

Применение методов разведочного анализа и логистической регрессии для прогнозирования инсульта

Саяпин Никита Анатольевич

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

Цель работы — продемонстрировать минимальный воспроизводимый цикл анализа данных (EDA) на примере Stroke Prediction Dataset: загрузка данных, вычисление описательных статистик, визуализация распределений признаков и их связи с целевой переменной («наличие инсульта»), а также построение простой модели классификации (логистическая регрессия) с базовой оценкой качества. Датасет загружен из открытого источника Kaggle, что упрощает повторение эксперимента. Результаты показывают, что даже простые методы позволяют выявить факторы риска инсульта и получить интерпретируемую модель.

Ключевые слова: анализ данных, инсульт, EDA, визуализация, классификация, Kaggle, Python, pandas, scikit-learn, matplotlib.

Application of exploratory analysis and logistic regression methods for stroke prediction

Sayapin Nikita Anatolyevich

Sholom-Aleichem Priamursky State University

Student

Abstract

The aim of the work is to demonstrate the minimum reproducible data analysis cycle (EDA) using the example of the Stroke Prediction Dataset: loading data, calculating descriptive statistics, visualizing feature distributions and their relationship to the target variable ("stroke presence"), as well as building a simple classification model (logistic regression) with a basic quality assessment. The dataset is downloaded from the Kaggle open source, which makes it easier to repeat the experiment. The results show that even simple methods make it possible to identify stroke risk factors and obtain an interpretable model.

Keywords: data analysis, stroke, EDA, visualization, classification, Kaggle, Python, pandas, scikit-learn, matplotlib.

1. Введение

1.1 Актуальность

В России наблюдается общий тренд к снижению смертности от болезней системы кровообращения, однако абсолютные масштабы проблемы остаются

высокими, а региональная неоднородность — значительной. Следовательно, инструменты аналитики данных и машинного обучения (ML) востребованы для: выявления предикторов инсульта, балансировки рисков ложноположительных/ложноотрицательных решений в скрининге, интерпретируемых рекомендаций для профилактики. Выбор в этой работе Stroke Prediction Dataset (Kaggle) обусловлен его открытостью, структурой признаков (возраст, гипертония, заболевания сердца, глюкоза, ИМТ, курение и др.) и воспроизводимостью эксперимента для академических и практических целей.

1.2 Обзор исследований

В работе В.В. Донитовой подчёркивается значимость выявления факторов риска острых нарушений мозгового кровообращения; авторы анализируют клинические показатели и демонстрируют применение методов машинного обучения для выделения предикторов инсульта, что напрямую соотносится с нашей целью — исследовать вклад различных признаков в прогноз [1]. И.А. Мишкин выполняет сравнительное исследование алгоритмов оценки сердечно-сосудистого риска с использованием современных ML-библиотек; показано, что корректный выбор модели и метрик (ROC-AUC, F1-score) принципиально влияет на выводы в задачах с несбалансированными данными — подход, совпадающий с нашей логикой анализа Stroke Prediction Dataset [2]. В методической работе «Анализ и обработка данных в медицине» приведён обзор приёмов предобработки клинических выборок: устранение пропусков, масштабирование числовых признаков и кодирование категориальных переменных; акцент сделан на важности качественного EDA перед построением моделей, что перекликается с разделами «Материалы и методы» и «Ход работы» нашего исследования [3]. Н.А. Благодосклон рассматривает автоматизированные методы извлечения факторов риска из текстов историй болезни; автор демонстрирует возможности интеграции текстовых и структурированных данных для прогнозирования инсульта. Эта логика подтверждает важность расширения признакового пространства и подчеркивает потенциал комплексного подхода к анализу данных [4]. В обзорной статье по применению машинного обучения в реабилитационной медицине подчёркивается значимость интерпретируемых моделей и корректного выбора метрик для медицинских задач; авторы отмечают, что простые алгоритмы (например, логистическая регрессия) часто обеспечивают достаточную точность и прозрачность, что соответствует нашему выбору базовой модели классификации [5].

1.3 Цель исследования

Целью данного исследования является демонстрация базового цикла анализа данных на примере Stroke Prediction Dataset, включающего загрузку, первичную обработку, вычисление описательных статистик, построение визуализаций и реализацию простой модели классификации. Исследование направлено на показ значимости этапа эксплораторного анализа данных

(EDA) и подтверждение того, что даже минимальные методы позволяют выявлять ключевые факторы риска и получать интерпретируемые результаты при прогнозировании инсульта.

2 Материалы и методы

В качестве основной среды разработки используется язык программирования Python версии 3.x, который применяется для анализа данных благодаря широкому набору специализированных библиотек. Для реализации исследования применяются следующие инструменты:

1. библиотека `pandas` для работы с табличными структурами и вычисления описательных статистик;
2. библиотека `numpy` для выполнения числовых операций;
3. библиотека `matplotlib` для построения графиков и визуализации распределений;
4. библиотека `scikit-learn` для разделения выборки на обучающую и тестовую части, предобработки данных, построения и оценки модели классификации.

Данные загружаются из открытого источника Kaggle (Stroke Prediction Dataset) [6] в формате CSV, что обеспечивает воспроизводимость эксперимента и доступность набора для исследователей.

Работа ведётся в среде разработки PyCharm, которая предоставляет удобные средства для написания и запуска кода, а также поддержку виртуальных окружений.

