Уральский федеральный университет
имени первого Президента России Б.Н.Ельцина
Физико-технологический институт

Уральский федеральный университет

Использование методов машинного обучения для идентификации заболеваний печени

Презентация выпускной квалификационной работы

Студент: Шишкина Е.А.

Группа: Фт-480012

Руководитель: Смирнов А.А.

АКТУАЛЬНОСТЬ

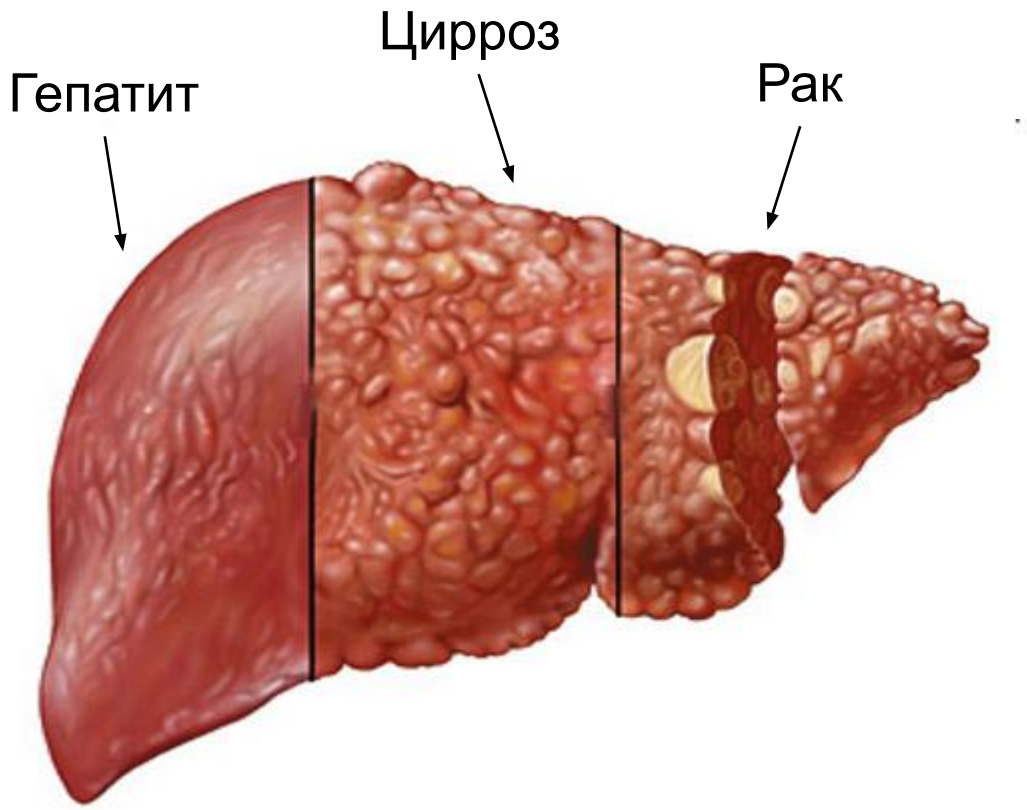
- На заболевания печени приходится 3,5 % всех смертей во всем мире
- Большая нагрузка на врачей
- Большая длительность ручной обработки подобного объема данных

Цель: использование методов машинного обучения для бинарной классификации заболеваний печени.

Задачи:

- проанализировать наиболее распространенные заболевания печени;
- рассмотреть существующие методы классификации в машинном обучении;
- реализовать выбранные методы классификации на Python.

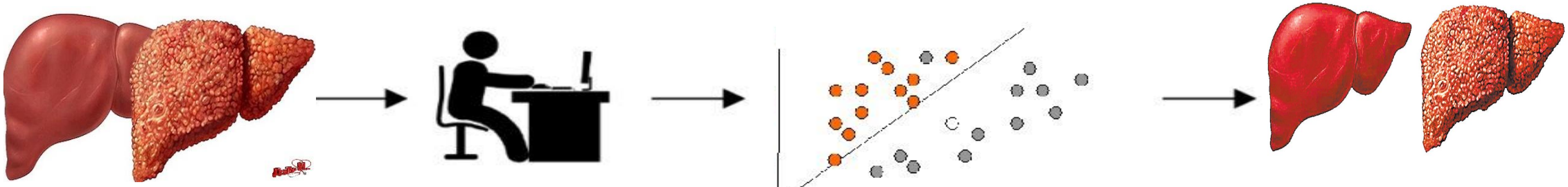
ФУНКЦИИ И ПАТАЛОГИИ ПЕЧЕНИ



- Регулирует объем крови
- Образование веществ для свертывания крови
- Синтез витаминов
- Поддержание уровня сахара
- Обмен железа
- Обезвреживание токсинов

