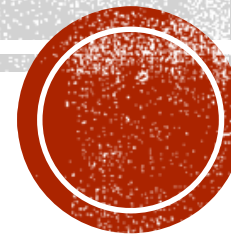


ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ФИЗИОЛОГИИ ТРУДА



Выполнил: к.м.н. Сапожников Кирилл Викторович

Руководитель: к.э.н. доцент Заграновская Анна Васильевна

ПРОБЛЕМА ДОКАЗАТЕЛЬНОСТИ В ФИЗИОЛОГИИ ТРУДА

Исследователь рассказывает о результатах своей работы научному сообществу при помощи публикаций.

Так возможен критический анализ его выводов – необходимый этап на пути принятия сообществом мнения исследователя.

Для этого выводы исследования должны основываться на объективном и полноценном исследовании.



ПРОБЛЕМНЫЕ ВОПРОСЫ

- Данные в физиологии труда представляют собой по большей части панельные данные малых групп (10-20 человек), имеющие высокую индивидуальную вариабельность, часто имеющие нелинейные взаимосвязи. Их распределение редко соответствует нормальному. Имеются пропуски в данных (дефект сбора, выброс, появление новых методик).
- Количественные данные анализируются с применением простого статистического инструментария (анализ средних), без учета их происхождения (счетные данные, **time-to-event**, временной ряд). К сложным методам (включая машинное обучение) прибегают нечасто.
- Для задач классификации чаще всего используются объяснимые линейные классификаторы. Нелинейные классификаторы используются неохотно ввиду трудностей интерпретации, легко интерпретируемый **KNN** не позволяет работать с пропущенными данными, а методы заполнения (включая **MICE**) далеко не всегда дают желаемый эффект.
- Центральные тенденции и дисперсия данных часто представляют в виде медианы и квартилей (иногда – с размахом выборки), что в значительной степени затрудняет обобщение результатов разных исследований (мета-анализ).

**ПРОПУСК ЗНАЧИМЫХ РЕЗУЛЬТАТОВ, ПОТЕРЯ ИНФОРМАТИВНОСТИ, НИЗКИЙ
УРОВЕНЬ ДОВЕРИЯ К ПУБЛИКАЦИЯМ!**

ЦЕЛЬ И ЗАДАЧИ

- Цель: определить область применения методов машинного обучения в сфере физиологии труда

Задачи:

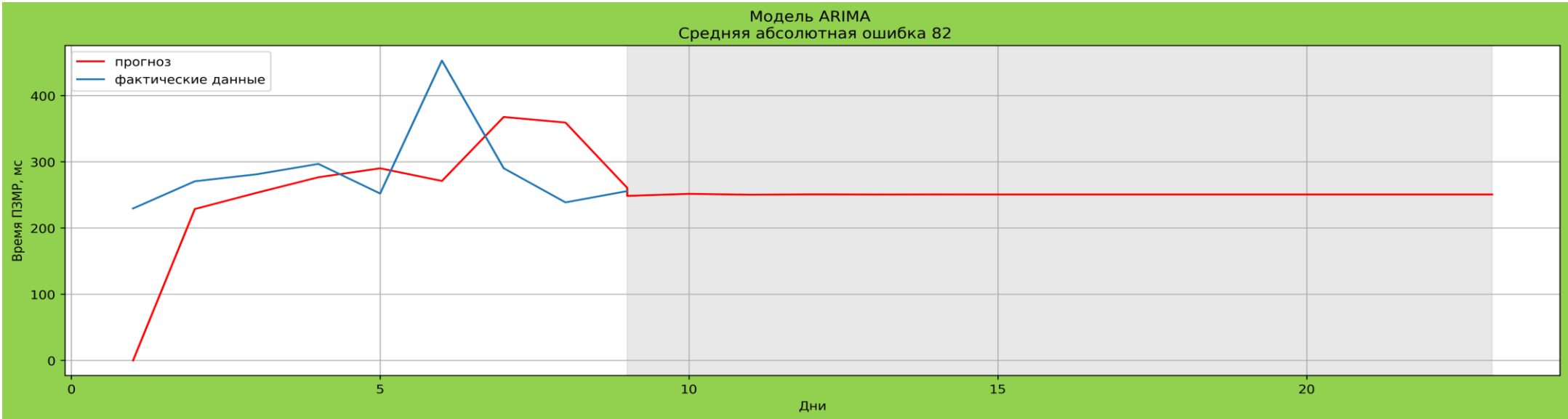
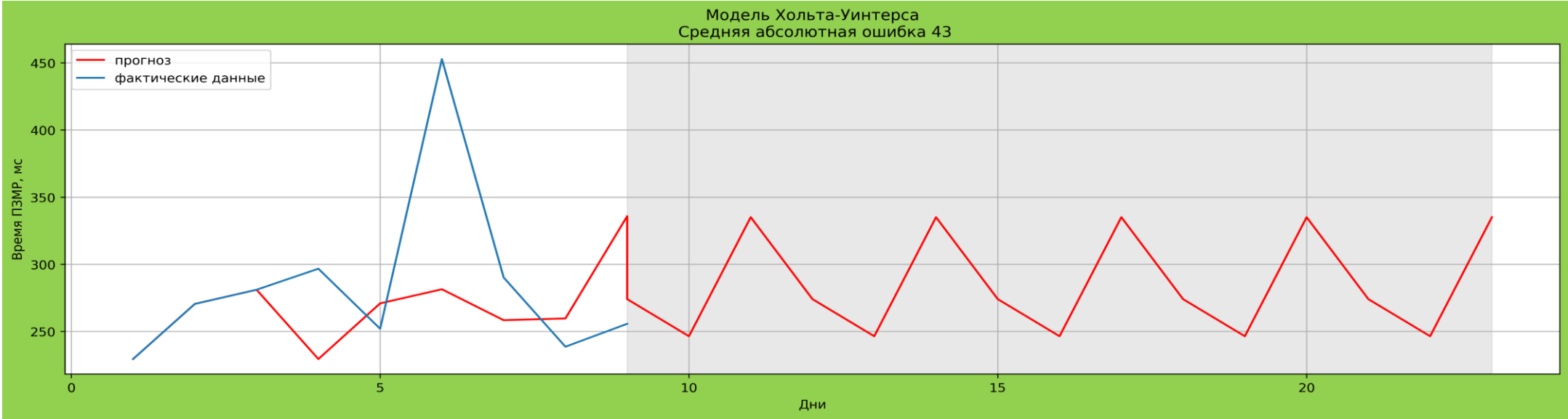
- Апробировать методы машинного обучения для анализа данных типа временных рядов, счетчиков и **time-to-event**
- Разработать аналог **knp**, позволяющий работать с данными, содержащими пропуски
- Разработать алгоритм аппроксимации среднего и стандартного квадратического отклонения по медиане, квартилям и размаху выборки