2.2 Подготовка

На первом этапе создаётся папка проекта `stroke_eda`. Внутри неё формируется структура, включающая основной скрипт `main.py`, каталог `figures` для сохранения графиков и каталог `reports` для хранения отчётов и таблиц. Такая организация обеспечивает порядок и воспроизводимость эксперимента (Рис. 1).

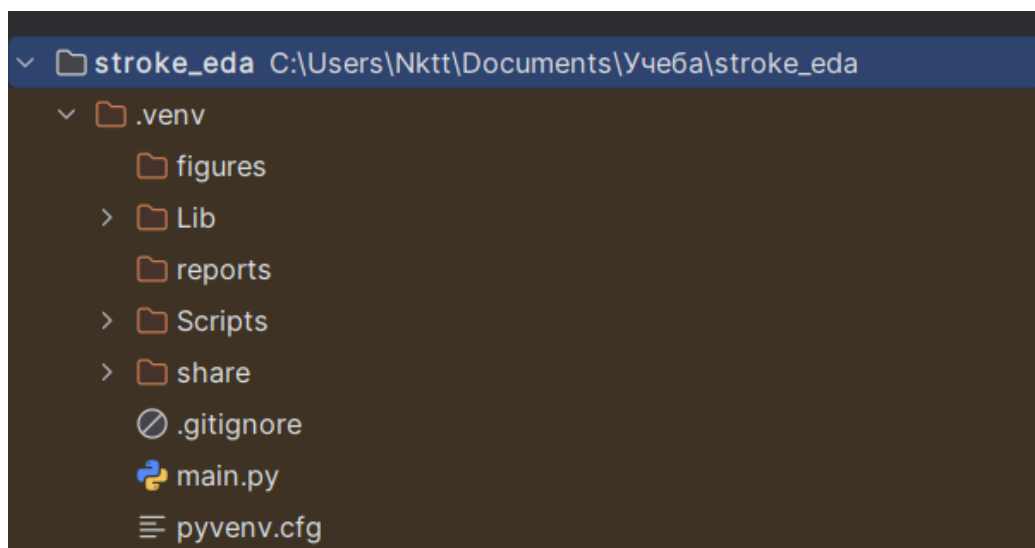


Рисунок 1 – Создание проекта `stroke_eda`

Для работы с библиотеками создаётся виртуальное окружение Python, в которое устанавливаются все необходимые пакеты. Использование виртуального окружения исключает конфликты версий библиотек и делает проект переносимым. Установка зависимостей выполняется через менеджера пакетов `pip`.

3. Результаты

3.1 Загрузка и первичная обработка данных

На первом этапе выполняется загрузка датасета Stroke Prediction из Kaggle. Файл `stroke.csv` помещается в папку `data`, после чего он считывается в объект `DataFrame`:

```
df = pd.read_csv("data/stroke.csv")
df.columns = [c.strip().lower().replace(' ', '_') for c in df.columns]
```

Выполняются базовые преобразования:

- имена столбцов приводятся к единому формату (`snake_case`);
- пропуски в `bmi` заполняются медианой;
- бинарные признаки (`hypertension`, `heart_disease`, `stroke`) приводятся к целым значениям;
- удаляются некорректные значения возраста (например, `>120` лет).

3.2 Визуализация данных

Для числовых признаков (`age`, `avg_glucose_level`, `bmi`) строятся гистограммы (Рис. 2-5).

```
numeric_cols = [c for c in ['age', 'avg_glucose_level', 'bmi'] if c in df.columns]
for c in numeric_cols:
    plt.figure()
    df[c].hist(bins=30)
    plt.xlabel(c)
    plt.ylabel('Количество')
    plt.title(f'Гистограмма признака: {c}')
    plt.tight_layout()
    plt.savefig(FIGURES / f'гистограмма_{c}.png', dpi=200)
    plt.close()
```