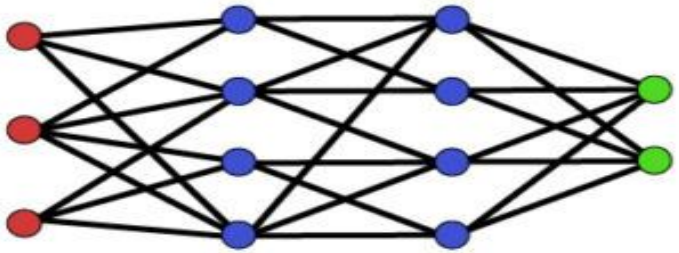
МЕТОДЫ

Машинное обучение



Входные данные → Человек → Классификатор → Выходные данные

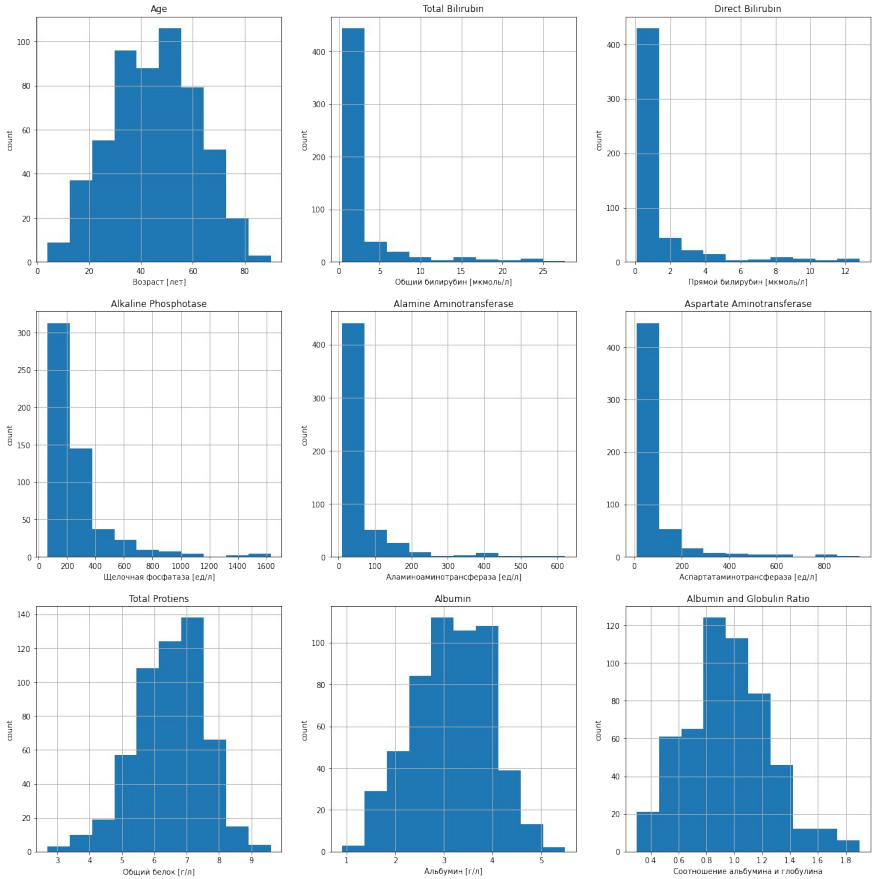
Глубокое обучение

Входные данные →  → Выходные данные

Нейронная сеть

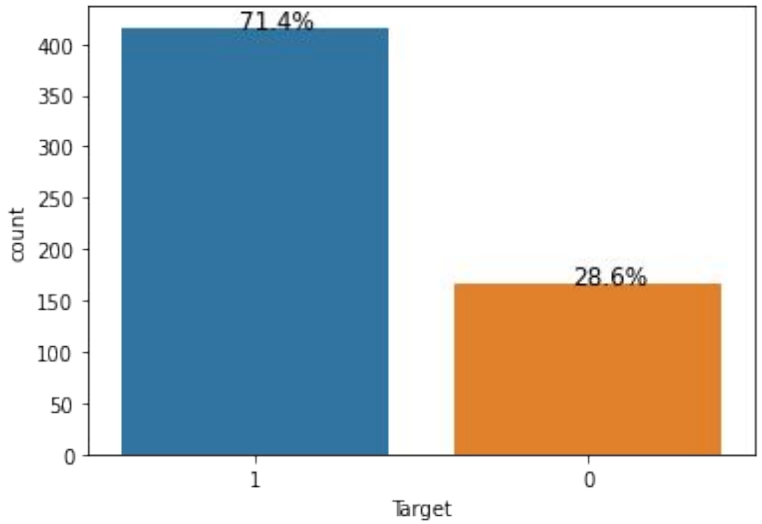
ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИДЕНТИФИКАЦИИ ЗАБОЛЕВАНИЙ ПЕЧЕНИ

