

# Прогнозирование медицинских кодов по симптомам

Решение задачи RuMedTop3: автоматическое определение ТОП-3 наиболее вероятных медицинских кодов на основе текстового описания симптомов пациентов

## Состав группы

- Ким Дмитрий
- Костин Виталий
- Гусев Егор
- Михайлов Артём





# Цель и метрики задачи

## Задача

По текстовому описанию симптомов на русском языке предсказать три наиболее вероятных медицинских кода из набора 105 возможных диагностических категорий.

## Метрики оценки

**Hit@3** – основная метрика, показывает процент случаев, когда правильный код находится в тройке предсказанных

**Accuracy@1** – дополнительная метрика точности первого предсказания

# Характеристика датасета RuMedTop3

## Источник данных

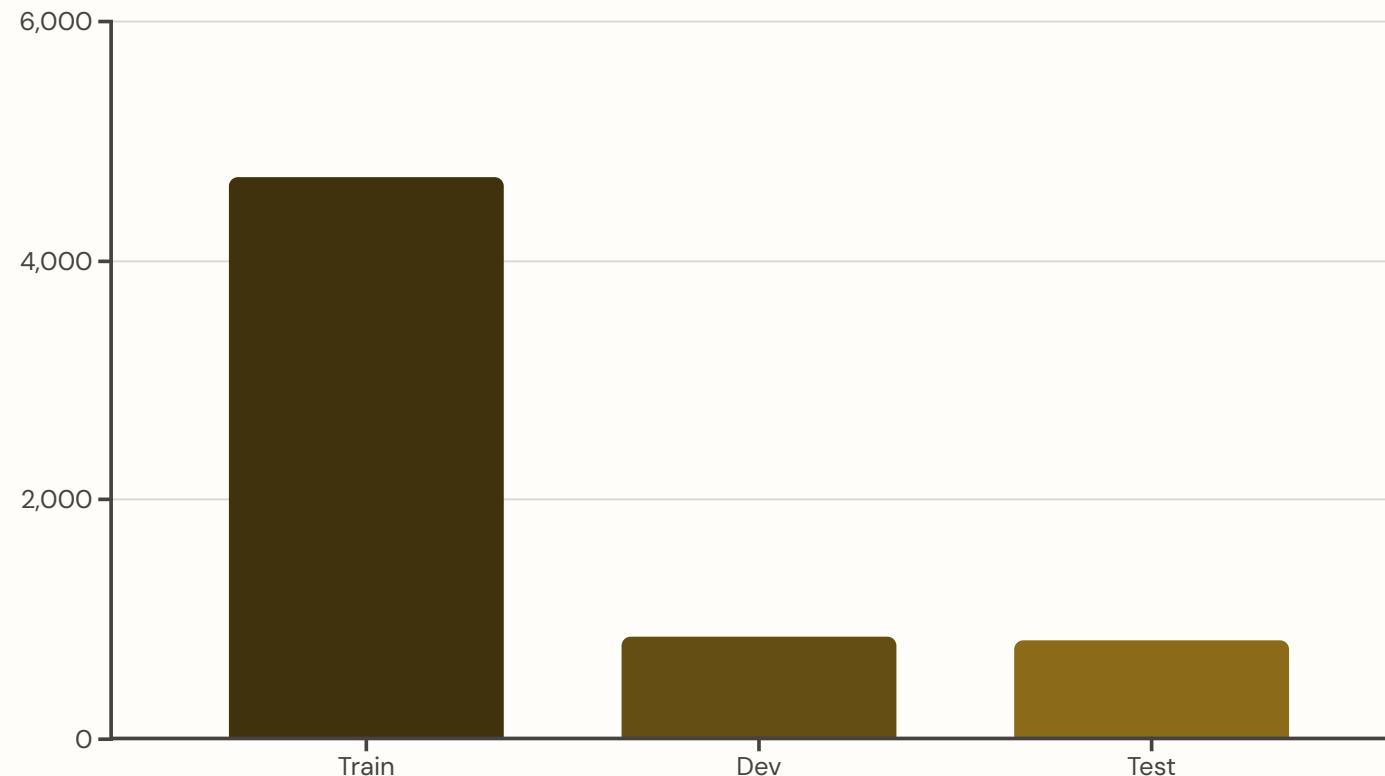
Датасет RuMedTop3 для многоклассовой классификации с ранжированием результатов

## Классы

105 уникальных медицинских кодов для классификации заболеваний

## Структура записи

idx (идентификатор), symptoms (описание), code (целевой класс)





