Министерство образования и науки Российской Федерации
РГУ нефти и газа (НИУ) имени И.М. Губкина

Факультет автоматики и вычислительной техники
Кафедра автоматизированных систем управления

Применение нейронных сетей в задаче классификации уровней загрязнения воздуха

Магистерская диссертация

Выполнил: студент гр. АСМ-23-05 А.С. Семенов

Руководитель: д.э.н., профессор А.А. Алетдинова

Москва, 2025

Актуальность

Загрязнение воздуха

Роль машинного обучения

Необходимость точного мониторинга

Практическая значимость

Цель работы

Разработка универсальной модели для решения задачи классификации уровней загрязнения воздуха на основе нейронных сетей

Задачи

- Провести обзор и анализ существующих подходов к решению задачи классификации уровней загрязнения воздуха
- Сформулировать математическую постановку задачи классификации уровней загрязнения воздуха
- Провести сбор, предобработку и анализ данных
- Выбрать и обучить классические модели машинного обучения
- Разработать модель на основе нейронных сетей и провести её оценку
- Провести сравнительный анализ метрик качества предложенных моделей
- Оценить устойчивость и обобщающую способность разработанных моделей

Подходы к решению задачи классификации

Классические алгоритмы ML	Ансамблевые модели	Нейронные сети
• Применяются наивный Байес, метод ближайших соседей (KNN), дерево решений, логистическая регрессия, SVM • Хорошие результаты для базовых задач классификации, простая реализация	• CatBoost, XGBoost, Random Forest • Хорошая точность на табличных данных, устойчивость к шуму • Пример: CatBoost для мониторинга загрязнения в Джакарте	• Рекуррентные сети (RNN), полносвязные и сверточные сети • Учёт временных и пространственных зависимостей • Пример: SMOTEDNN для классификации уровней загрязнения в Саудовской Аравии

Постановка задачи классификации

Классификация уровня загрязнения воздуха в заданном городе

Использование исторических данных и различных факторов, таких как выбросы загрязняющих веществ

Определение категории уровня загрязнения:

низкая, повышенная, высокая, очень высокая

Описание исходных данных

Источник данных	Объем данных	Основные атрибуты
• Набор данных предоставлен Центральным советом по контролю за загрязнением окружающей среды Индии и охватывает период с 2015 по 2020 годы	• Включает 29 531 наблюдение, охватывающее 26 различных городов	• Дата наблюдения • Город • Уровень загрязнения воздуха (низкий, повышенный, высокий, очень высокий) • Объем выбросов различных веществ, таких как: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene и Xylene

Обоснование использования нейронных сетей

- Обработка многомерных данных
- Нелинейные зависимости
- Адаптивное обучение
- Обработка временных рядов
- Гибкость и масштабируемость

Математическая постановка задачи классификации

Цель: Определить уровень загрязнения воздуха в заданном городе по историческим данным

Пространство входных и выходных данных:

$$f: X \to Y, \qquad Y = \{1, 2, 3, 4\}$$

Выборка:

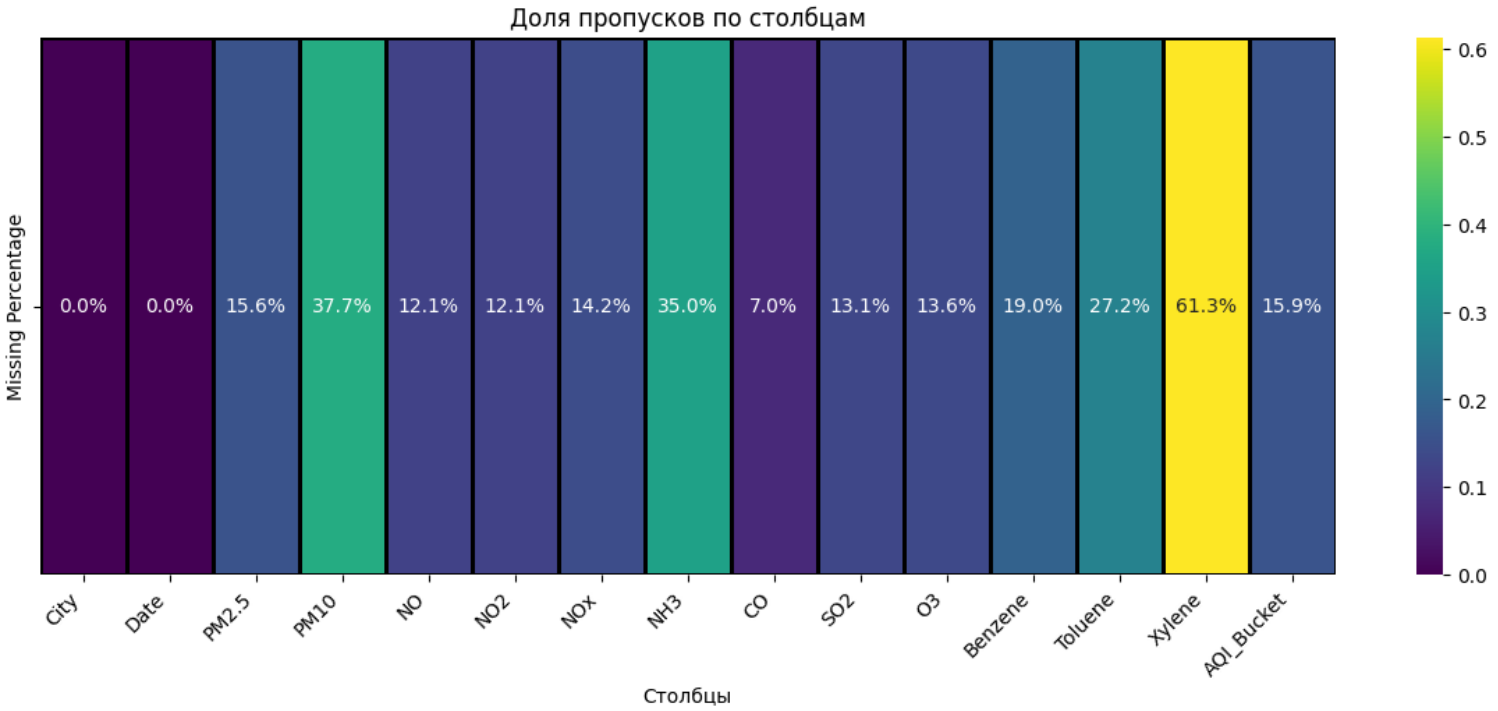
$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^n, \qquad y_i \in Y$$