ВРЕМЕННЫЕ РЯДЫ В ФИЗИОЛОГИИ ТРУДА

- ▶ Показатель: динамика скорости реакции на зрительный стимул у оператора пульта управления, миллисекунды.
- ▶ Частная задача: прогноз динамики адаптации оператора к условиям работы (индивидуальная оценка)
- ▶ Методы:
 - ▶ Интегрированная модель авторегрессии — скользящего среднего (ARIMA)
 - ▶ Модель Хольта-Уинтерса
 - ▶ Анализ трендов

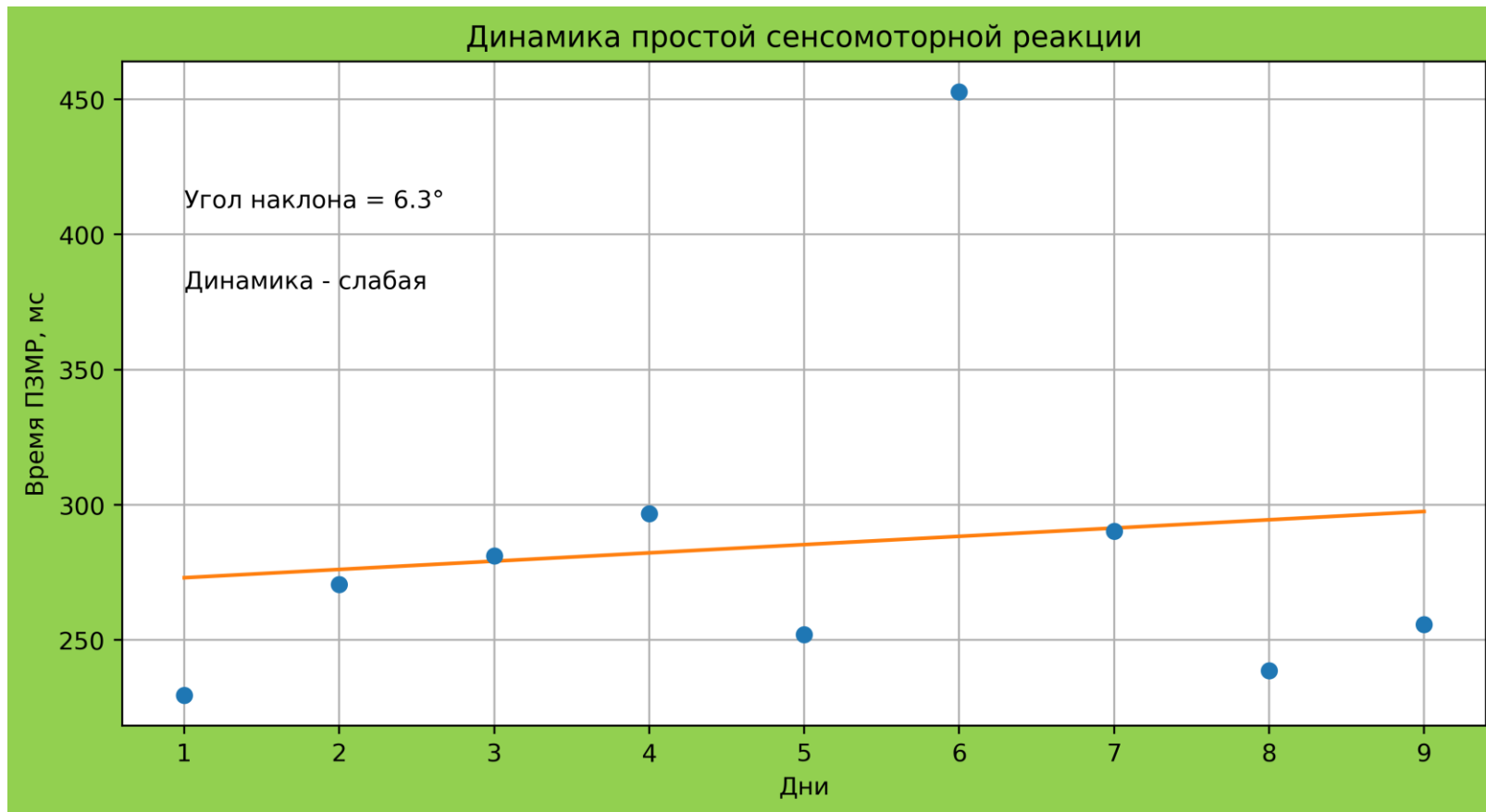


НЕ ВСЕ МОДЕЛИ МОГУТ ПРОГНОЗИРОВАТЬ АДАПТАЦИОННЫЙ ПРОЦЕСС



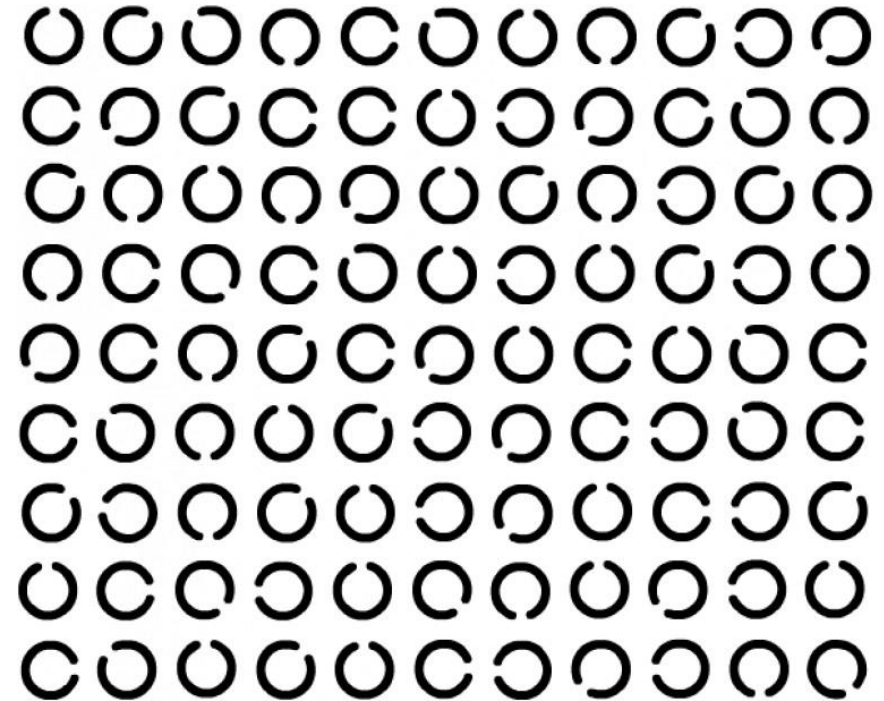
РЕШЕНИЕ – ВЫДЕЛЕНИЕ ЛИНЕЙНОГО ТРЕНДА, АНАЛИЗ УГЛА ЕГО НАКЛОНА

```
from scipy.optimize import curve_fit
```