# Базовое решение: подход и архитектура

01

---

## TF-IDF векторизация

Символьные n-граммы (3-8), извлечение признаков из текстового описания симптомов

02

---

## Логистическая регрессия

One-vs-Rest стратегия, L2-регуляризация с параметром  $C = 10$

03

---

## Получение вероятностей

Расчёт `predict_proba` для всех 105 классов

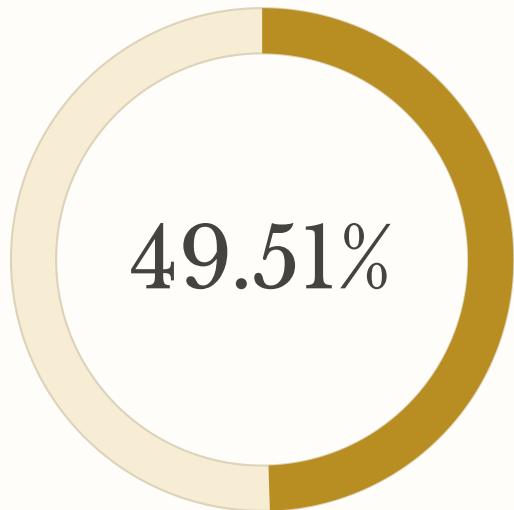
04

---

## Ранжирование

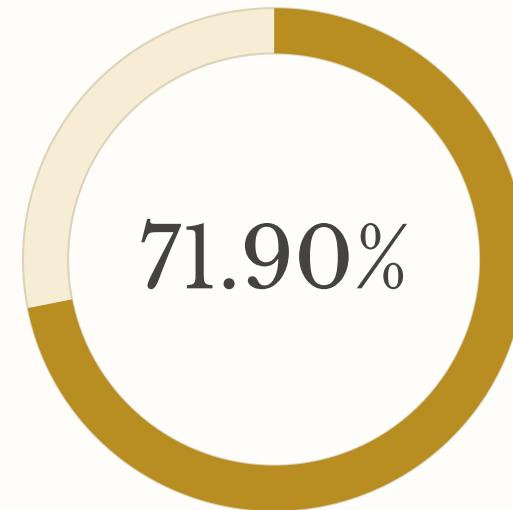
Сортировка классов по убыванию вероятности для формирования ТОП-3

# Результаты базового решения



Accuracy@1

Точность предсказания первого кода на валидационной выборке



Hit@3

Процент попадания правильного кода в тройку предсказанных

- ❑ **Ограничения базового решения:** использование только одного типа признаков (символьные n-граммы), единственный классификатор с ограниченной выразительностью, отсутствие ансамблирования и прямое использование вероятностей без учёта рангов

# Улучшенное решение: расширение признаков

## Двухуровневая векторизация

Для повышения выразительности модели используются два комплементарных типа признаков:

- **Символьные n-граммы** (3-8): захват морфологических паттернов, опечаток и суффиксов – max\_features = 50 000
- **Словарные n-граммы** (1-2): захват семантических связей и медицинских терминов – max\_features = 30 000

Итоговая размерность признакового пространства: **4690 × 80 000**

2

Типа признаков

80K

Признаков



# Ансамбль моделей и калибровка

1

## Logistic Regression

Базовая линейная модель с высокой интерпретируемостью и стабильными результатами

2

## LinearSVC + Calibration

Margin-based классификация с калибровкой через CalibratedClassifierCV для получения вероятностей

3

## Multinomial Naive Bayes

Быстрая вероятностная модель, эффективная на текстовых данных

4

## Random Forest

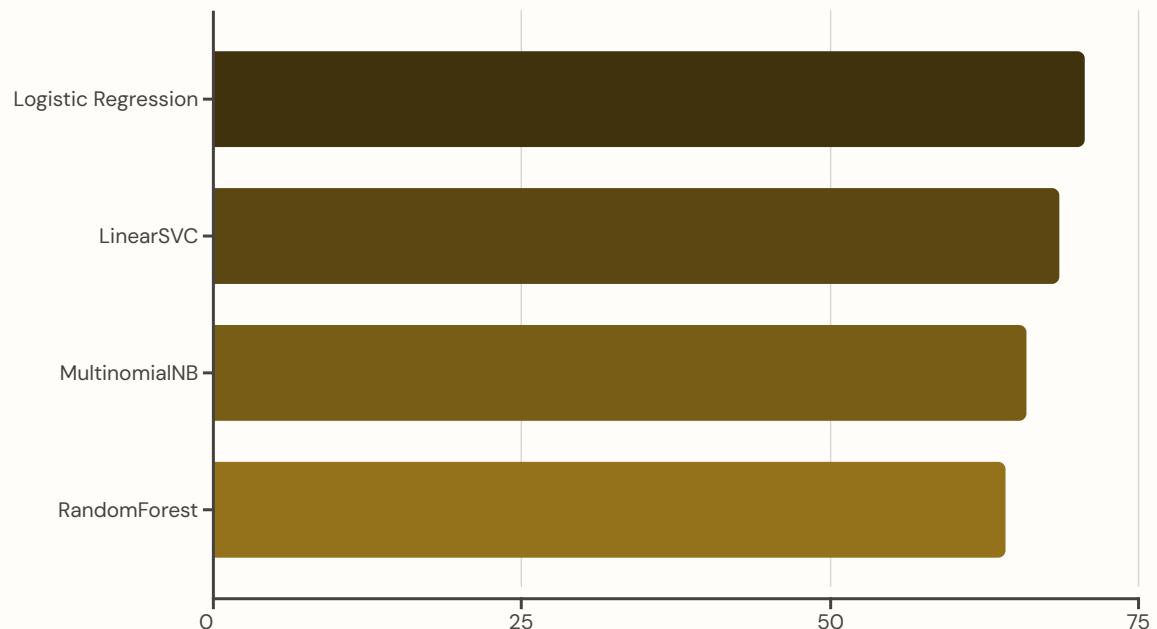
Нелинейная ансамблевая модель для захвата сложных взаимодействий признаков

