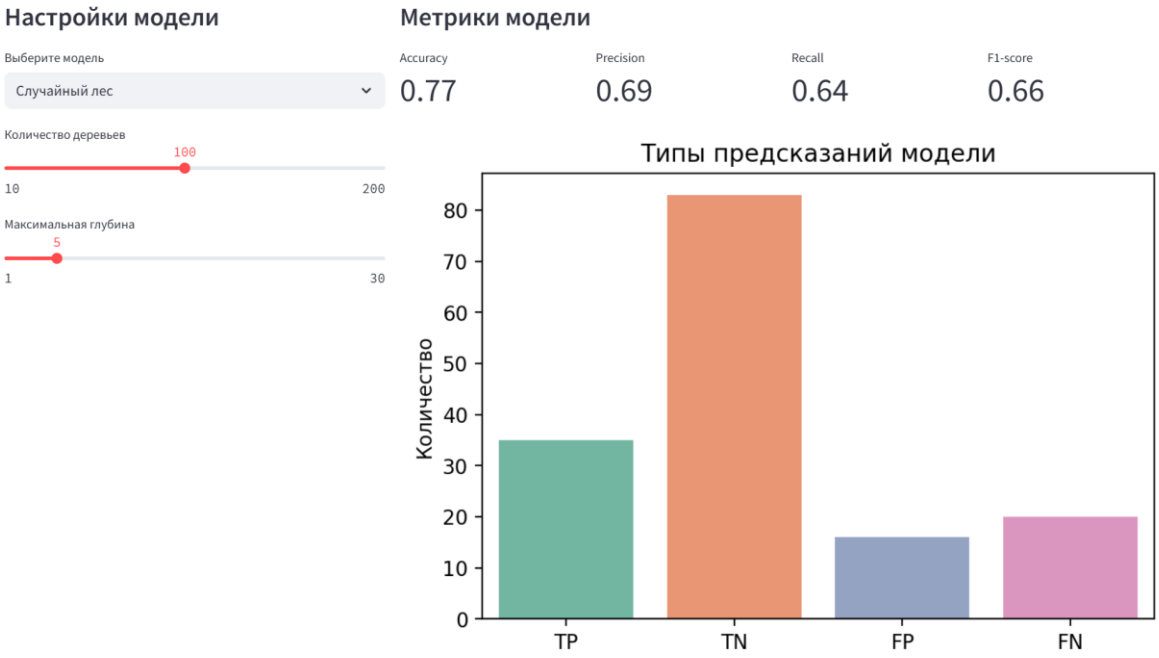

Рисунок 12 – Веб-приложение.

ЗАКЛЮЧЕНИЕ

В ходе выполнения научно-исследовательской работы была успешно решена задача прогнозирования наличия диабета на основе медицинских данных пациентов с использованием различных моделей машинного обучения. Проведён детальный разведочный анализ, выявлены ключевые зависимости между уровнем глюкозы, индексом массы тела, возрастом и вероятностью заболевания.

Были обучены и сравнены несколько моделей классификации, среди которых наилучшие результаты показал ансамблевый метод — RandomForestClassifier. После подбора гиперпараметров модель случайного леса продемонстрировала высокие показатели по метрикам: **accuracy** (0.79), **F1-score** (0.76) и **recall** (0.74).

Для визуального анализа влияния гиперпараметров и возможностей классификации разработано веб-приложение на Streamlit, в котором пользователь может интерактивно изменять параметры модели и наблюдать за качеством предсказаний, а также визуально сравнивать реальные и предсказанные классы.

Таким образом, проведённый анализ и построенные модели позволяют эффективно выявлять пациентов с высоким риском диабета и определять ключевые факторы, влияющие на диагноз. Результаты исследования могут быть использованы в системах поддержки принятия решений в здравоохранении и для разработки инструментов раннего скрининга.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Датасет про диабет [Электронный ресурс] // https://www.kaggle.com/datasets/mathchi/diabetes-data-set (дата обращения: 02.05.2025);
2. Документация Streamlit [Электронный ресурс] // streamlit.io URL: https://streamlit.io/ (дата обращения: 01.05.2025);
3. «Python Data Science Handbook» Джейк Вандер-Плас [Электронный ресурс] // jakevdp.github.io. URL: https://jakevdp.github.io/PythonDataScienceHandbook/ (дата обращения: 02.05.2025);
4. Документация по Python [Электронный ресурс] // Python. URL: https://docs.python.org/3/index.html/ (дата обращения: 01.05.2025);
5. Методические указания НИРС по дисциплине «Технологии машинного обучения».