Целевая функция:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)$$

Предобработка и анализ данных

Работа с пропусками



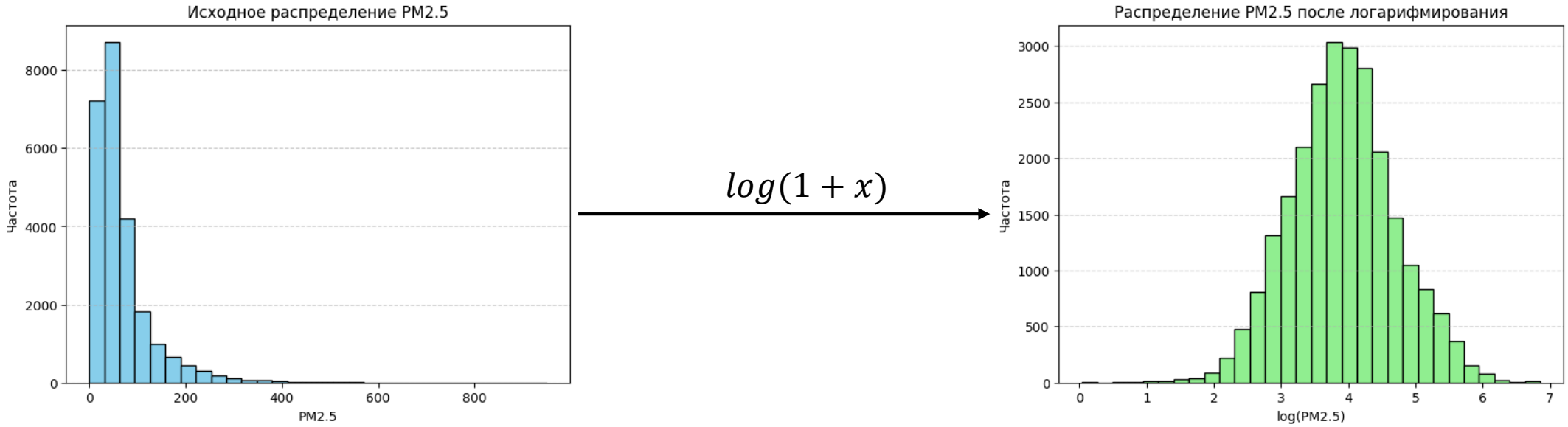
Признак "**Xylene**" (Ксилол) имеет **61.3%** пропусков. Принято решение **удалить** его, чтобы избежать искажения данных

Числовые признаки: пропуски заполнены **медианой по городу** для сохранения устойчивости к выбросам

Категориальные признаки: пропуски заполнены **модой по городу**, чтобы сохранить преобладающее значение

