# Heart Disease Classification

Doronin I.V  Aristarkhov Z.G

National Research Nuclear University MEPhI

## Информация о датасете

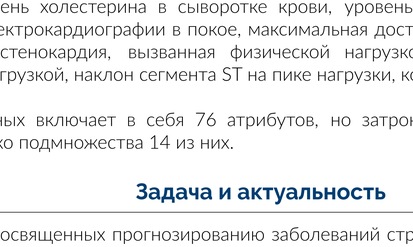
Датасет взят из официального UCI репозитория. Это многомерный набор данных, состоящий из 14 атрибутов: возраст, пол, тип боли в груди, артериальное давление в покое, уровень холестерина в сыворотке крови, уровень сахара в крови натощак, результаты электрокардиографии в покое, максимальная достигнутая частота сердечных сокращений, стенокардия, вызванная физической нагрузкой, депрессия, вызванная физической нагрузкой, наклон сегмента ST на пике нагрузки, количество крупных сосудов и талассемии.
Вся база данных включает в себя 76 атрибутов, но затронутые здесь исследования касаются только подмножества 14 из них.

## Задача и актуальность

Число работ посвященных прогнозированию заболеваний стремительно растет по мере появления статистики, позволяющей произвести анализ. Особенно это актуально для определения болезней сердца, которые занимают лидирующие места среди причин смертности в России и во всем мире.
Настоящая работа представляет обзор моделей машинного обучения для прогнозирования наличия болезней сердца у пациентов. Также одной из основных целей работы является выделение **наиболее важных признаков**, которые являются определяющими при классификации.

## Предобработка и визуализация данных

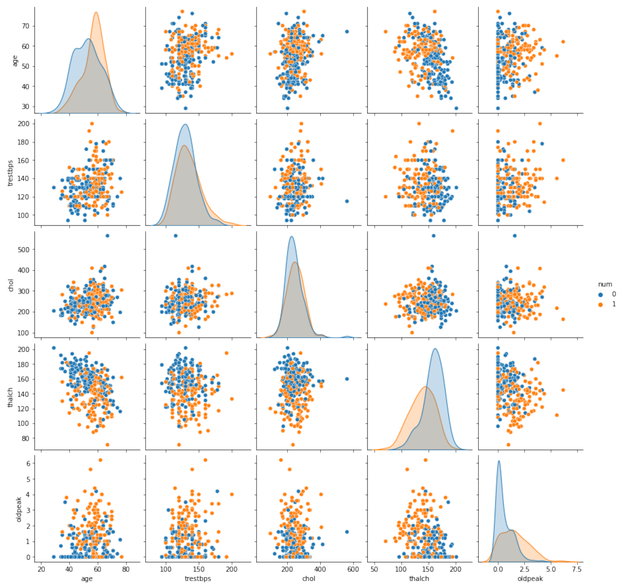
- **Предобработка.** Данные были проверены на невалидные вещественные значения(NaN) и на наличие пропусков в признаках.
- **Выделение категориальных признаков.** Признаки были разбиты на категориальные и вещественные.
- **Визуализация данных.** Построены круговые диаграммы для категориальных признаков и визуализирована попарная кореляция признаков.
- **Бинаризация целевой переменной.** Так как основной задачей являлась классификация - наличие болезни, мы опускаем степень заболевания и определяем целевую переменную как: 1 - человек болен, 0 - нет.
- **Стандартизация данных.** Для вещественных признаков была применена нормировка при помощи StandartScaler, для категориальных - OneHotEncoding.




0.5

Figure 1. Heatmap

С помощью heatmap, мы можем заметить, что основной вклад в целевую переменную дают признаки: **ca**, **age**, **oldpeak**.




0.5

Figure 2. Попарная корелляция данных

При помощи pairplot мы можем сравнивать распределения пар числовых переменных. Из этих парных диаграмм, мы можем заметить, что для классов '1' и '0' распределения одинаковых признаков имеют явные закономерности. Данные разделены вертикально/горизонтально, т.е при некотором значении на оси x или y вероятность наличия болезней сердца повышается. Это означает, что можно перейти к выбору метрик и моделей.

## Модели

Каждая из моделей использовалась в совокупности с процедурой кросс-валидации. Помимо стандартных моделей, таких как K-Nearest Neighbors, Logistic Regression, Decision Tree, использовались и композиции алгоритмов: Gradient Boosting, Random Forest, Stacking. Помимо этого использовался метод опорных векторов SVM, который при использовании различных ядер способен быть как бинарным, так и многоклассовым классификатором, благодаря переходу в пространство более высокой размерности. **Классифицирующая функция** строится в виде:

$$F(x) = sign(< w, x > + b)$$

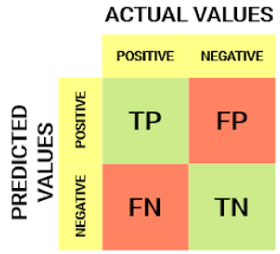
В случае линейной неразделимости задается такое отображение в пространство более высокой размерности, причем $\phi$ такое, что в в новом пространстве $X$ выборка становится линейно разделимой: $\phi : \mathbb{R}^n \to X$.
Классифицирующая функция примет вид $F(x) = sign(< w, \phi(x) > + b)$ , а выражение $k(x, x') =< \phi(x), \phi(x') >$ называется **ядром** классификатора. С математической точки зрения ядром может служить любая положительно определенная симметричная функция двух переменных.
В процессе обучения моделей были добавлены новые признаки такие как: произведение давления на холестерин и частота сердечных сокращений умноженная на депрессию, вызванную нагрузкой. После добавления этих признаков, площадь под ROC кривой стала 0.975.

## Метрики качества классификации

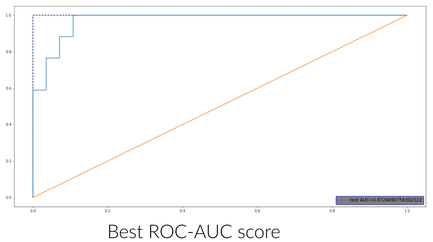
В задачах классификации очень важным шагом является выбор метрики качества. Это связано с рядом ньюансов, например, в задачах медицинской диагностики очень часто наблюдается несбалансированность выборки, по итогу чего цены ошибок неравнозначны. Чтобы различать ошибки разных типов рационально ввести матрицу ошибок.
Одним из способов оценить модель в целом, не привязываясь к конкретному порогу, является **AUC-ROC** (или ROC AUC) — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve ). Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR):

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

$$FPR = \frac{FP}{FP + TN}$$




Error Matrix



Best ROC-AUC score

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.89 | 0.91 | 28 |
| 1 | 0.83 | 0.88 | 0.86 | 17 |
| accuracy | - | - | 0.89 | 45 |
| macro avg | 0.88 | 0.89 | 0.88 | 45 |
| weighted avg | 0.89 | 0.89 | 0.89 | 45 |

Table 1. Precision-recall.

## Результаты

Таким образом, были выявлены лучшие модели для прогнозирования заболеваний сердца на UCI наборе данных. Ими являются Logistic Regression И SVM с применением сигмоидальной функции ядра, показавшие результаты 0.91 и 0.89, что является наивысшим показателем среди всех ноутбуков, представленных в этом соревновании на kaggle.
Отобраны самые важные признаки для диагностики заболеваний сердца: количество крупных сосудов, окрашенных при рентгеноскопии, возраст, депрессия ST, вызванная физической нагрузкой по сравнению с отдыхом.

## Ссылки

[1] Федотов Станислав, Синицин Кирилл, ШАД ML-handbook

[2] `https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data` - Dataset

[3] Google colab with the notebook