НАБОР ДАННЫХ



Распределение числовых признаков

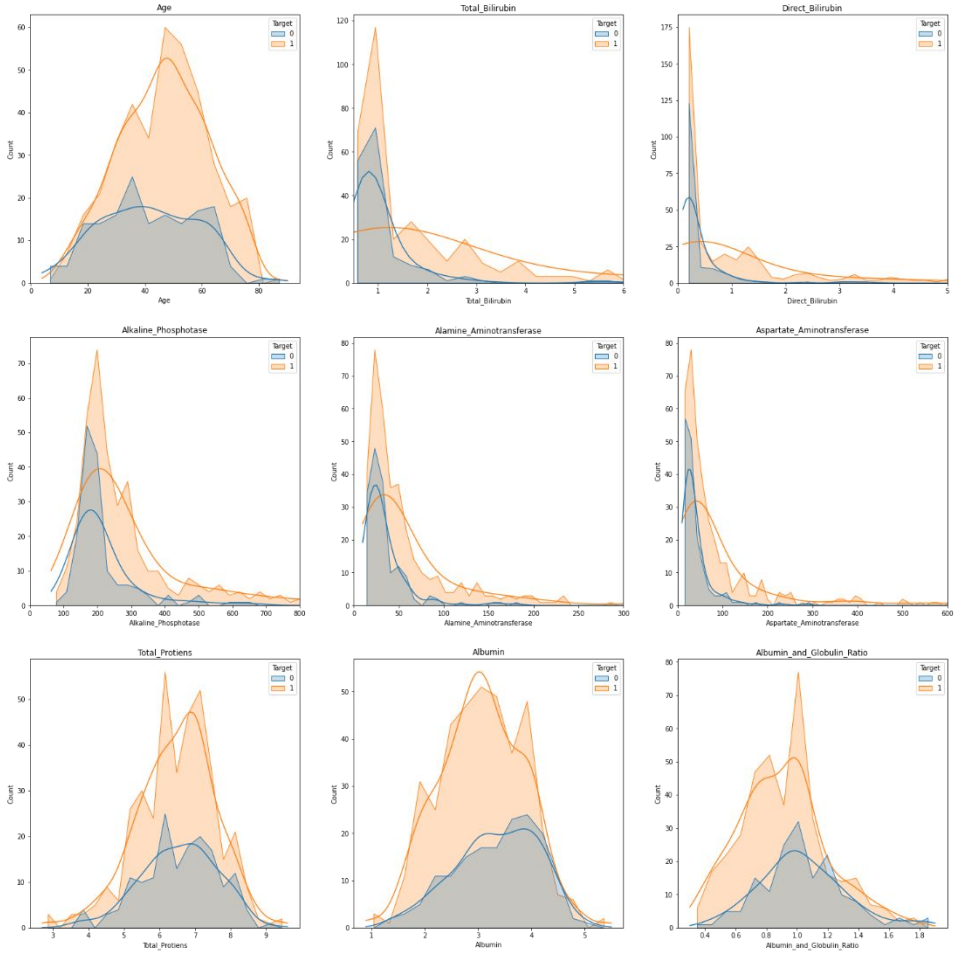
- 583 записи о пациентах
- 416 записей о пациентах с заболеваниями
- 167 записей о пациентах без заболеваний печени
- 10 параметров



Распределение целевой переменной

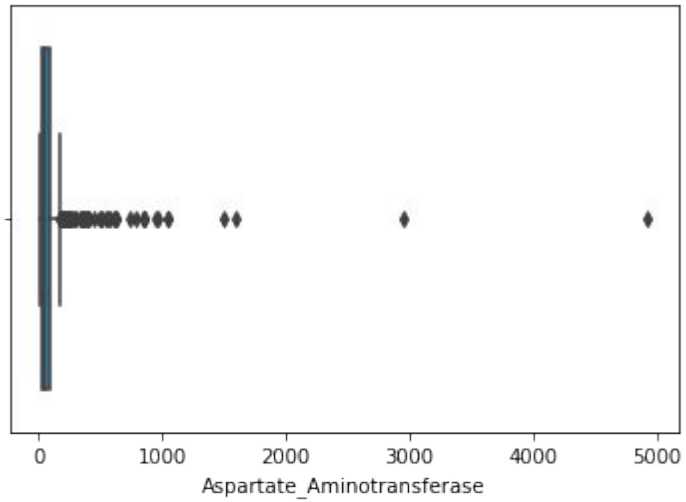
http://archive.ics.uci.edu/IndianLiverPatientDataset(ILPD)