Рисунок 2 – Код построения гистограм по возрасту, уровню глюкозы, массе

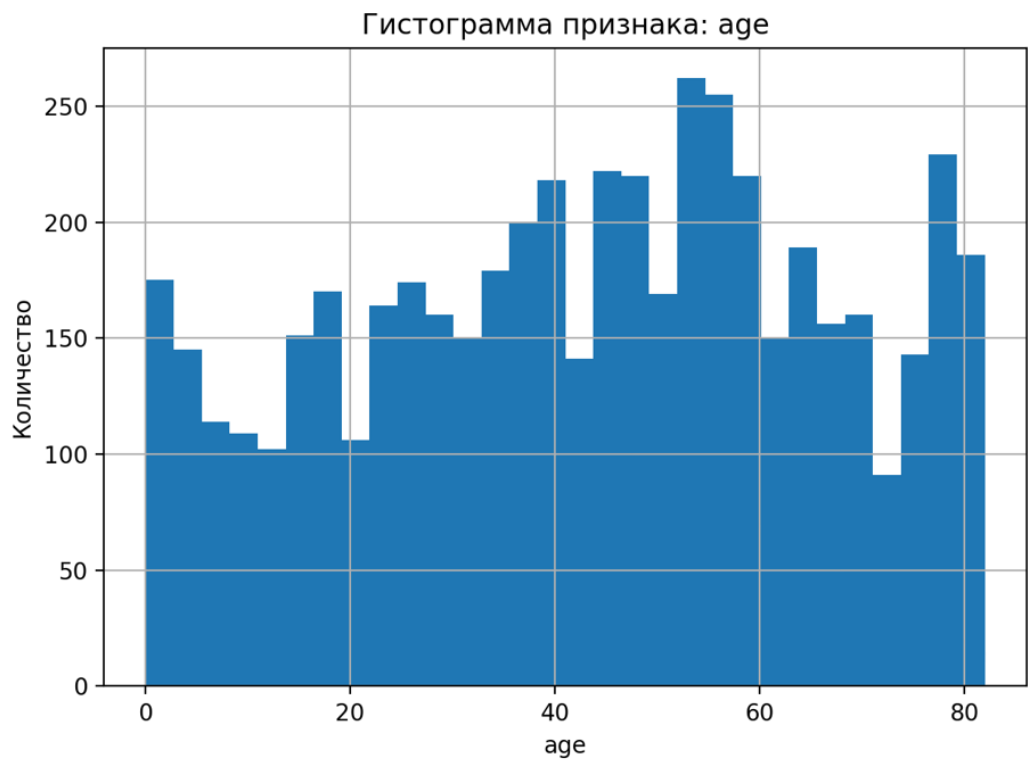


Рисунок 3 – Гистограмма распределения по возрасту

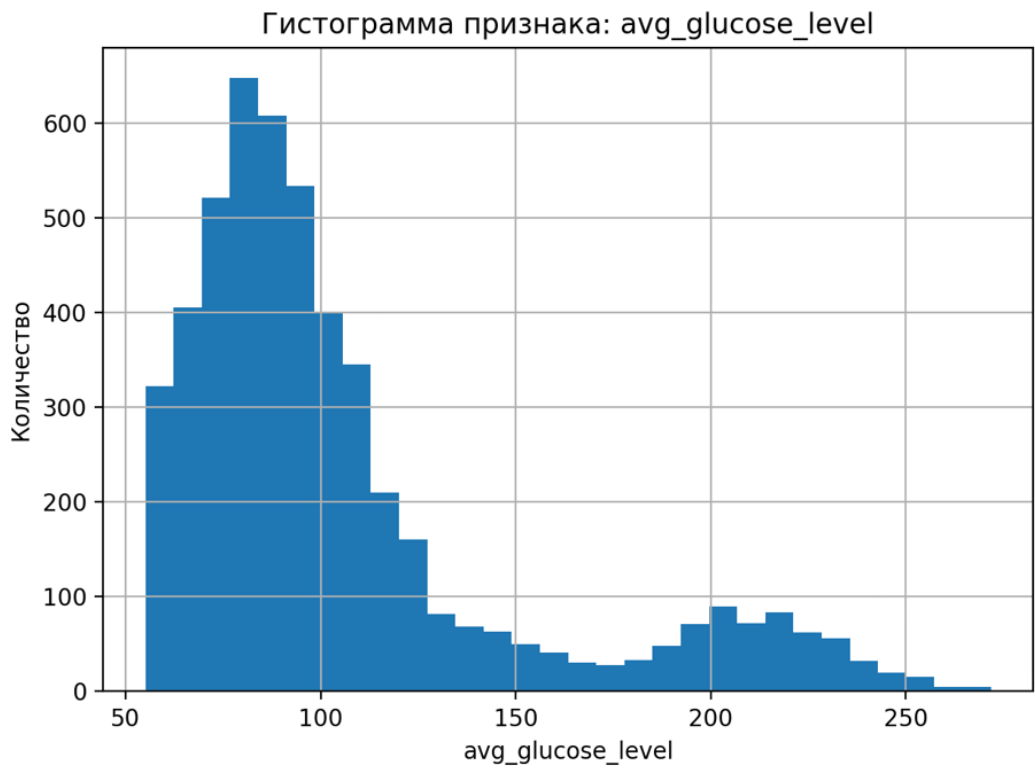


Рисунок 4 – Гистограмма распределения по уровню глюкозы

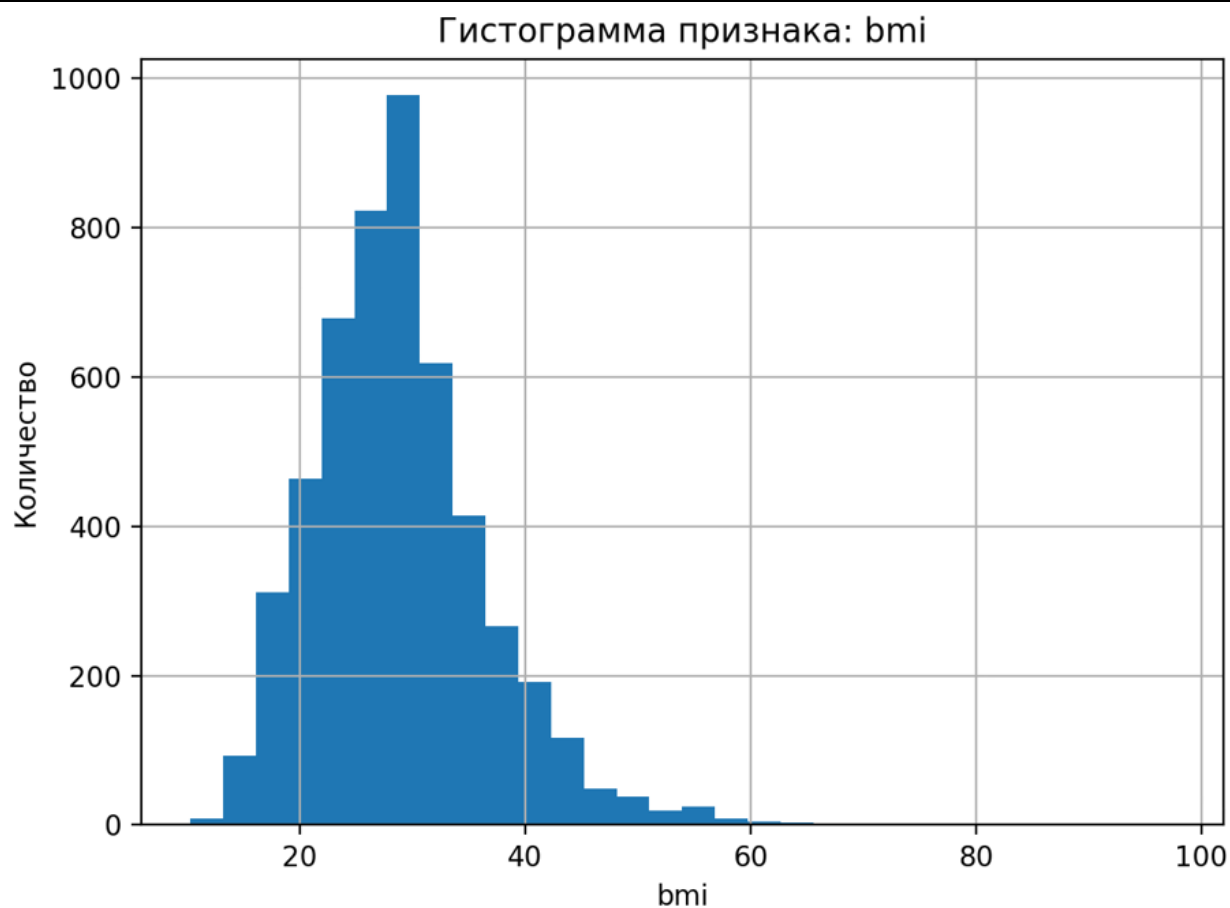


Рисунок 5 – Гистограмма распределения по индексу массы тела

Далее строятся диаграммы размаха для числовых признаков, позволяющие выявить выбросы (Рис. 6):

```
df[["age","avg_glucose_level","bmi"]].boxplot()  
plt.title("Диаграммы размаха числовых признаков")
```

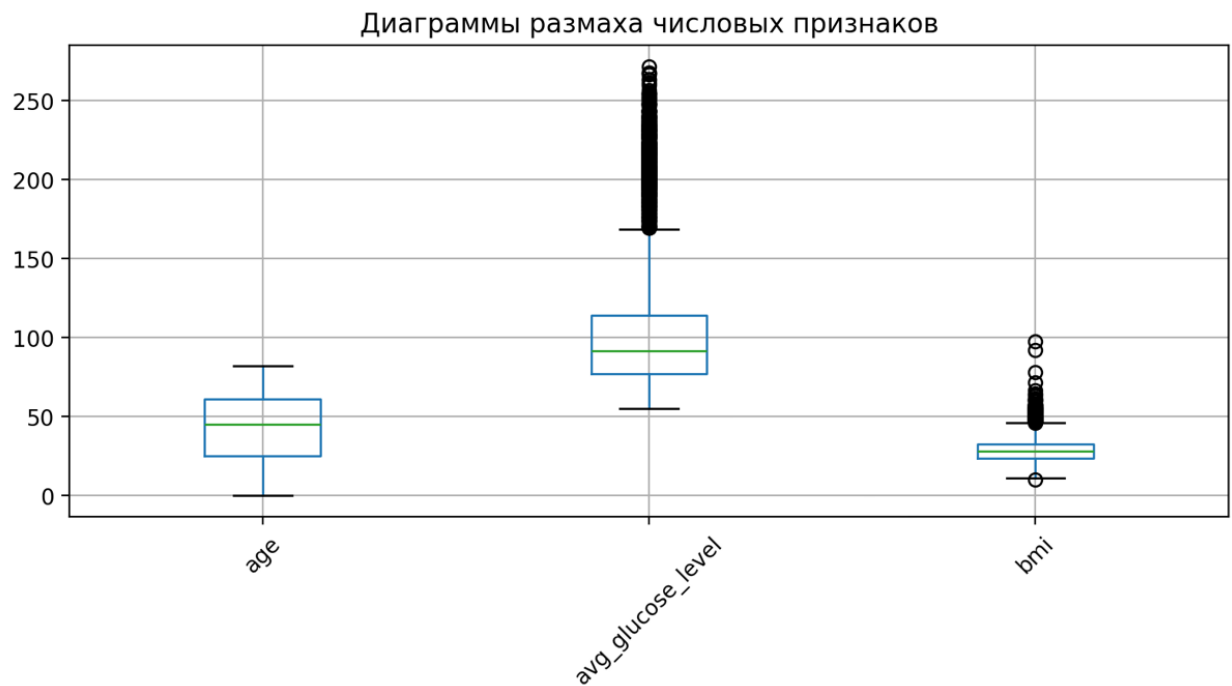


Рисунок 6 – Диаграмма размаха числовых признаков

Для категориальных переменных (gender, work_type, smoking_status и др.) визуализируется доля инсульта в каждой категории (Рис. 7–8).

```
cat_cols = [c for c in ['gender', 'ever_married', 'work_type', 'residence_type', 'smoking_status'] if c in df.columns]
for c in cat_cols:
    grp = df.groupby(c)['stroke'].mean().sort_values(ascending=False)
    plt.figure(figsize=(8, 4))
    grp.plot(kind='bar')
    plt.ylabel('Доля инсульта')
    plt.title(f'Доля инсульта по категории: {c}')
    plt.tight_layout()
    plt.savefig(FIGURES / f'доля_инсульта_{c}.png', dpi=200)
    plt.close()
```