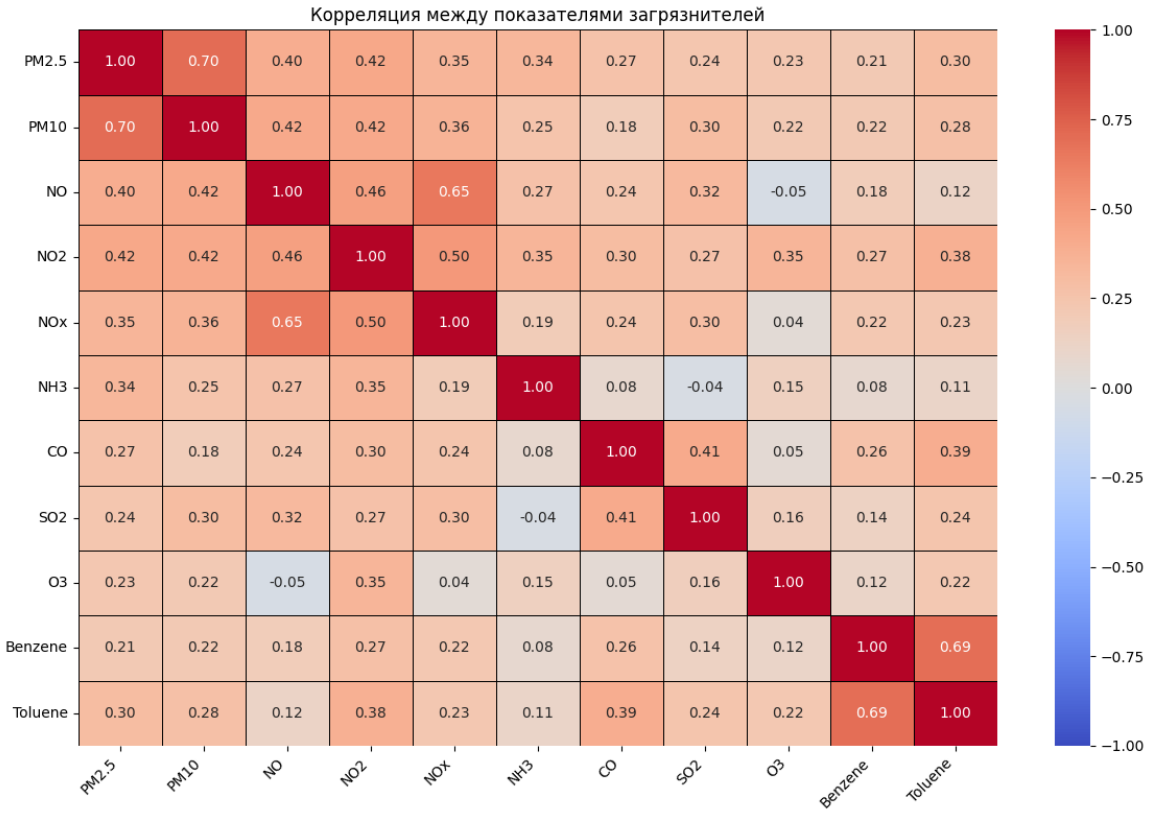
Предобработка и анализ данных

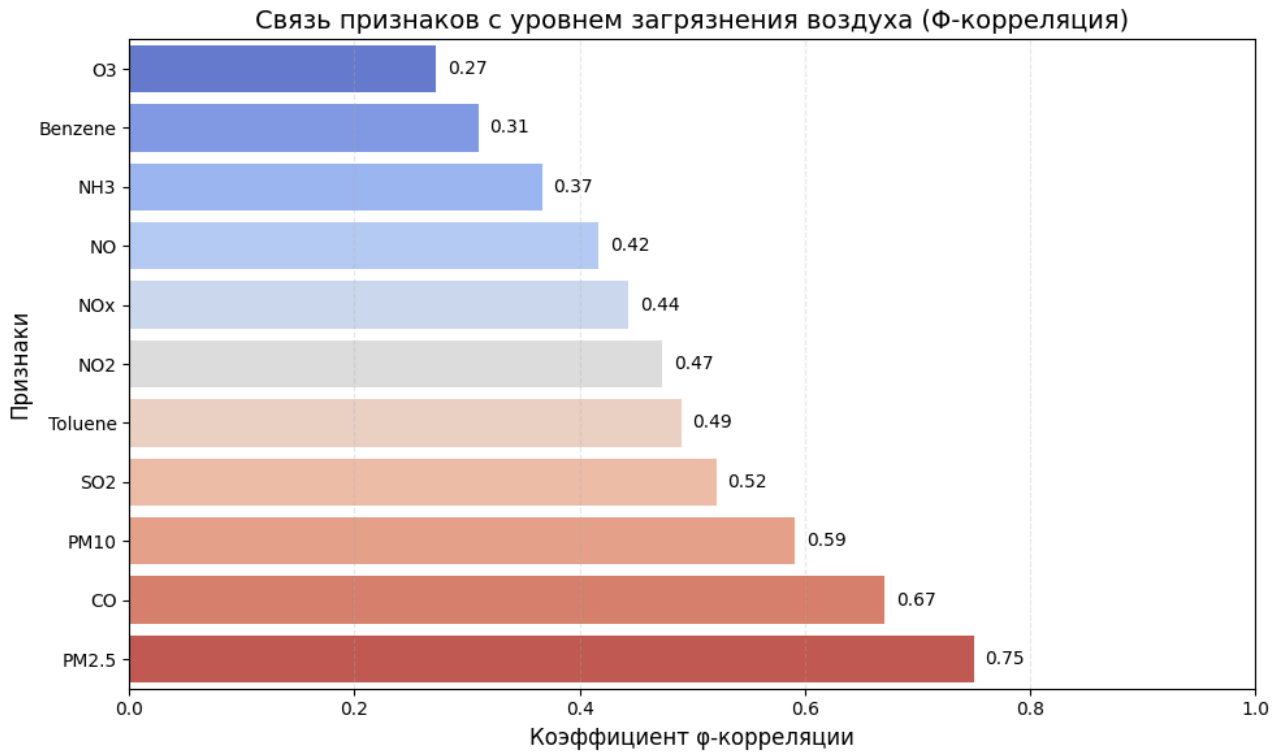
Улучшение распределений на примере признака PM2.5



