

Александров Александр Александрович, магистрант, Белгородский государственный национальный исследовательский университет, г. Белгород

Гахова Нина Николаевна, научный руководитель, кандидат технических наук, доцент кафедры информатики, Белгородский государственный национальный исследовательский университет, г. Белгород

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМА ОЦЕНКА РИСКА СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ НА ОСНОВЕ ГРАДИЕНТНОГО БУСТИНГА

Аннотация

В работе рассматривается программная реализация первичного скрининга рисков сердечно-сосудистых заболеваний (ССЗ). Описан процесс разработки предиктивной модели с использованием алгоритма CatBoost и её интеграция в интерактивное веб-приложение на базе фреймворка Streamlit. На наборе данных Statlog (Heart) проведена оптимизация гиперпараметров, обеспечившая точность AUC-ROC 0,91. Особое внимание уделено интерпретации результатов через призму клинических рекомендаций по стратификации риска.

Annotation

The paper discusses the software implementation of a primary screening system for cardiovascular risks. It describes the development process of a predictive model using the CatBoost algorithm and its integration into an interactive web application built on the Streamlit framework. Hyperparameter optimization was performed on the Statlog (Heart) dataset, achieving an AUC-ROC accuracy of 0.91. Particular

attention is paid to the interpretation of results through the prism of clinical guidelines for risk stratification.

Ключевые слова: машинное обучение, CatBoost, градиентный бустинг, кардиоваскулярный риск, Streamlit, системы поддержки принятия врачебных решений.

Keywords: machine learning, CatBoost, gradient boosting, cardiovascular risk, Streamlit, clinical decision support systems (CDSS).

Традиционные методы оценки кардиоваскулярного риска зачастую ограничены линейными моделями, которые не в полной мере учитывают сложные взаимосвязи между метаболическими и диагностическими показателями пациента [1]. Цифровая трансформация здравоохранения открывает возможности для внедрения ансамблевых методов машинного обучения в повседневную врачебную практику. Целью данной работы является программная реализация инструмента, обеспечивающего высокую точность оценки риска возникновения сердечно-сосудистых заболеваний и удобство интерпретации результатов для специалиста.

Информационным базисом исследования был выбран набор данных Statlog (Heart) Data Set. Выборка включает 270 клинических случаев, описанных 13 признаками, включая антропометрические данные, показатели артериального давления и результаты ЭКГ-диагностики.

Программная подготовка данных включала следующие этапы:

1) типизация: категориальные признаки (пол, тип боли, ЭКГ и др.) приведены к целочисленному формату для нативной обработки классификатором.

2) валидация: применение стратифицированного разделения выборки (80:20) для сохранения баланса классов.

3) оптимизация: использование метода RandomizedSearchCV для поиска оптимальной комбинации гиперпараметров в пространстве глубины деревьев, скорости обучения и коэффициентов регуляризации.

Для реализации предиктивного модуля выбран алгоритм CatBoostClassifier. Его преимущество заключается в эффективной обработке категориальных данных без расширения признакового пространства, что минимизирует риск переобучения на малых медицинских выборках. Процесс подготовки, обучения и достигнутые метрики представлены на рисунке 1.

```
Размер выборки: 270, признаки: 13
Категориальные признаки (индексы): [1, 2, 5, 6, 8, 10, 12]
Типы данных после подготовки:
  age          float64
  sex          int64
  cp          int64
  trestbps    float64
  chol        float64
  fbs        int64
  restecg     int64
  thalach     float64
  exang       int64
  oldpeak    float64
  slope      int64
  ca         float64
  thal       int64
dtype: object
RandomizedSearchCV результаты:
Лучшие параметры: {'learning_rate': 0.01, 'l2_leaf_reg': 3, 'iterations':
  500, 'depth': 4, 'border_count': 128}
Лучшее CV ROC-AUC: 0.9135
ROC-AUC на валидации: 0.8819
Accuracy на валидации: 0.8519
F1-score на валидации: 0.8333
```

Рисунок 1 – Результаты подготовки данных и обучения модели

Итоговое значение ROC-AUC составило 0,913, что подтверждает высокую разделительную способность алгоритма в задачах бинарной классификации состояния пациента.

Программный комплекс реализован на языке Python с использованием библиотеки Streamlit. Схема компонентов программного комплекса представлена на рисунке 2.