СЧЕТНЫЕ ДАННЫЕ В ФИЗИОЛОГИИ ТРУДА

- ▶ Показатель: количество ошибок при корректурной пробе Ландольта, штук за пробу
- ▶ Частная задача: сравнение числа операторских ошибок в группе, принимающей препарат для улучшения операторской работоспособности, и группе плацебо
- ▶ Методы:
 - ▶ Пуассоновская регрессия
 - ▶ Отрицательная биномиальная регрессия



РЕШЕНИЕ:

- ▶ Пуассоновская регрессия при равенстве среднего и дисперсии

```
...family=statsmodels.families.Poisson()
```

- ▶ Отрицательная биномиальная регрессия при невыполнении условия

```
...family= statsmodels.families.NegativeBinomial()
```

- ▶ Достоинство: легко интерпретируемый вывод, скорректированный по включенным ковариатам:

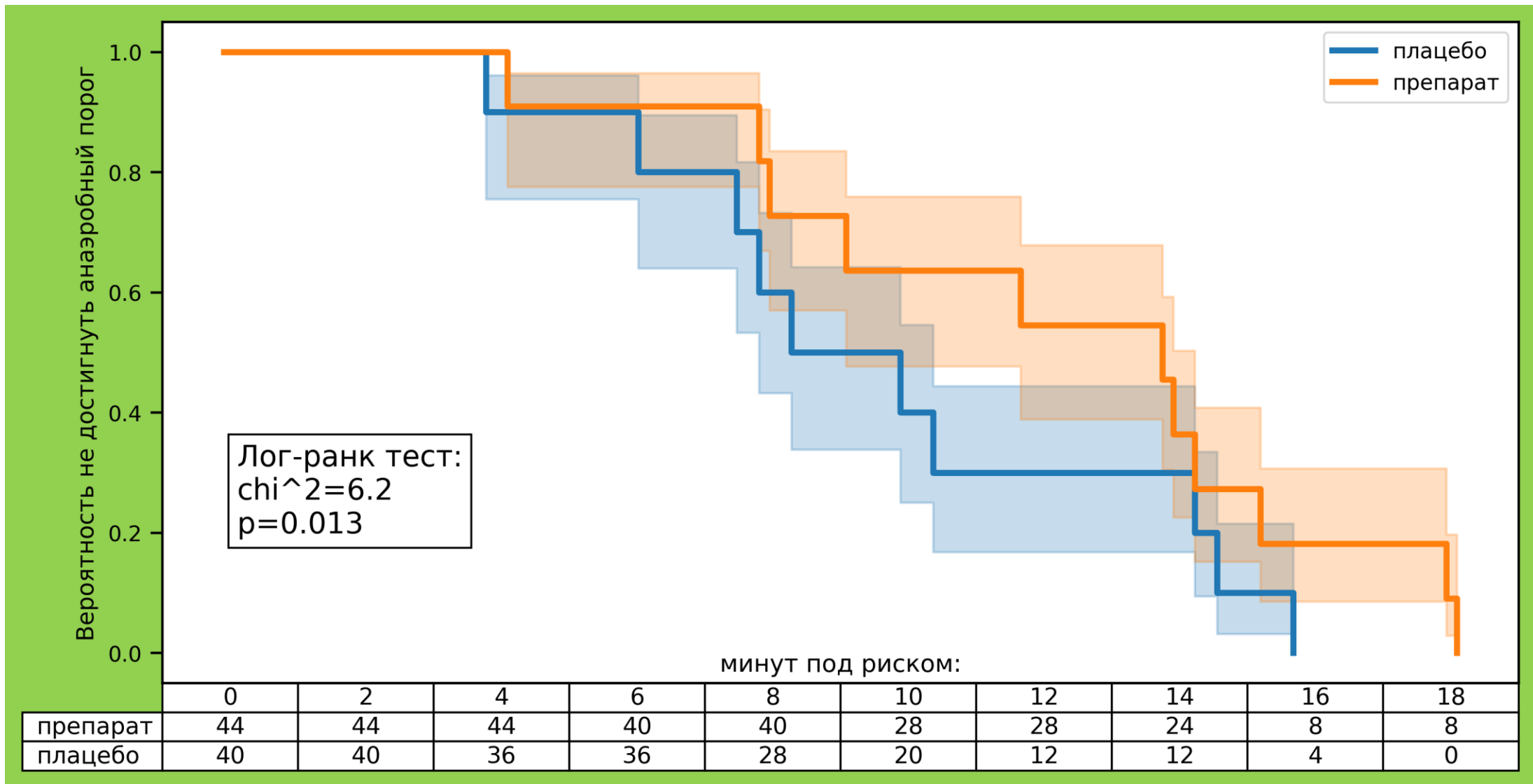
```
IRR = 1,99 (1,35; 2,93)
```

ДАННЫЕ TIME-TO-EVENT

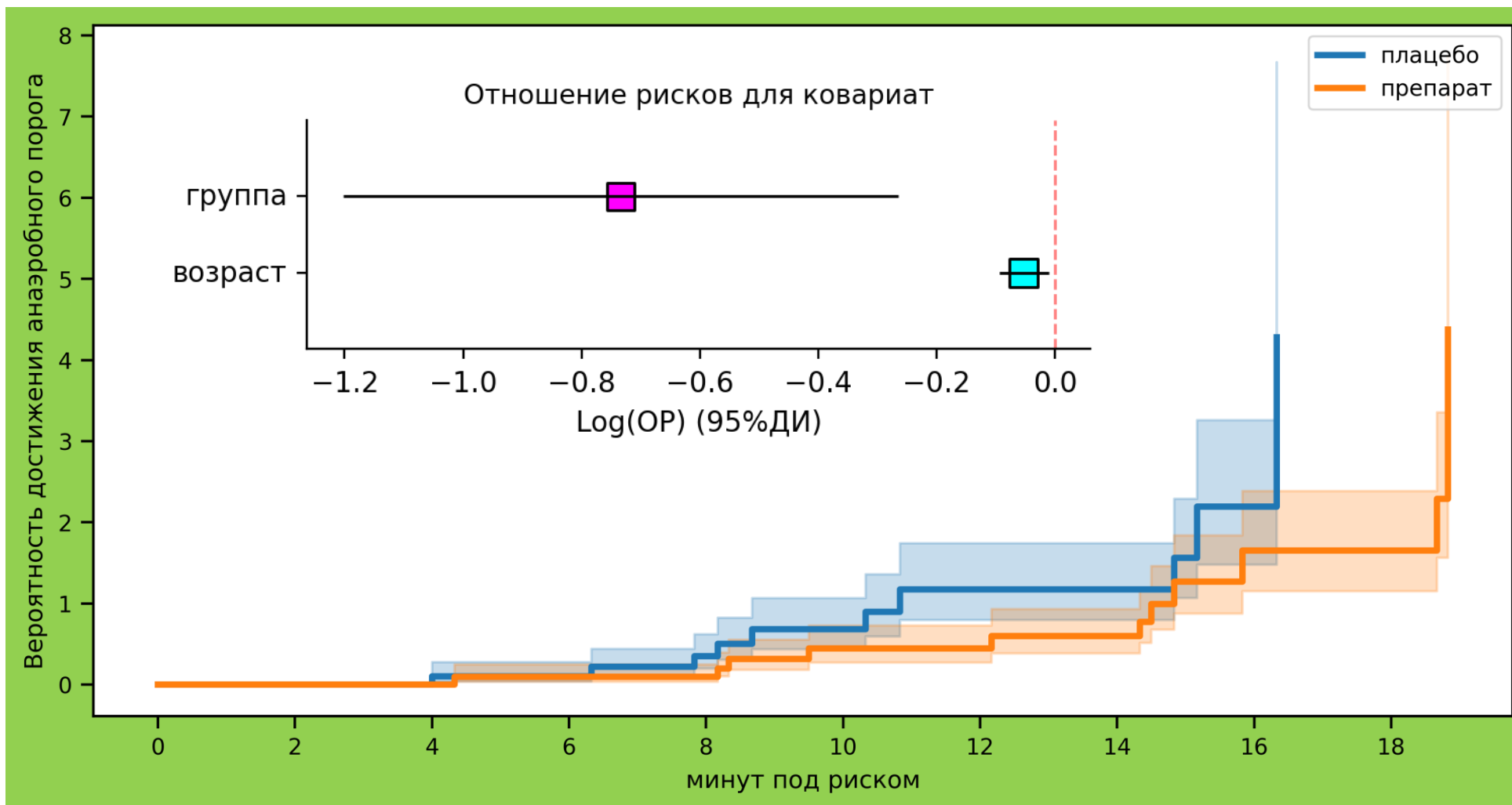
- ▶ Показатель: время до достижения анаэробного порога при эргоспирометрии, минут.
- ▶ Важно: зависит от возраста!
- ▶ Частная задача: сравнение данного показателя в группе, принимающей препарат для улучшения физической работоспособности, и группе плацебо
- ▶ Методы:
 - ▶ Кривые Каплан-Мейера
 - ▶ Пропорциональные модели Кокса