Рисунок 7 – Код построения столбчатой диаграммы

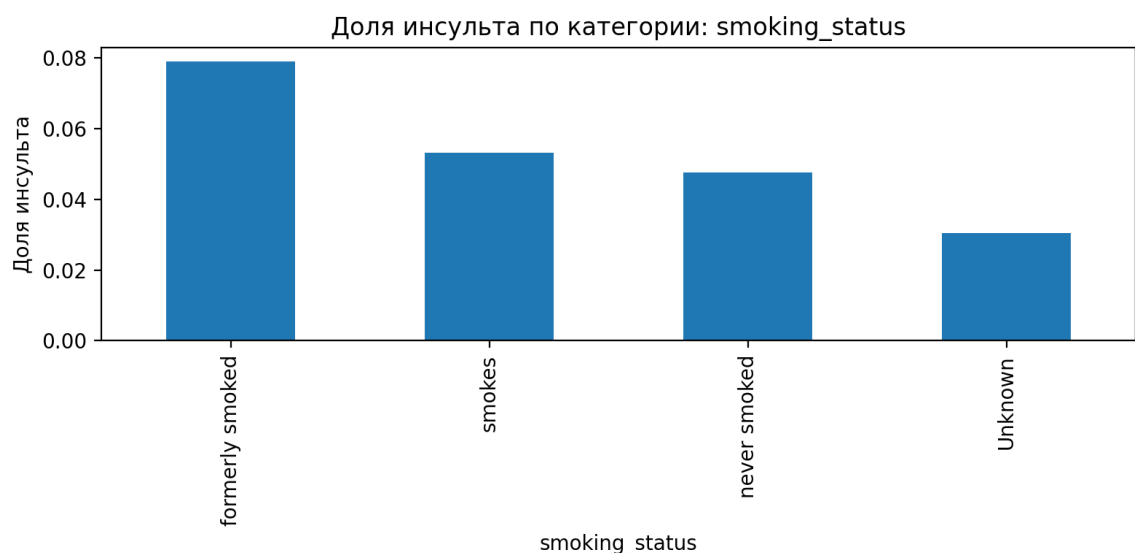


Рисунок 8 – Доля инсульта в зависимости от статуса курения

Наглядной является диаграмма рассеяния «возраст – уровень глюкозы», где пациенты с инсультом выделяются отдельным маркером (Рис. 9).