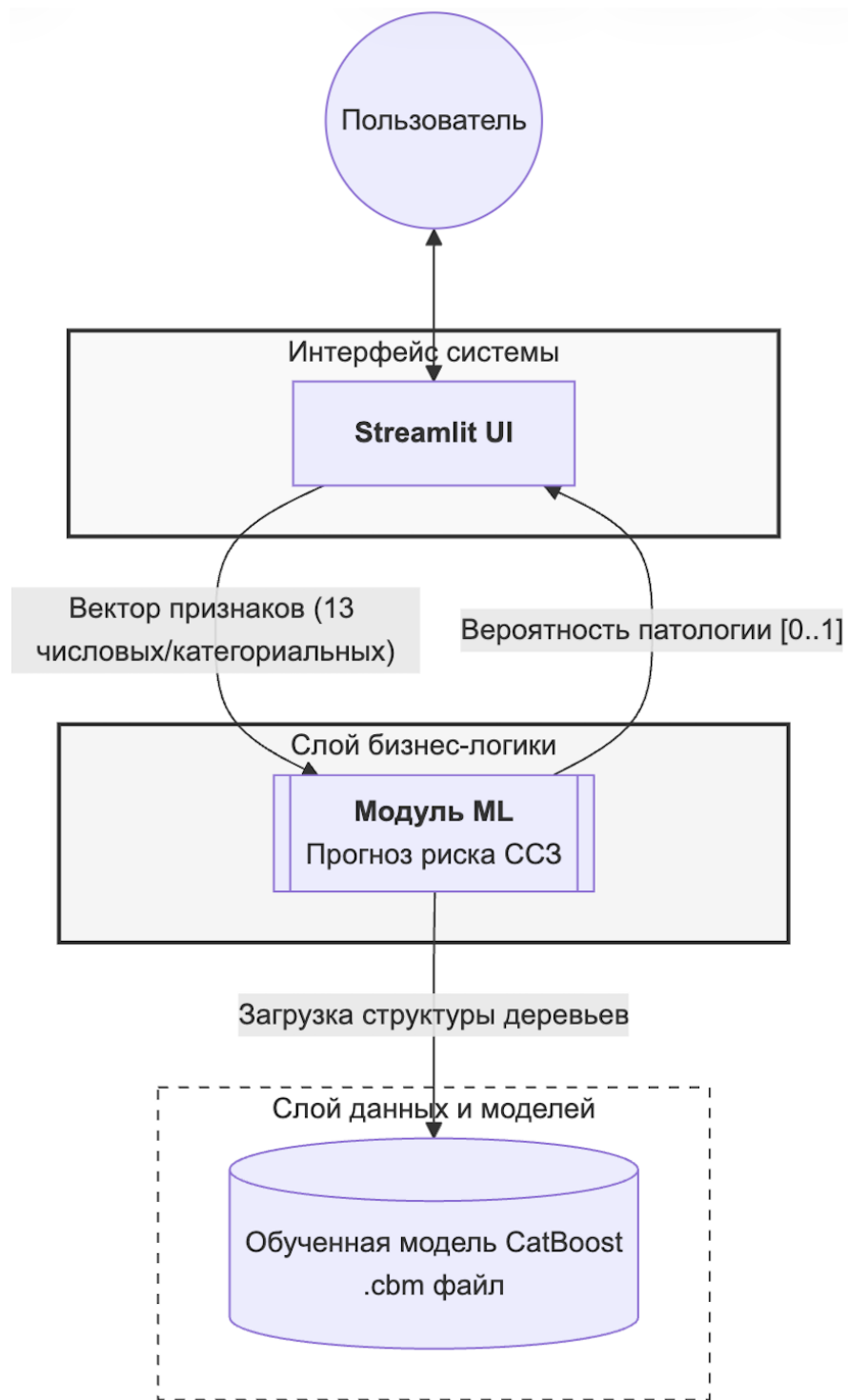


Рисунок 2 – Схема компонентов программного комплекса

Интерфейс программы (рисунок 3) включает:

- информационную панель (Sidebar): визуализация метрик качества (AUC, F1-score) и текущих параметров модели.
- диагностический блок: интерактивные поля ввода клинических параметров.