КРИВЫЕ КАПЛАН-МЕЙЕРА



ПРОПОРЦИОНАЛЬНЫЕ МОДЕЛИ КОКСА



РАЗРАБОТКА АНАЛОГА KNN

Новая метрика, устойчивая к пропускам, с обобщением путем взятия среднего. Есть возможность добавления весов исследователем.

$$m_{i,j}^k = \frac{w^k + (1 - |a_i^k - b_j^k|)}{2}, \text{ где}$$

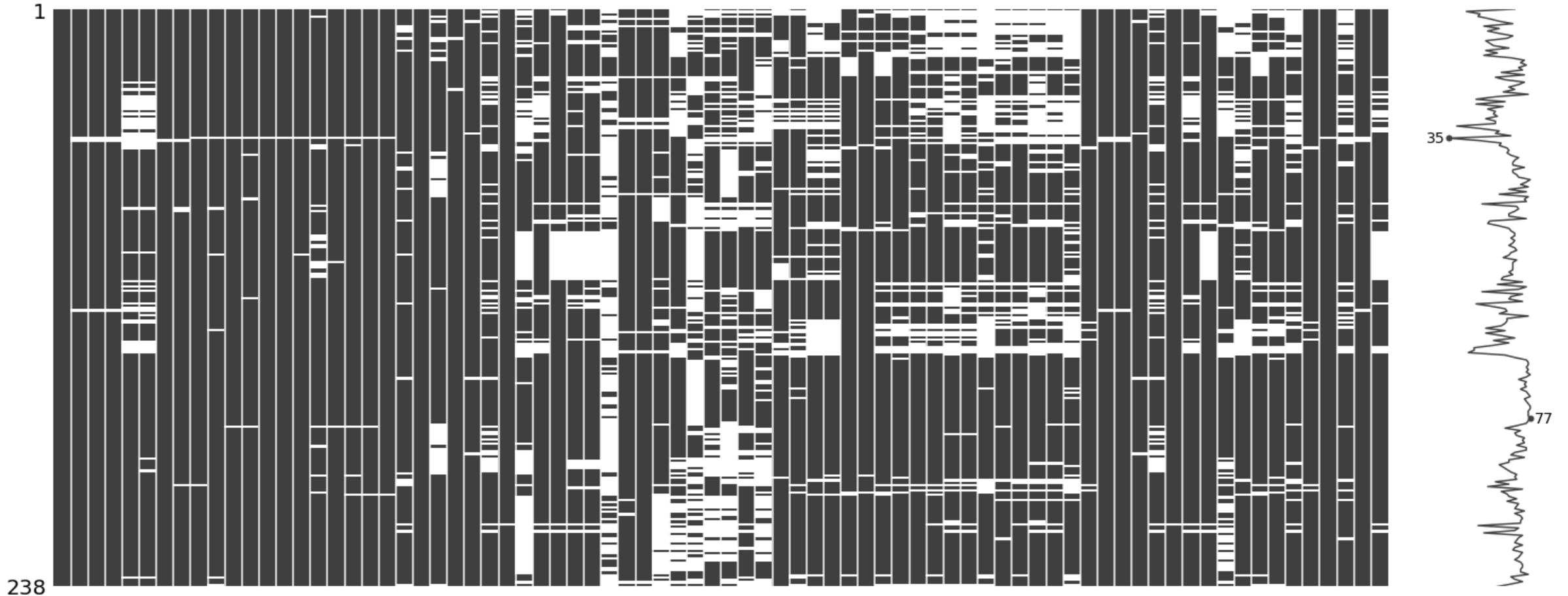
- w^k – вес k -го показателя. задается исследователем от 0 до 1, где 0 – не значимый вес, 1 – максимально значимый вес.
- $|a_i^k - b_j^k|$ – разница между нормализованными k -ми показателями нового случая a_i и случая b_i из базы. Нормализация – минимакс.
- $m_{i,j}^k$ – степень близости i -го нового случая к j -му случаю в базе данных по показателю k . От 0 до 1, где 0 – низкая степень близости, 1 – максимальная близость

Обобщение: вычисление средней величины расстояний по всем доступным показателям между новым случаем и случаем из базы.

Далее – как в классическом knn.