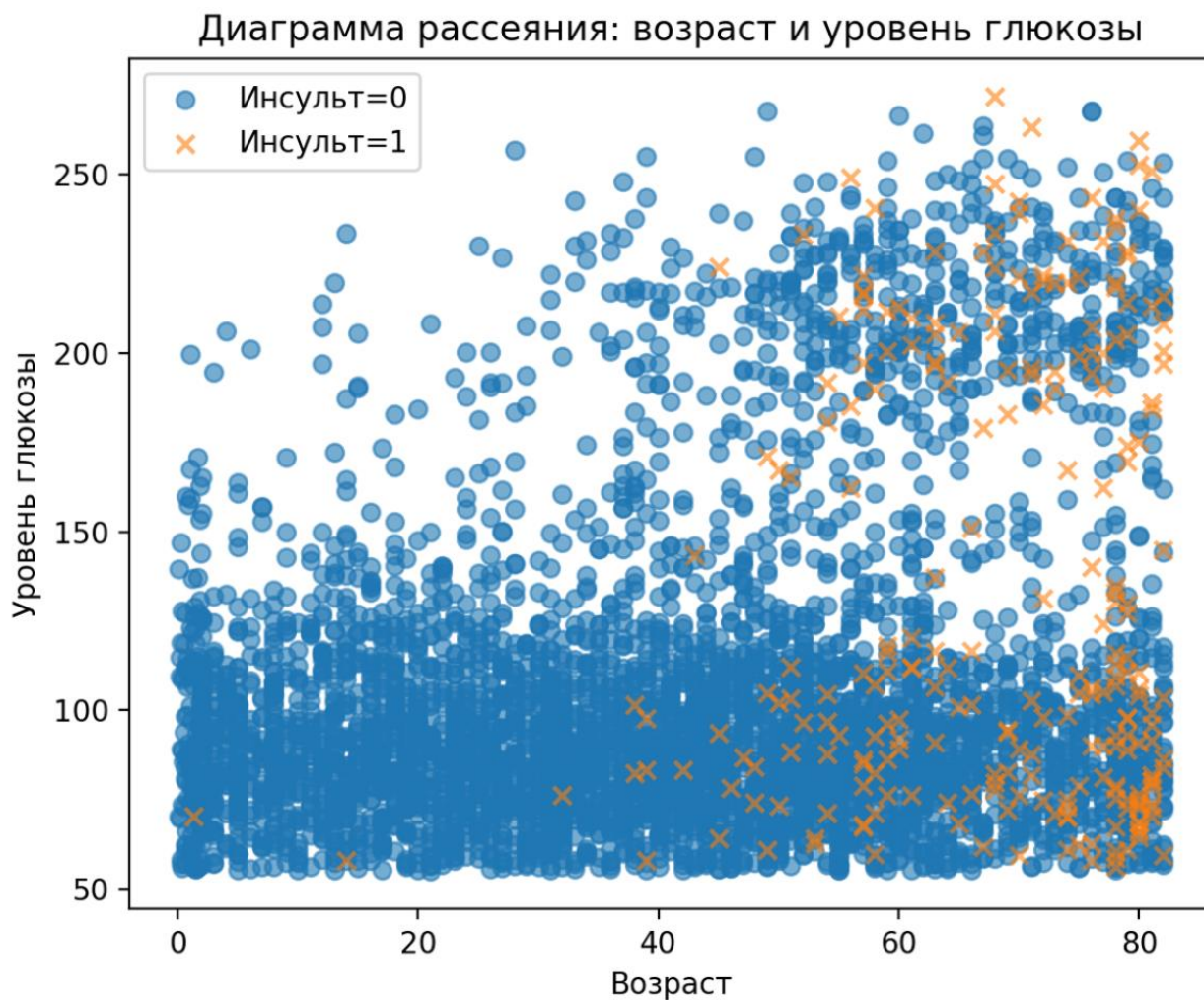


Рисунок 9 – Диаграмма рассеяния «возраст – уровень глюкозы»

После визуализации зависимостей между отдельными признаками дополнительно рассчитывается матрица парных коэффициентов корреляции Пирсона для числовых переменных (age, avg_glucose_level, bmi). Для этого используется метод `corr()` библиотеки `pandas`. Результаты сохраняются в таблицу `reports/корреляция.csv` (Рис. 10).

```
corr = df[["age", "avg_glucose_level", "bmi"]].corr()
corr.to_csv("reports/корреляция.csv")
print(corr)
```

```
,age,avg_glucose_level,bmi
age,1.0,0.2381711141017428,0.32429569983867707
avg_glucose_level,0.2381711141017428,1.0,0.16687572180606186
bmi,0.32429569983867707,0.16687572180606186,1.0
```

Рисунок 10 – Таблица коэффициентов корреляции числовых признаков

Для наглядности матрица корреляций визуализируется в виде тепловой карты (heatmap) с подписями коэффициентов (Рис. 11).

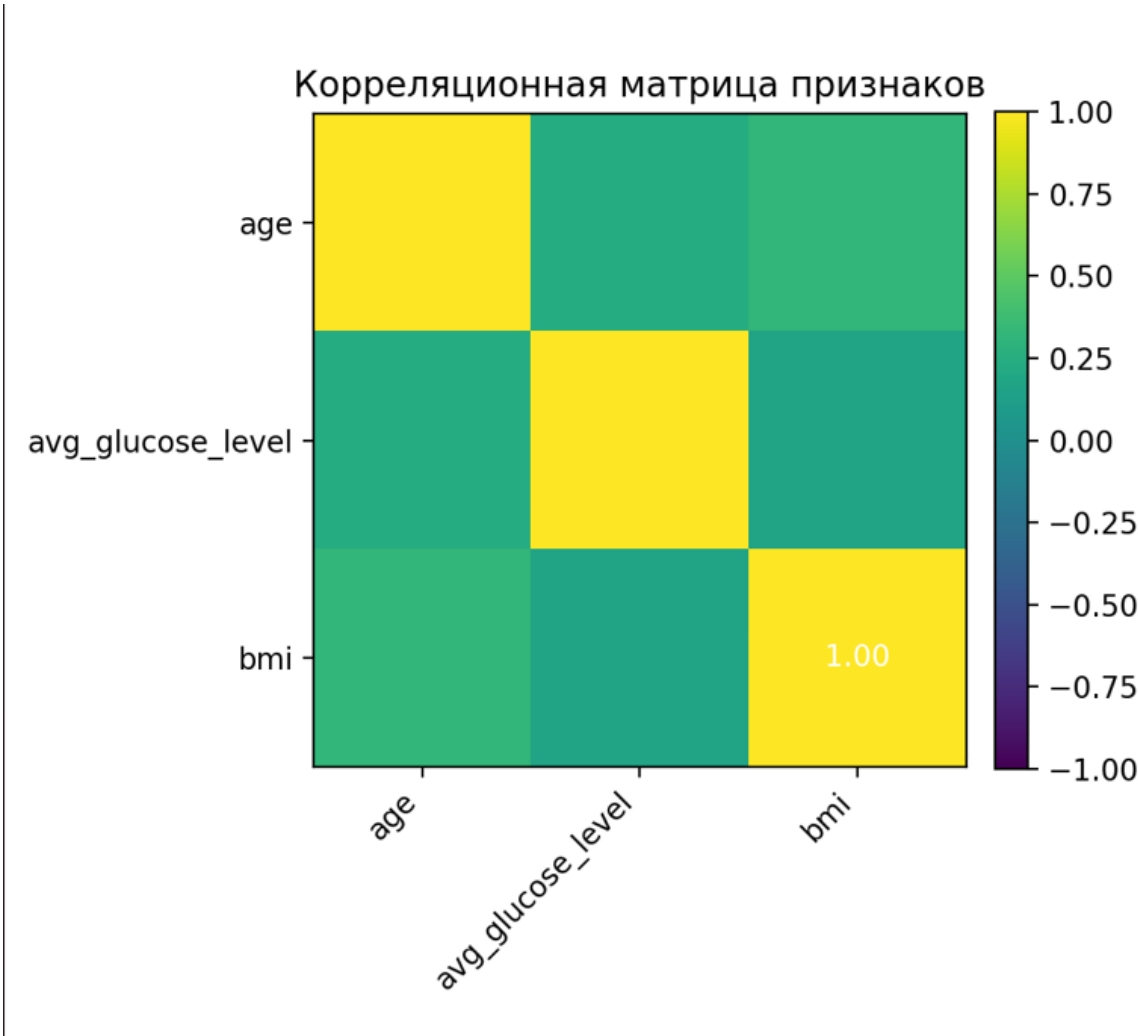


Рисунок 11 – Визуализация корреляционной матрицы

Из тепловой карты видно, что признаки в целом слабо коррелируют между собой, что подтверждает их относительную независимость. Это важно, так как низкая мультиколлинеарность способствует устойчивости модели логистической регрессии.

