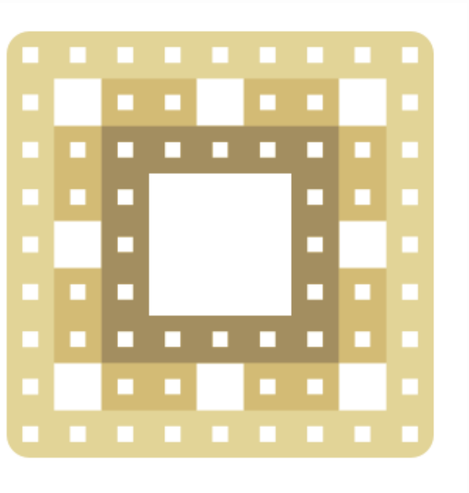


АНАЛИЗ ДАННЫХ В МЕДИЦИНСКИХ ЗАДАЧАХ

Выполнил: Гершов М.Д.
Руководитель: Кушнеров А.В.

ВВЕДЕНИЕ В ПРОБЛЕМУ

Машинное обучение в медицинской сфере будет призвано решить сразу несколько основных проблем:

- Недостаток квалифицированного медицинского персонала
- Стоимость и доступность качественной диагностики
- Раннее выявление особо опасных заболеваний для повышения шансов на выздоровление

ЦЕЛЬ

Качественный **анализ** медицинских **данных** для **обучения моделей МО** и применение полученных результатов для прогнозирования наличия ССЗ.

ЗАДАЧИ

1. Подбор, анализ и предварительная обработка данных для обучения модели.
2. Изучение классических методов МО для решения задачи классификации.
3. Обучение нескольких видов моделей, оптимизация и оценка качества.
4. Сравнительный анализ обученных моделей.

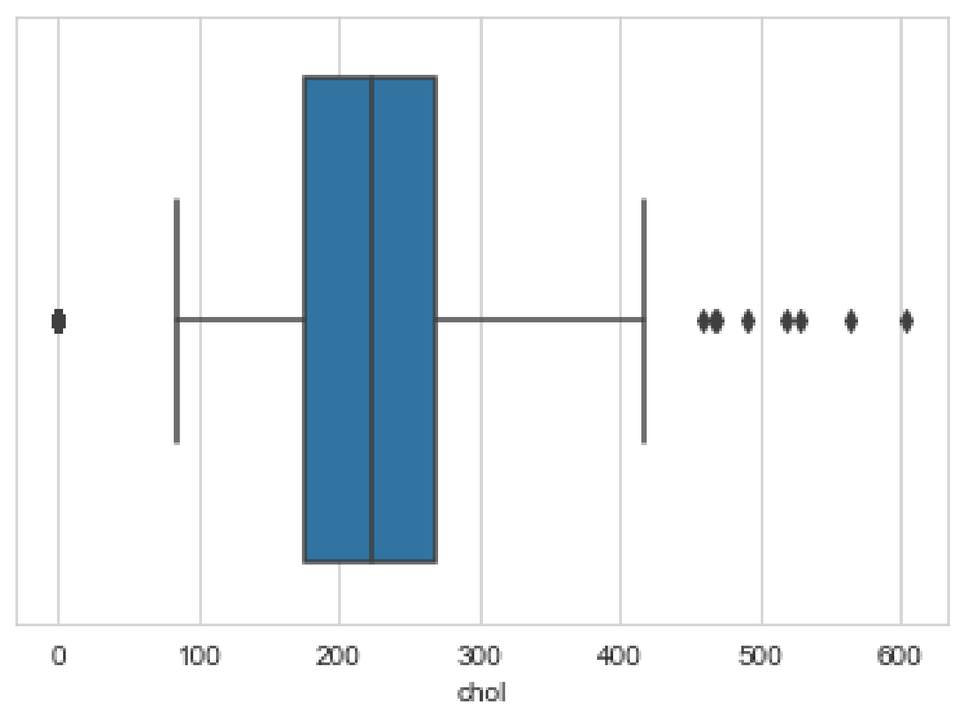
НАБОР ДАННЫХ

	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	63	Male	Cleveland	typical angina	145.0	233.0	1	lv hypertrophy	150.0	0.0	2.3	downsloping	0.0	fixed defect	0
1	67	Male	Cleveland	asymptomatic	160.0	286.0	0	lv hypertrophy	108.0	1.0	1.5	flat	3.0	normal	2
2	67	Male	Cleveland	asymptomatic	120.0	229.0	0	lv hypertrophy	129.0	1.0	2.6	flat	2.0	reversable defect	1
3	37	Male	Cleveland	non-anginal	130.0	250.0	0	normal	187.0	0.0	3.5	downsloping	0.0	normal	0
4	41	Female	Cleveland	atypical angina	130.0	204.0	0	lv hypertrophy	172.0	0.0	1.4	upsloping	0.0	normal	0