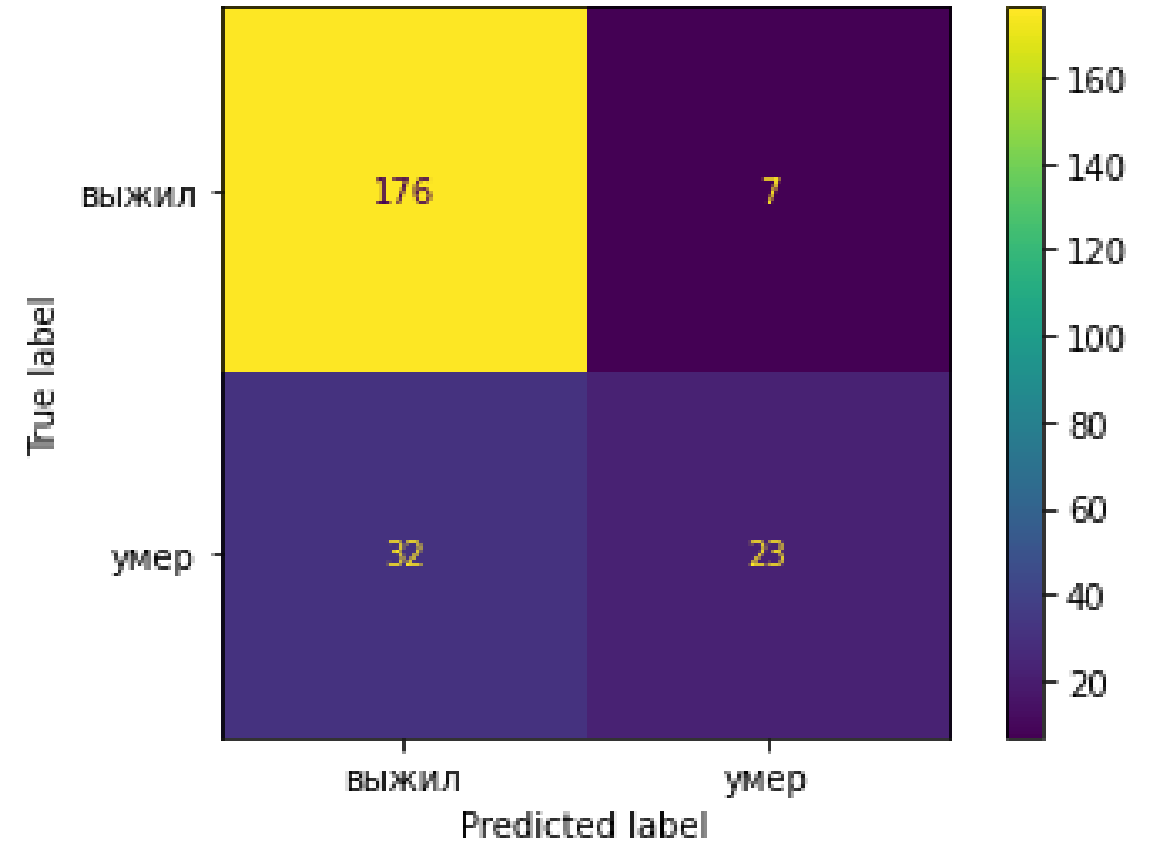
АПРОБАЦИЯ

- Датасет с реальными данными. Число полных строк = 0!



АПРОБАЦИЯ

Модель	Точность	СКО
логистическая регрессия	-	-
линейный дискриминантный анализ	-	-
knn	-	-
найвный Байес	-	-
решающее дерево	-	-
not-knn	0,84	0,37