# Rank-based ансамблирование

## Принцип работы

Вместо простого усреднения вероятностей используется **ранговое голосование**:

1. Каждый классификатор сортирует классы по уверенности
2. Начисляются баллы по формуле:  $score += weight / (rank + 1)$
3. Веса пропорциональны Hit@3 на Dev
4. Random Forest штрафуется ( $\times 0.7$ ) для балансировки



# Сравнение решений

## Baseline

**Accuracy@1:** 49.51%

**Hit@3:** 71.90%

- Один тип признаков
- Единственный классификатор
- Простое ранжирование

## Улучшенное решение

**Accuracy@1:** 50.00%

**Hit@3:** 72.63%

- Два типа признаков (80К)
- Ансамбль из 4 моделей
- Ранговое голосование

□ **Вывод:** Ансамбль показывает стабильные результаты, сопоставимые с базовым решением. Logistic Regression остаётся сильным одиночным классификатором. Хотя прирост качества минимален (+0.49% по Accuracy@1, +0.73% по Hit@3), ансамбль обеспечивает большую устойчивость за счёт комбинирования разных моделей и подходов к векторизации.



# Итоги и перспективы развития

## Достигнутые результаты

- Повышена устойчивость модели через ансамблирование
- Протестированы различные ML-подходы
- Реализован кастомный rank-based ансамбль под метрику Hit@3
- Расширено признаковое пространство до 80К признаков

## Направления улучшения

- **Fine-tuning RuBERT/MedBERT** для учёта контекста
- **Learning-to-Rank (LambdaMART)** для оптимизации ранжирования
- **Error analysis** по редким и сложным классам
- **Data augmentation** симптомов для балансировки датасета

# Бот для прогнозирования медицинских кодов

## Входные данные

Бот принимает текст на русском языке с описанием симптомов пациента

## Выходные данные

- ТОП-3 наиболее вероятных медицинских кодов (МКБ-10)
- Расшифровку кодов в виде названий заболеваний

The screenshot shows a series of messages from a bot named 'RuMedTop3' on a dark-themed interface. The messages are as follows:

- Message 1:** 'Результат носит справочный характер и не является медицинским заключением.' (Result is for reference only and does not constitute a medical conclusion.)  
Time: 23:29  
Text: 'высокая температура (выше 38°C), озноб, сильная слабость, ломота в мышцах и суставах, головная боль (особенно в области лба и глаз). Позже присоединяются сухой кашель, першение в горле, насморк, а также может быть потливость, боли за грудиной, светобоязнь, покраснение лица и глаз.'  
Time: 23:29
- Message 2:** 'ТОП-3 возможных диагноза:'  
List:
  1. J06 — Острые инфекции верхних дыхательных путей множественной и неуточненной локализации
  2. J00 — Острый назофарингит [насморк]
  3. M54 — Дорсалгия

Результат носит справочный характер и не является медицинским заключением.

Time: 23:29
- Message 3:** 'на чувство комка в горле и периодическая боль в горле'  
Time: 23:30  
Text: 'нашение температуры до 38.5, насморк, заложенность носа, головные боли и боли в горле, слабость.'

Результат носит справочный характер и не является медицинским заключением.

Time: 23:30
- Message 4:** 'ТОП-3 возможных диагноза:'  
List:
  1. J31 — Хронический ринит, назофарингит и фарингит
  2. E06 — Тиреоидит
  3. E04 — Другие формы нетоксического зоба

Результат носит справочный характер и не является медицинским заключением.

Time: 23:30
- Message 5:** 'ТОП-3 возможных диагноза:'  
List:
  1. J06 — Острые инфекции верхних дыхательных путей множественной и неуточненной локализации
  2. J00 — Острый назофарингит [насморк]
  3. J30 — Вазомоторный и аллергический ринит

Результат носит справочный характер и не является медицинским заключением.

Time: 23:30