Краткое описание основных признаков:

- **cp** – тип боли в груди
- **trestbps** – кровяное давление в состоянии покоя в мм. рт. ст. при поступлении в больницу
- **fbs** – уровень сахара в крови натощак > 120 мг/дл
- **restecg** – результаты электрокардиографии в состоянии покоя
- **thalch** – достигнутая максимальная частота сердечных сокращений
- **slope** – наклон пикового сегмента ST
- **num** – целевая переменная

ОБРАБОТКА И АНАЛИЗ ДАННЫХ



Основные **этапы** обработки данных:

- Приведение булевских значений к числовому типу
- Обработка выбросов
- Работа с пропущенными значениями
- Кодирование категориальных признаков
- Приведение целевой переменной к бинарному типу

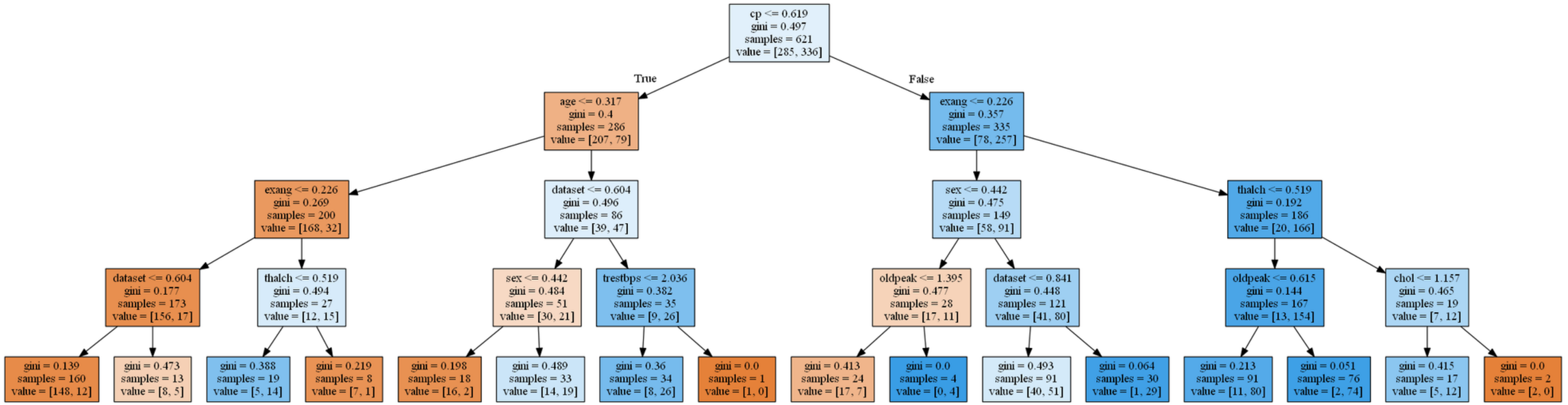
После проведения обработки набора данных количество признаков сократилось до 12, а число записей – до 888.