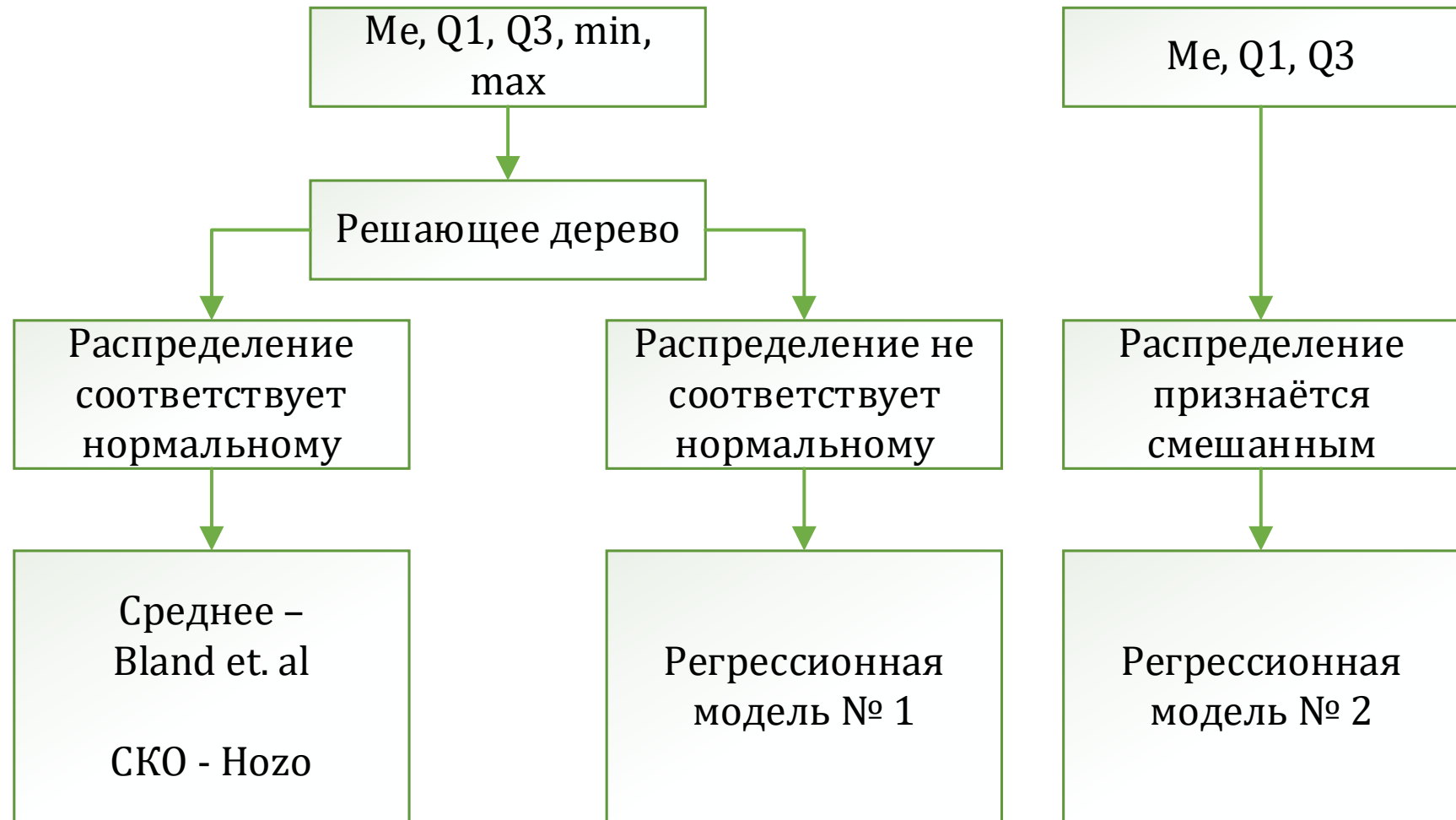
АЛГОРИТМ АППРОКСИМАЦИИ СРЕДНЕГО И СКО

- Актуальность: имеющиеся решения (Кокран, Wan et al., Hozo et al., Bland et al.) хорошо работают только на нормальном распределении.
- Данные: генерация данных смешанного распределения (нормальное и скошенные - логнормальное, бета-распределение, экспоненциальное, Вейбулла) – 3 итерации:
 - Для обучения моделей
 - Для их валидации
 - Для апробации разработанного решения
- Предлагаемые метрики: отклонение медианы от середины межквартильного интервала, середины размаха выборки, отклонение квартилей от их ожидаемых значений
- Методы: регрессия с регуляризацией «elastic net», деревья классификации

АПРОБАЦИЯ АЛГОРИТМА

- Решающее дерево для определения типа распределения

(accuracy = 94%,
precision = 94%,
recall = 94%)



- Линейная регрессия с регуляризацией типа «elastic net» для расчета среднего и СКО при скошенном ($R^2 = 98\%$ и 93%) и смешанном распределении ($R^2 = 98\%$ и 83%)
- Решение в целом:
 - $R^2 = 99\%$ и 98% для среднего и СКО соответственно (есть **min, max**)
 - $R^2 = 98\%$ и 89% для среднего и СКО соответственно (нет **min, max**)

ВЫВОДЫ

- Для анализа коротких временных рядов рекомендуется использовать построение линейного тренда и оценку угла его наклона для приблизительного прогноза динамики адаптации оператора
- Для анализа счетных данных и получения показателя **Incidence rare ratio** рекомендуется использовать пуассоновскую (отрицательную биномиальную) регрессию с ковариатами.
- Для анализа времени до наступления события рекомендуется использовать кривые Каплан-Мейера с анализом отношения риска, уточненного по имеющимся ковариатам

ВЫВОДЫ

- Разработанный аналог `knn` может применяться в данных, имеющих пропуски, когда методы заполнения не применимы.
- Разработанный алгоритм аппроксимации среднего и СКО показал хорошие результаты по точности, его применение рекомендуется при условии контроля результата (проведения анализа чувствительности).