Предобработка и анализ данных

Оценка корреляции



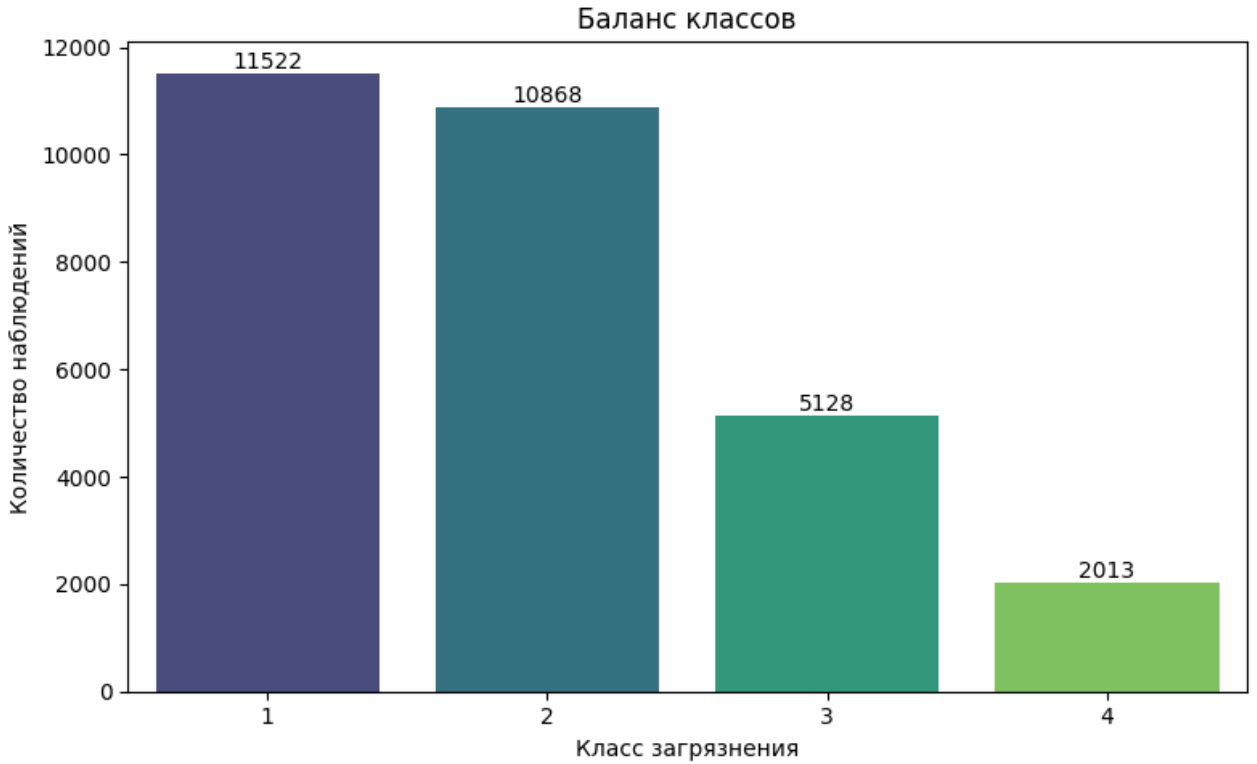


Предобработка и анализ данных

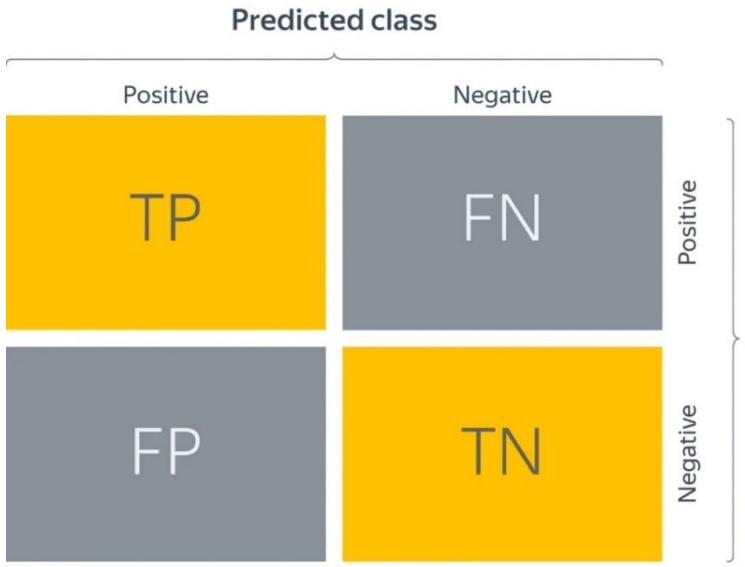
Масштабирование и синтетическое расширение данных