Сведения о модели

Статус: Загружена из файла

ROC-AUC: **0.913** F1-Score: **0.835**

Ассигнатура: **0.852**

Детали обучения

Алгоритм: CatBoost Classifier

Данные: Statlog (Heart) Data Set (UCI)

Выборка: 270 строк

Гиперпараметры

Разработка системы поддержки принятия
врачебных решений (СППВР)

Оценка риска возникновения сердечно-сосудистых заболеваний человека

Показатели здоровья

| | | | | | |
|----------------------------------|--------|---------------------------------------|---------|--|----------|
| Возраст (лет) | 40,00 | Пол | Мужчина | Тип боли | Типичная |
| Артериальное давление (мм.рт.ст) | 120,00 | Холестерин (мг) | 200,00 | Уровень глюкозы натощак > 120 | Нет |
| ЭКГ | Норма | Частота сердечных сокращений (уд/мин) | 0,00 | Стенокардия при нагрузке | Нет |
| Депрессия сегмента ST (мм) | 0,50 | Наклон сегмента ST | Вверх | Количество сосудов окрашенных флюороскопией () | 0,00 |
| Состояние сосудов | Норма | | | | |

Рассчитать риск

Уровень риска: Низкий

Вероятность патологии: 10.0%

При низком риске сердечно-сосудистых заболеваний рекомендуется тактика активного наблюдения и коррекция образа жизни согласно рекомендациям ESC по первичной профилактике

Расчет произведен на основе классификатора CatBoost Classifier

Рисунок 3 – Интерфейс программы

Логика интерпретации выходных данных опирается на расчет апостериорной вероятности наличия патологии. В программной реализации применена трехфакторная цветовая индикация, соответствующая международным протоколам стратификации риска [2,3]:

- низкий риск (вероятность < 30%, зеленая зона): соответствует тактике активного наблюдения и коррекции образа жизни согласно рекомендациям ESC по первичной профилактике.

- средний риск (30% – 70%, оранжевая зона): пограничное состояние, требующее расширенной диагностики и учета дополнительных клинических факторов (семейный анамнез, биомаркеры).

- высокий риск (вероятность > 70%, красная зона): критический уровень, требующий немедленного внимания профильного специалиста. Установление порога в 70% обеспечивает высокую специфичность системы, минимизируя ложноположительные срабатывания [2].

Разработанный программный инструмент позволяет автоматизировать процесс первичного кардиоскрининга, переводя сложные математические расчеты в интуитивно понятную для пользователя форму. Сочетание высокой точности (AUC 0,913) и клинически обоснованной интерпретации рисков делает данную реализацию перспективной базой для создания полноценных систем поддержки принятия врачебных решений.

Список литературы

1. Goldstein B. A., Navar A. M., Carter R. E. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges // *European Heart Journal*. – 2017. – Vol. 38, № 23. – P. 1805–1814.
2. Arnett D. K. et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines // *Circulation*. 2019. Vol. 140, No. 11.
3. Visseren F. L. J. et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice // *European Heart Journal*. 2021. Vol. 42, Iss. 34. P. 3227–3337.
4. Аджиева З. И., Кубанов М. Б., Аджиева Х. М. Цифровая трансформация здравоохранения: вызовы и новые возможности // Альманах «Крым». 2023. №40. URL: <https://cyberleninka.ru/article/n/tsifrovaya-transformatsiya-zdravoohraneniya-vyzovy-i-novye-vozmozhnosti> (дата обращения: 19.04.2026).
5. Statlog (Heart) Data Set. UCI Machine Learning Repository [Электронный ресурс]. — URL: <https://archive.ics.uci.edu/ml/datasets/heart+disease> (дата обращения: 20.04.2026).

References

1. Goldstein B. A., Navar A. M., Carter R. E. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges // *European Heart Journal*. – 2017. – Vol. 38, № 23. – P. 1805–1814.
2. Arnett D. K. et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines // *Circulation*. 2019. Vol. 140, No. 11.
3. Visseren F. L. J. et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice // *European Heart Journal*. 2021. Vol. 42, Iss. 34. P. 3227–3337.
4. Adzhieva Z. I., Kubanov M. B., Adzhieva Kh. M. Digital transformation of healthcare: challenges and new opportunities // *Almanac "Crimea"*. 2023. No. 40. URL: <https://cyberleninka.ru/article/n/tsifrovaya-transformatsiya-zdravoohraneniya-vyzovy-i-novye-vozmozhnosti> (accessed: 19.04.2026).
5. Statlog (Heart) Data Set. UCI Machine Learning Repository [Electronic resource]. — URL: <https://archive.ics.uci.edu/ml/datasets/heart+disease> (accessed: 20.04.2026).