ПРЕДОБРАБОТКА



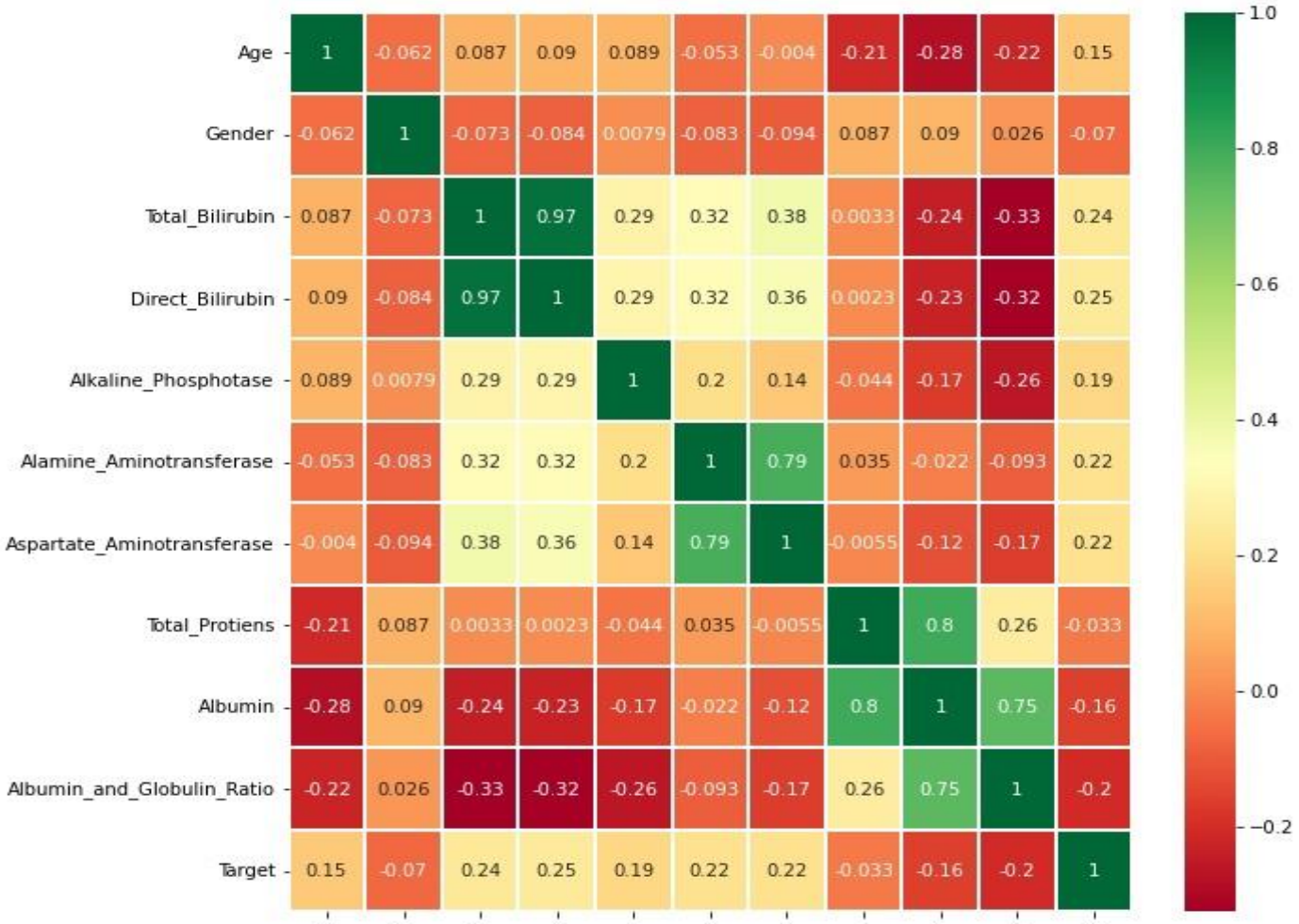
Распределение целевой переменной в каждом параметре

- Удаление пустых ячеек
- Заполнение пропусков
- Удаление выбросов
- Удаление дубликатов