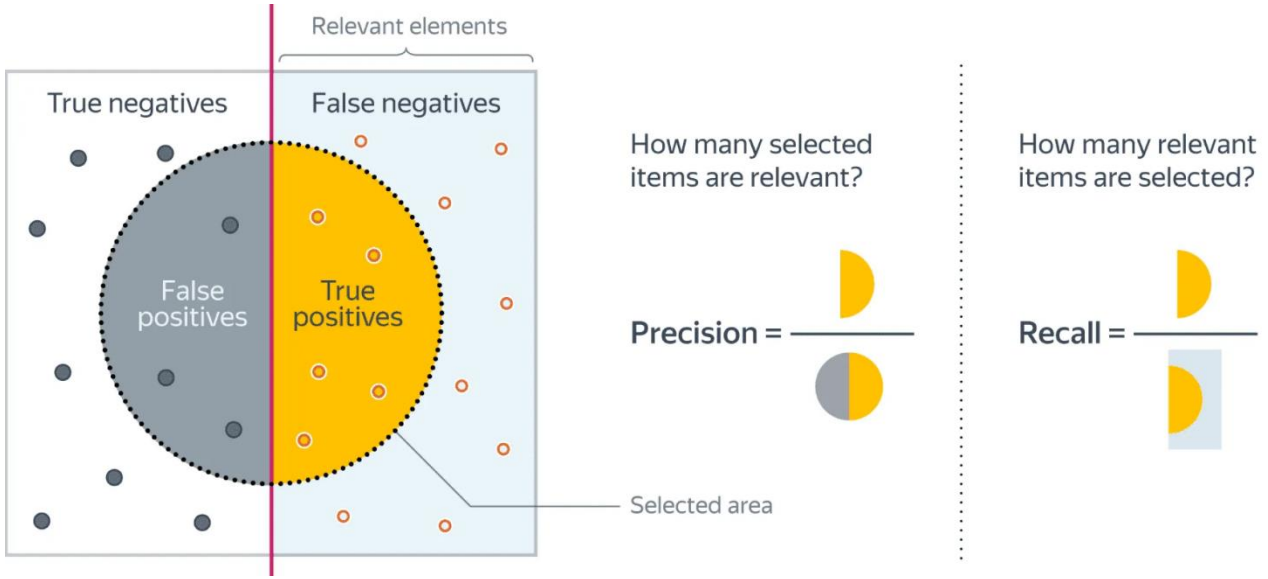
StandardScaler – преобразует значения числовых признаков к нулевому среднему и единичному стандартному отклонению для улучшения сходимости и повышения точности

SMOTE (Synthetic Minority Over-sampling Technique) – создаёт новые синтетические примеры для миноритарного класса, интерполируя значения между ближайшими соседями



Метрики оценки





$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

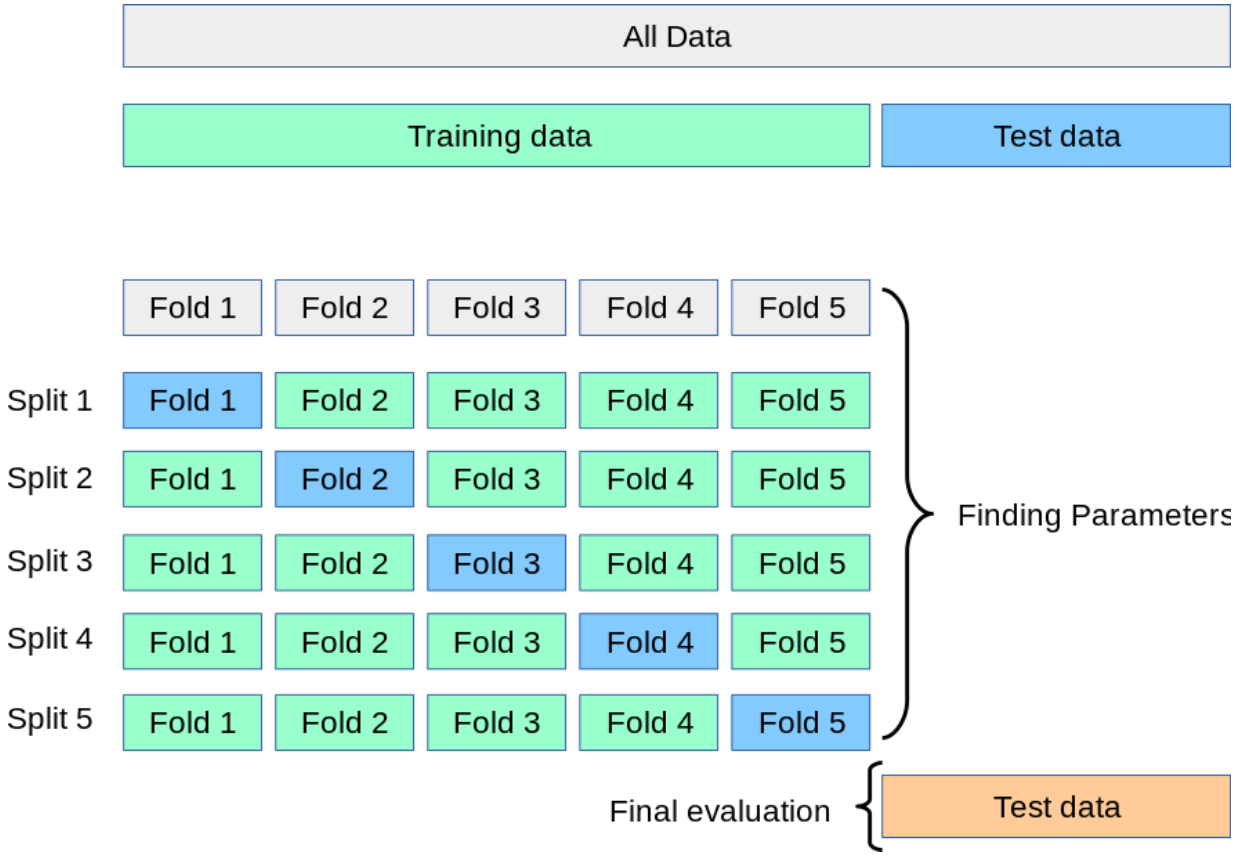
$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$$