МОДЕЛИ МО: РЕШАЮЩЕЕ ДЕРЕВО

7 Моделей классификатора

80% Recall-точность дерева для *7 главных компонент*

0,84 Площадь под *ROC-кривой* лучшей модели



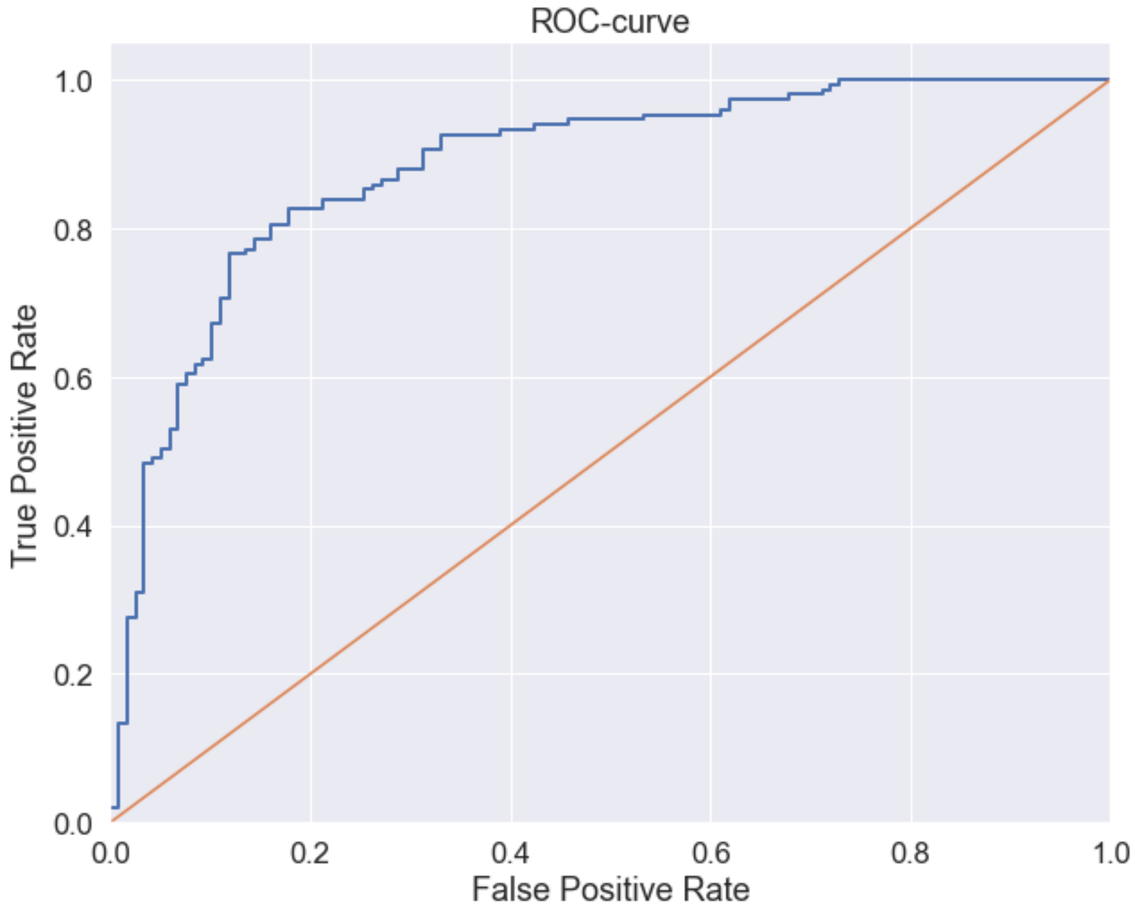
МОДЕЛИ МО: СЛУЧАЙНЫЙ ЛЕС

7 Моделей классификатора

0,88 Площадь под *ROC-кривой* лучшей модели

Важные преимущества использования моделей, основанных на **бэггинге**:

- уменьшение дисперсии прогноза модели
- улучшение производительности и надёжности
- обработка сложных и нелинейных взаимосвязей в данных



МОДЕЛИ МО: ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

	precision	recall	f1-score	support
Здоровый	0.84	0.79	0.81	125
Больной	0.83	0.87	0.85	142
accuracy			0.83	267
macro avg	0.83	0.83	0.83	267
weighted avg	0.83	0.83	0.83	267

6 Моделей классификатора

0,89 Площадь под *ROC-кривой* лучшей модели

РЕЗУЛЬТАТЫ

Что было сделано?

- Полноценное погружение в область анализа данных и МО
- Изучение основных методов предварительной обработки данных
- Исследование базовых способов выявления взаимосвязей в данных
- Изучение нескольких техник кодирования, метода главных компонент
- Обучение наиболее популярных моделей МО: дерево решений, случайный лес, логистическая регрессия
- Поиск и подбор гиперпараметров
- Ознакомление с метриками оценки качества моделей

Что было получено в цифрах?

20	87%	0,89
Обученных моделей	Recall-точность лучшей модели	Площадь под *ROC-кривой* лучшей модели