Гистограмма выбросов Аспартатаминотрансферазы

КОРРЕЛЯЦИЯ ПРИЗНАКОВ

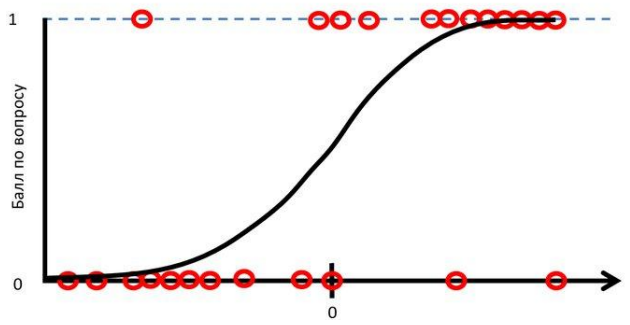


Корреляционная матрица

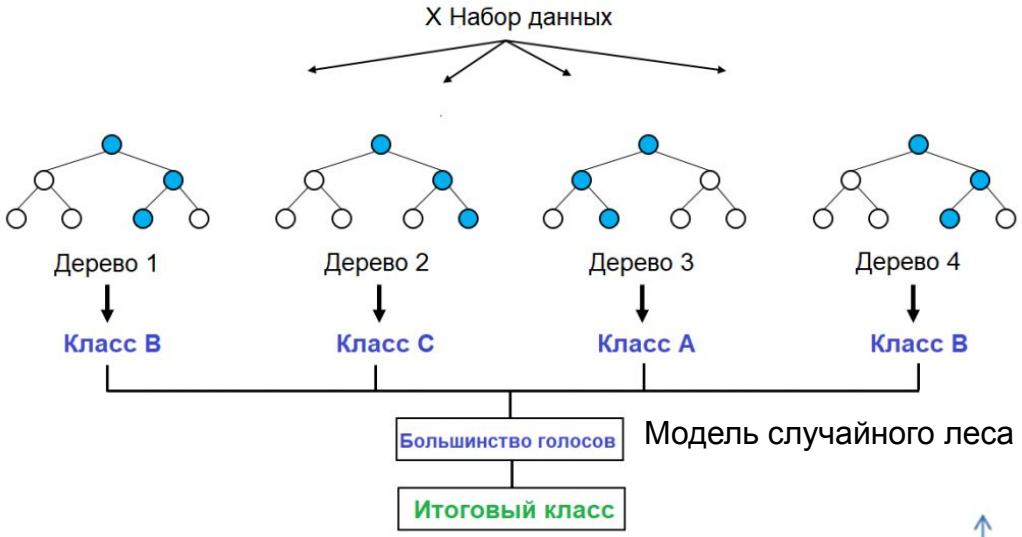
Корреляция показывает, насколько близко значения для двух отдельных функций изменяются одновременно.

- около -1 или 1 – сильная связь
- ближе к 0 – слабая.