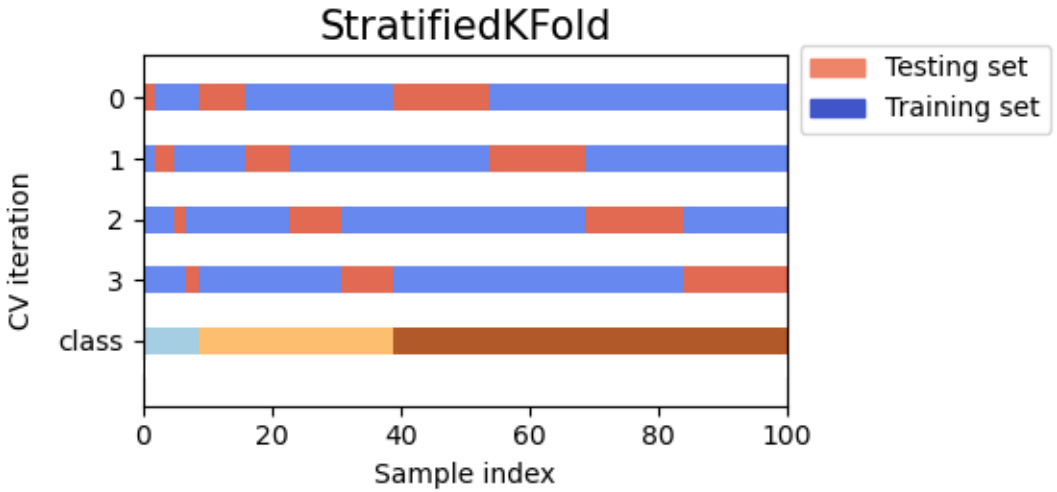
Классические модели ML

Многоклассовая логистическая регрессия	Метод опорных векторов (SVM)	Категориальный бустинг (CatBoost)
• Идея: Модель строит линейную комбинацию признаков и преобразует её в вероятности с помощью логистической функции (сигмоиды) • Accuracy: 77.11% • F1-macro: 75.79%	• Идея: Строится гиперплоскость, разделяющая данные с максимальным отступом между классами. В многоклассовом варианте используется подход "один против всех" • Accuracy: 85.17% • F1-macro: 84.96%	• Идея: Ансамблевый метод, использующий градиентный бустинг на деревьях решений для обучения, эффективен для табличных данных и автоматически обрабатывает категориальные признаки • Accuracy: 86.78% • F1-macro: 86.43%

Cross-validation



Stratified KFold Cross Validation



Модели нейросетей

Многослойный перцептрон (MLP)

- Описание: Несколько полносвязных слоев с вариациями в количестве слоев и нейронов
- Цель: Простая архитектура для выявления зависимостей между признаками
- Диапазон целевой метрики: 75–87%

Сверточная нейронная сеть (CNN)

- Описание: Сети с 2-4 сверточными слоями, Batch Normalization, MaxPooling и Dropout для регуляризации
- Цель: Извлечение более сложных паттернов и зависимостей в данных
- Диапазон целевой метрики: 75–89%