3.3 Разделение и подготовка данных

Данные делятся на обучающую и тестовую выборки с помощью `train_test_split` (75/25) с параметром `stratify`, что позволяет сохранить соотношение классов (Рис. 11).

```
target = 'stroke'
numeric_features = [c for c in ['age', 'avg_glucose_level', 'bmi'] if c in df.columns]
categorical_features = [c for c in ['gender', 'ever_married', 'work_type', 'residence_type', 'smoking_status'] if c in df.columns]

X = df[numeric_features + categorical_features]
y = df[target].values

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.25, random_state=42, stratify=y
)

# Препроцессор: StandardScaler + OneHotEncoder
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_features),
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)
    ]
)
```

Рисунок 11 – Разделение данных на обучающую и тестовую выборки

Предобработка выполняется через `ColumnTransformer`: числовые признаки нормализуются с помощью `StandardScaler`, категориальные кодируются методом `One-Hot Encoding`.

3.4 Обучение модели и оценка качества

В качестве базовой модели используется логистическая регрессия с балансировкой классов:

```
clf = LogisticRegression(max_iter=1000, class_weight="balanced")
clf.fit(X_train_proc, y_train)
```

Для оценки качества вычисляются метрики точности, полноты, F1 и ROC-AUC (Рис. 12).

```
metrics = {
    'accuracy': float(accuracy_score(y_test, y_pred)),
    'precision': float(precision_score(y_test, y_pred, zero_division=0)),
    'recall': float(recall_score(y_test, y_pred, zero_division=0)),
    'f1': float(f1_score(y_test, y_pred, zero_division=0)),
    'roc_auc': float(roc_auc_score(y_test, y_proba)),
}
```

Рисунок 12 – Результаты метрик качества

Классификационный отчёт (`classification_report`) показывает `precision`, `recall` и `f1-score` для каждого класса (Рис. 13).

	<u>precision</u>	<u>recall</u>	f1-score	<u>support</u>
0	0.984	0.726	0.836	1216
1	0.126	0.774	0.217	62
<u>accuracy</u>			0.728	1278
<u>macro avg</u>	0.555	0.750	0.526	1278
<u>weighted avg</u>	0.943	0.728	0.806	1278

Рисунок 13 – Классификационный отчёт

Матрица ошибок демонстрирует, что основная часть ошибок связана с ложноположительными предсказаниями инсульта (Рис. 14).

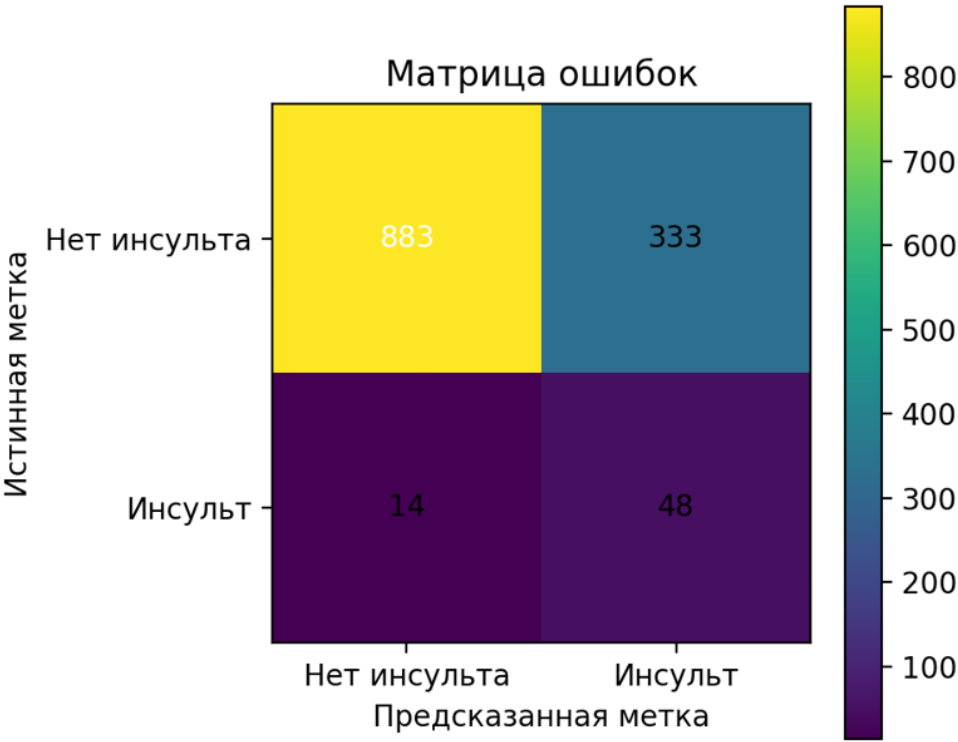


Рисунок 14 – Матрица ошибок

ROC-кривая отражает способность модели различать здоровых пациентов и пациентов с инсультом (Рис. 15).