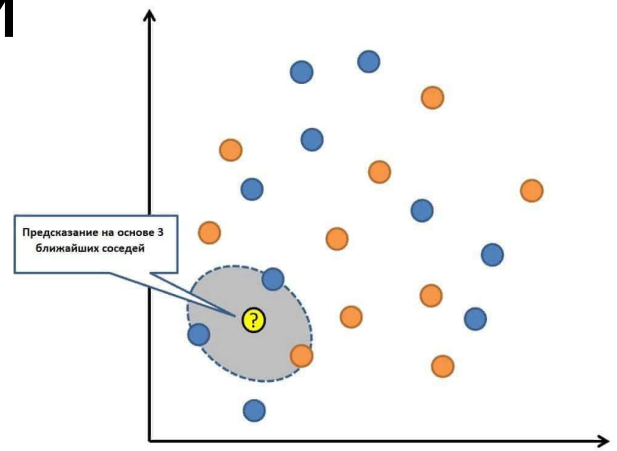
МЕТОДЫ КЛАССИФИКАЦИИ



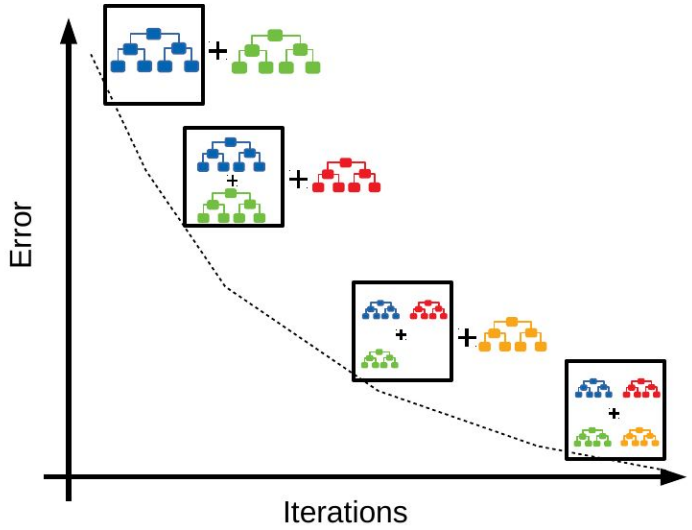
Модель логистической регрессии



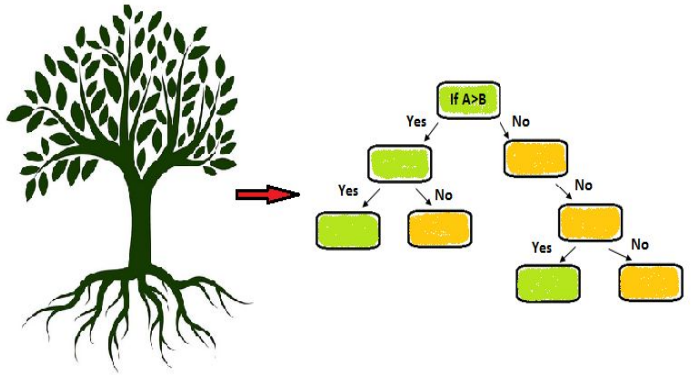
Модель случайного леса



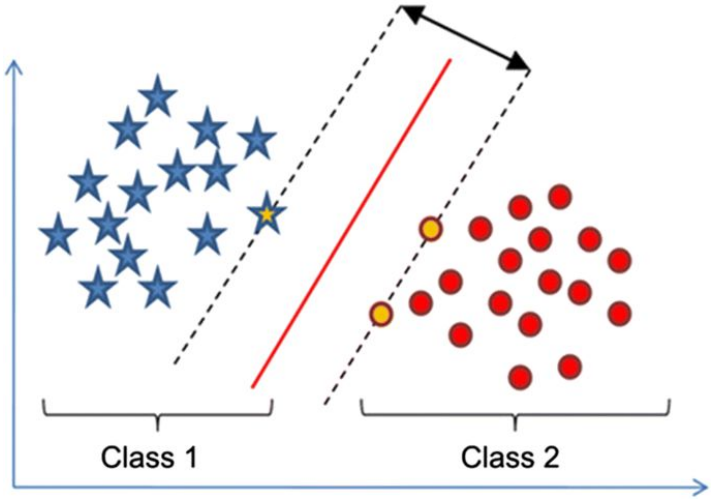
Метод k-ближайших соседей



Модель градиентного бустинга