Вариации и параметры, применённые к обеим архитектурам

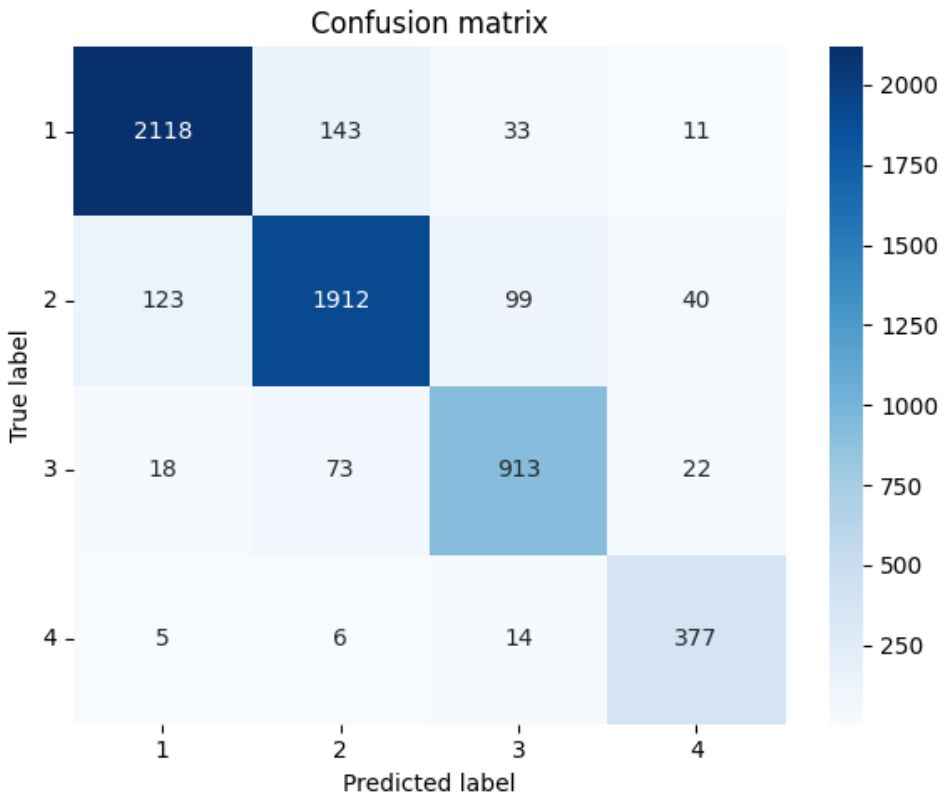
- SMOTE, ADASYN и SMOTETomek: Синтетическое увеличение данных для балансировки классов
- Создание новых признаков как линейную комбинацию уже существующих
- CatBoost: Обучение CatBoost для создания дополнительных признаков (предсказанные вероятности классов, листья деревьев)
- Изменения гиперпараметров: варьирование количества слоев, числа фильтров, коэффициентов Dropout, скорости обучения и других параметров для улучшения точности

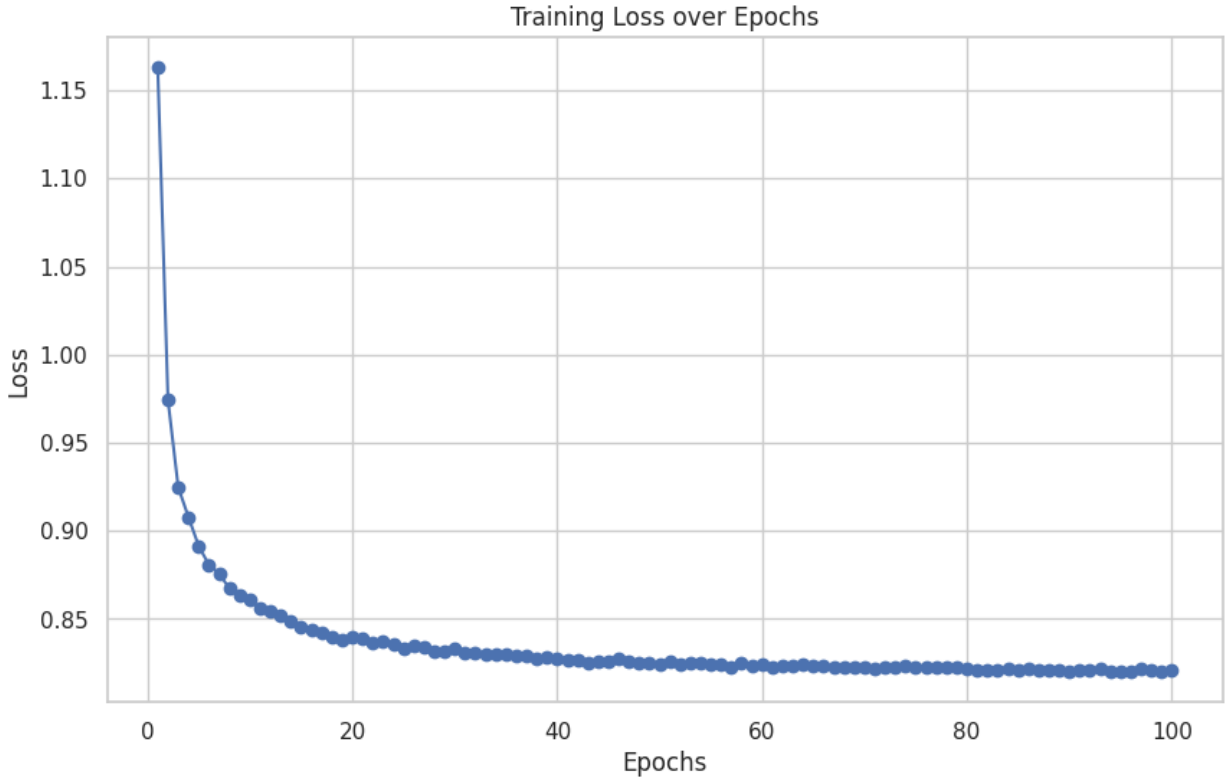
Структура лучшей модели

Feature-level Stacking

CatBoost	Сверточные слои	Полносвязные (Dense) слои	Выходной слой	Гиперпараметры (после оптимизации)
• 800 итераций • Скорость обучения 0.05 • Глубина 8 • L2-регуляризация 7 • Предсказанные вероятности классов добавляются как новые признаки для обучения нейронной сети	• Четыре слоя с ReLU, Batch Normalization и Dropout (0.1) • Первый слой: 64 фильтра, ядро 3 • Второй слой: 128 фильтров, ядро 3 • Третий слой: 256 фильтров, ядро 3 • Четвертый слой: 512 фильтров, ядро 3	• Два слоя с ReLU и Dropout (0.1) для регуляризации • Первый слой: 128 нейронов • Второй слой: 64 нейрона	• Softmax для многоклассовой классификации	• Функция потерь: CrossEntropyLoss • Оптимизатор: Adam • Скорость обучения 0.0003 • L2-регуляризацией 1e-3 • 100 эпох