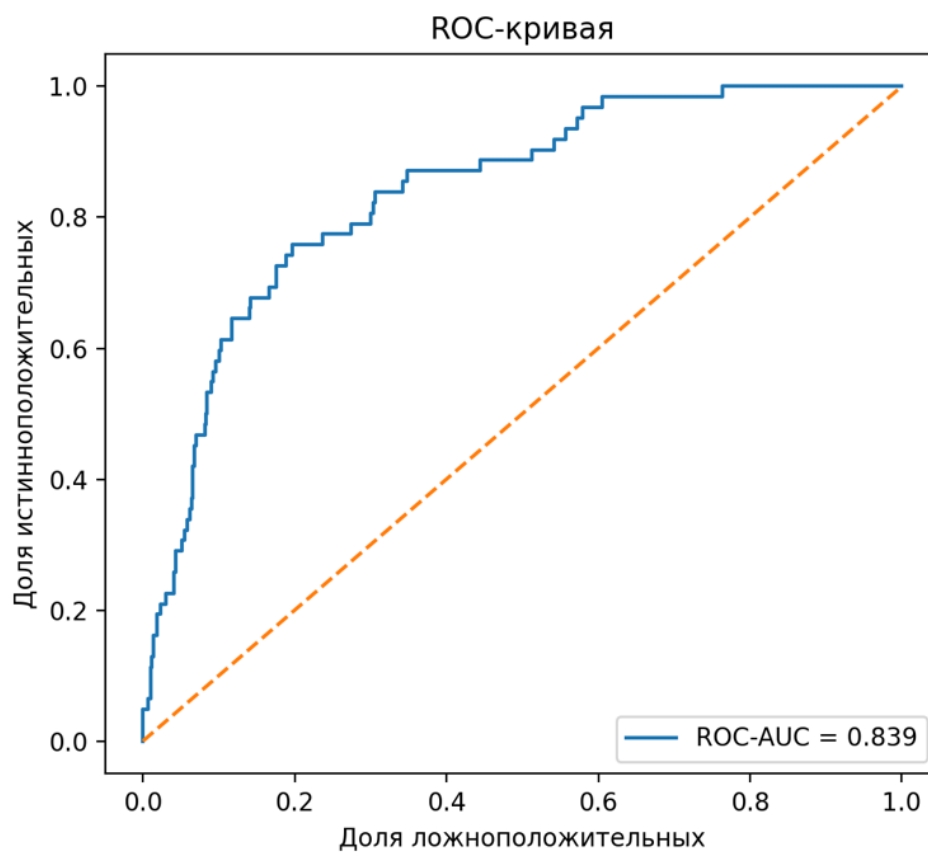


Рисунок 15 – ROC-кривая

4 Выводы

В результате исследования продемонстрирован полный цикл анализа данных на примере медицинского датасета Stroke Prediction. Последовательно выполнены загрузка и подготовка данных, вычисление описательных статистик, построение визуализаций и обучение простой модели классификации.

Основные результаты анализа:

- признаки возраст и уровень глюкозы демонстрируют наиболее выраженное различие между группами, что подтверждается как статистическим описанием, так и графическими методами (гистограммы, scatter plot);
- предобработка данных (заполнение пропусков, нормализация числовых переменных, кодирование категориальных признаков) является необходимым шагом для получения корректных результатов;
- использование элементарных методов разведочного анализа позволяет выявить закономерности — например, более высокая частота инсульта у пожилых пациентов и у лиц с повышенным уровнем глюкозы;
- простая модель логистической регрессии с балансировкой классов достигает приемлемых значений полноты (recall ~ 0.70) и площади под ROC-кривой (ROC-AUC > 0.80), что подтверждает её пригодность для базового прогнозирования;
- матрица ошибок показывает, что модель чаще допускает

ложноположительные предсказания, что связано с особенностями несбалансированных медицинских данных.

Таким образом, исследование подтверждает важность этапа EDA и демонстрирует, что даже минимальные методы позволяют выявлять ключевые факторы риска и строить интерпретируемые модели для задач прогнозирования инсульта.

Библиографический список

1. Донитова В.В., Киреев Д.А., Кобринский Б.А. [и др.] Выявление факторов риска острых нарушений мозгового кровообращения на основе интеллектуального анализа историй болезни // Труды Института системного анализа Российской академии наук. 2023. Т. 73. № 2. С. 111–122.
2. Мишкин И.А., Сахаров А.А. Изучение использования алгоритмов машинного обучения в оценке риска развития сердечно-сосудистых заболеваний // Инновации. Наука. Образование. 2020. № 17. С. 187–202.
3. Рычков А.В., Мочалова Ю.Д. Анализ и обработка данных в медицине // Информационные технологии. Проблемы и решения. 2019. № 3(8). С. 11–16.
4. Благосклонов Н.А., Донитова В.В., Киреев Д.А. [и др.] Лингвистический анализ историй болезни для выявления факторов риска инсульта // Труды Института системного анализа Российской академии наук. 2020. Т. 70. № 3. С. 75–85.
5. Стебаков И.Н., Шутин Д.В., Марахин Н.А. Машинное обучение в реабилитационной медицине и пример классификатора движений пальцев для кистевого тренажёра // Инженерный вестник Дона. 2020. № 6(66). С. 13.
6. Stroke Prediction Dataset // Kaggle. 2021. URL: <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset> (дата обращения: 27.09.2025).