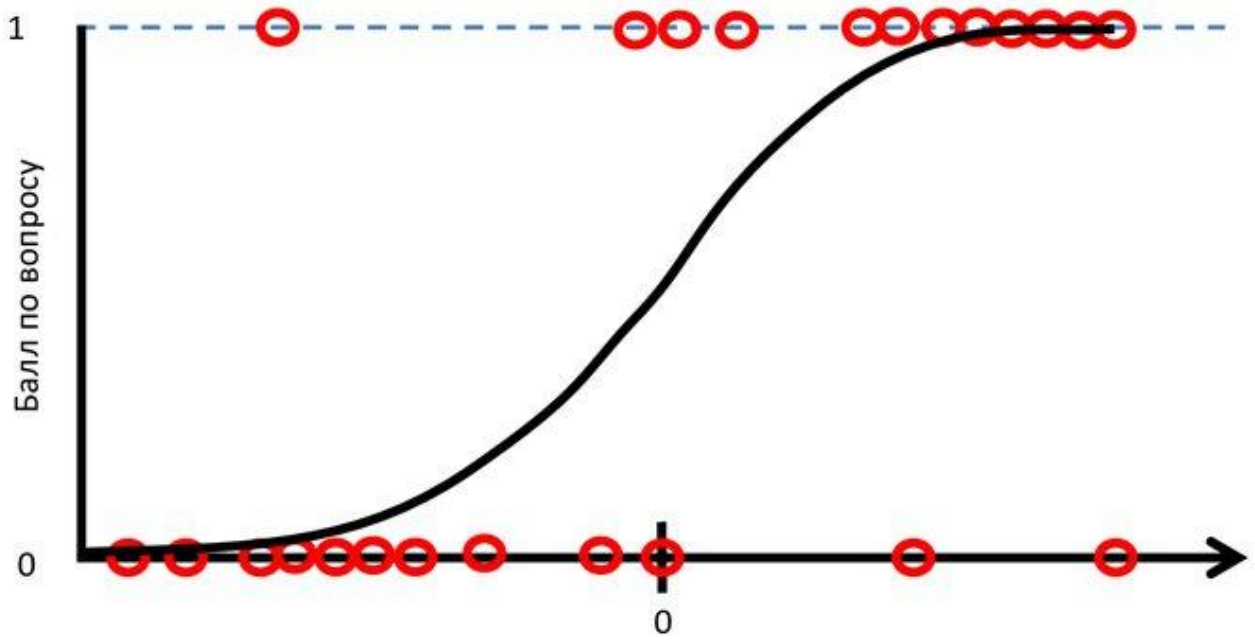


Модель дерево решений



Метод опорных векторов

Логистическая регрессия

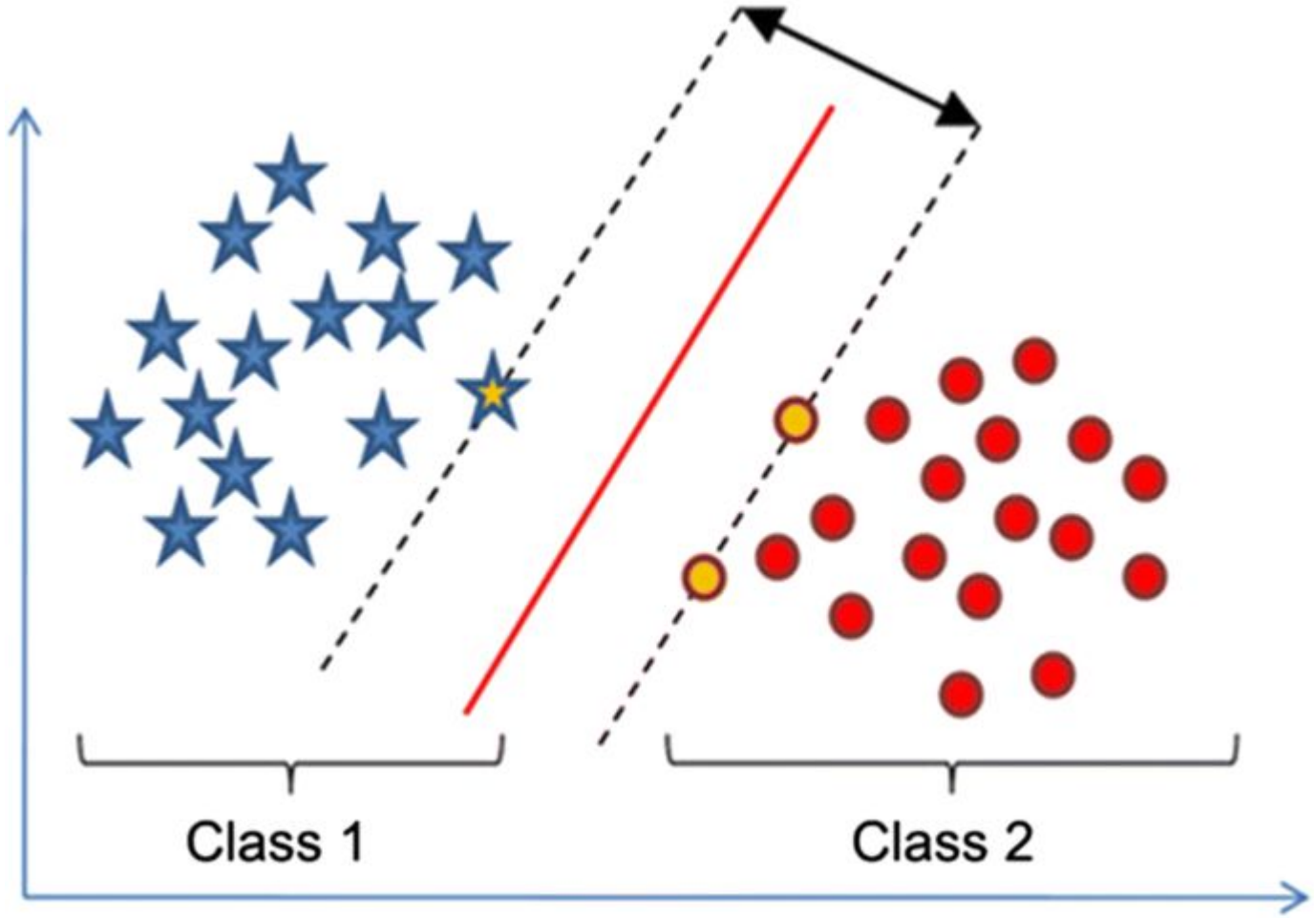


Логистическая регрессия вычисляет вероятность того, что данное исходное значение принадлежит к определенному классу.