Результаты лучшей модели

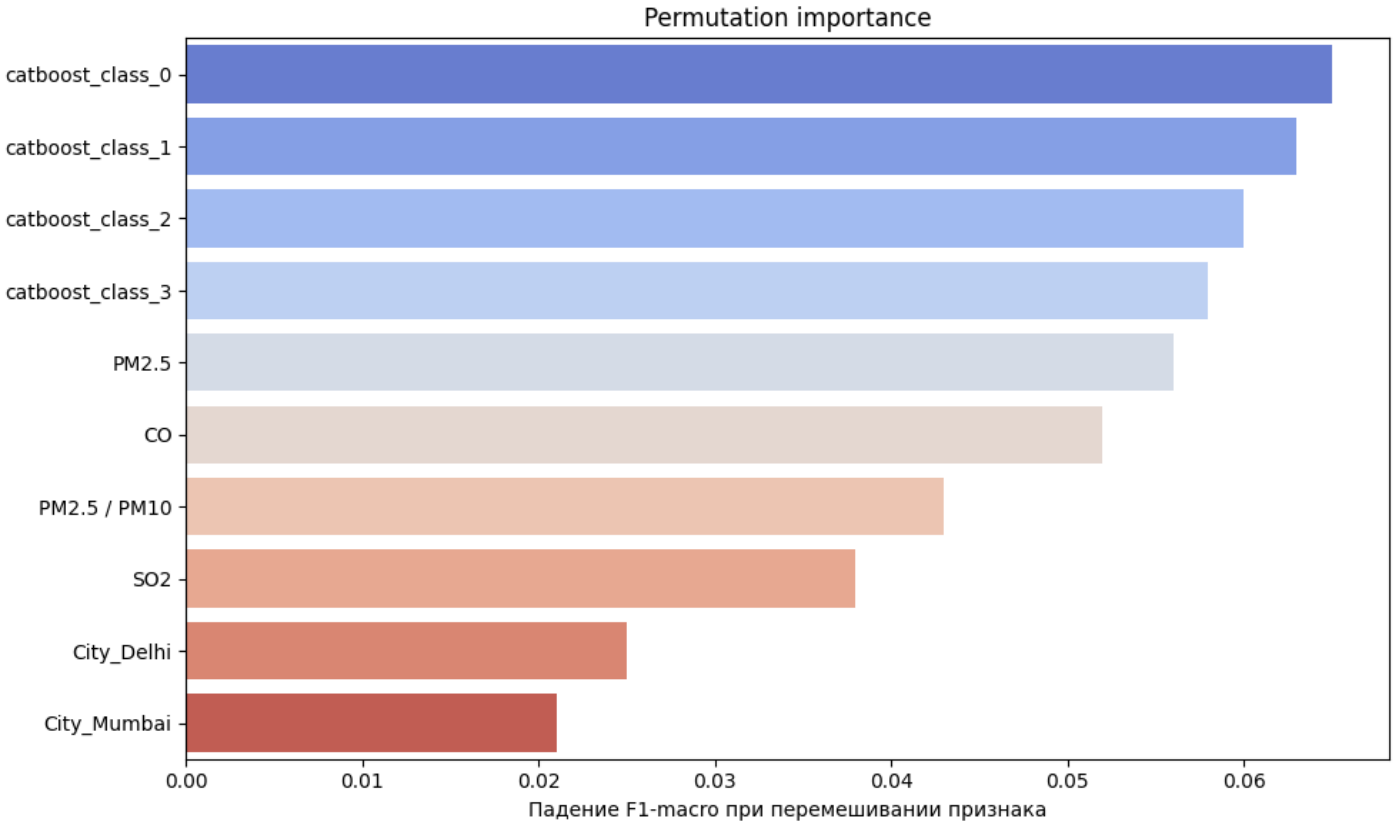




Accuracy – 90.06%

F1-macro – 89.39%

Анализ важности признаков



Без признаков CatBoost:
F1-macro = **85%**

С признаками CatBoost:
F1-macro = **89%**

Заключение

Разработана универсальная модель классификации уровней загрязнения воздуха на основе сверточной нейронной сети

В модель добавлены вероятности принадлежности к классам, предсказанные CatBoost – Feature-level Stacking

CatBoost-признаки усиливают модель, не заменяя, а дополняя исходные данные

Гибридный подход показал наилучший результат: F1-macro = 89%

Участие в конференции «Нефть и газ – 2025»

- Доклад в секции «Автоматизация, моделирование и искусственный интеллект» в рамках 79-й Международной молодежной научной конференции «Нефть и газ – 2025»

Будущие исследования

Проведение экспериментов с другими архитектурами (LSTM, TabNet, Transformer)

Исследование влияния дополнительных признаков:
влажность, скорость и направление ветра, температура, осадки, география

Применение модели на российских данных

Благодарю за внимание