В модели по умолчанию использовался параметр C = 1

При лучших параметрах C =0.1

Метод опорных векторов

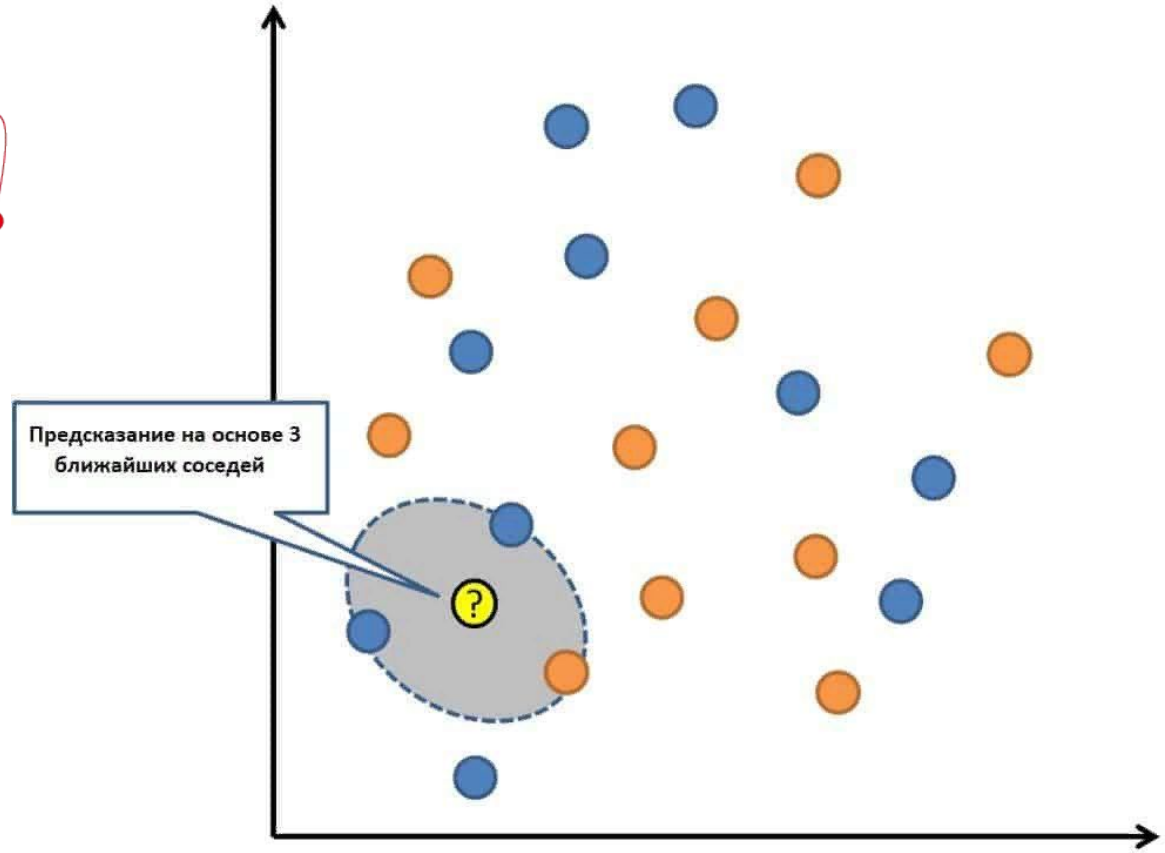


SVM стремится найти оптимальную гиперплоскость, которая разделяет данные на разные классы. Точки данных, расположенные по обе стороны от гиперплоскости, могут быть отнесены к разным классам.

Параметры базовой модели: C=1, gamma =1;

Параметры наилучшей модели: C=100, gamma = 2.

Метод k-ближайших соседей



«Посмотри на соседей вокруг,
какие из них преобладают,
таковым ты и являешься.»

Алгоритм:

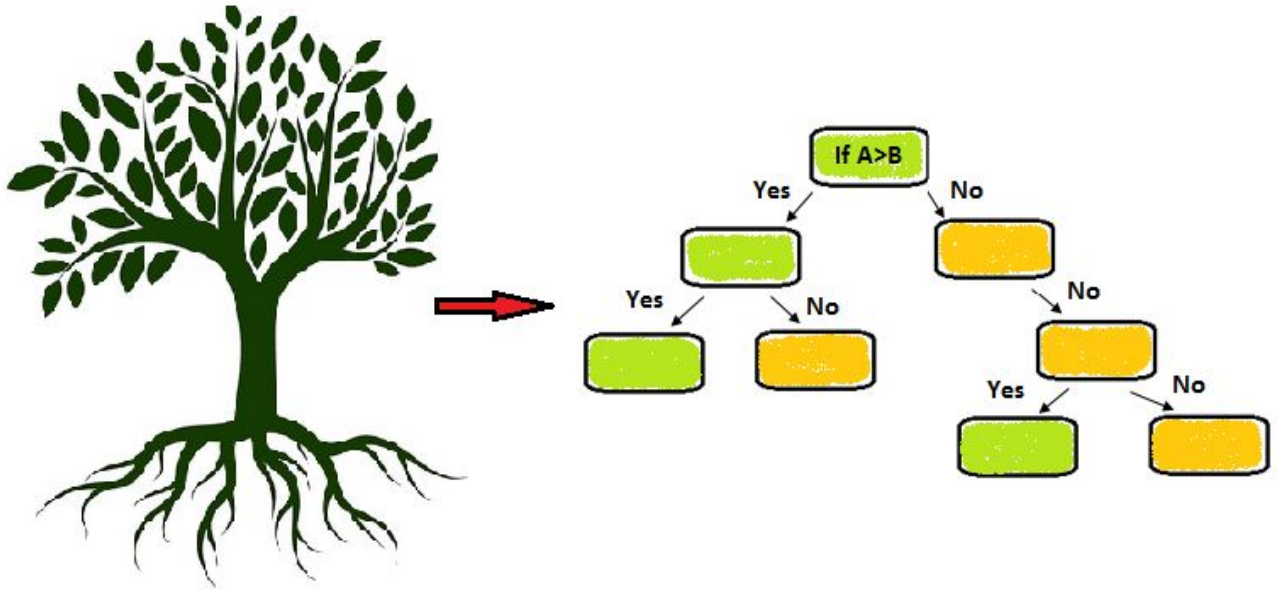
- Вычислить расстояние до каждого из объектов обучающей выборки;
- Отобрать k объектов обучающей выборки, расстояние до которых минимально;
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

Значение соседей в базовой модели = 5;

При наилучшем результате – 15.

Дерево решений

Данные непрерывно разделяются в соответствии с определенным параметром. Это древовидный классификатор, в котором внутренние узлы представляют характеристики набора данных, ветви представляют правила принятия решений, а каждый конечный узел представляет результат.



Базовые параметры модели:

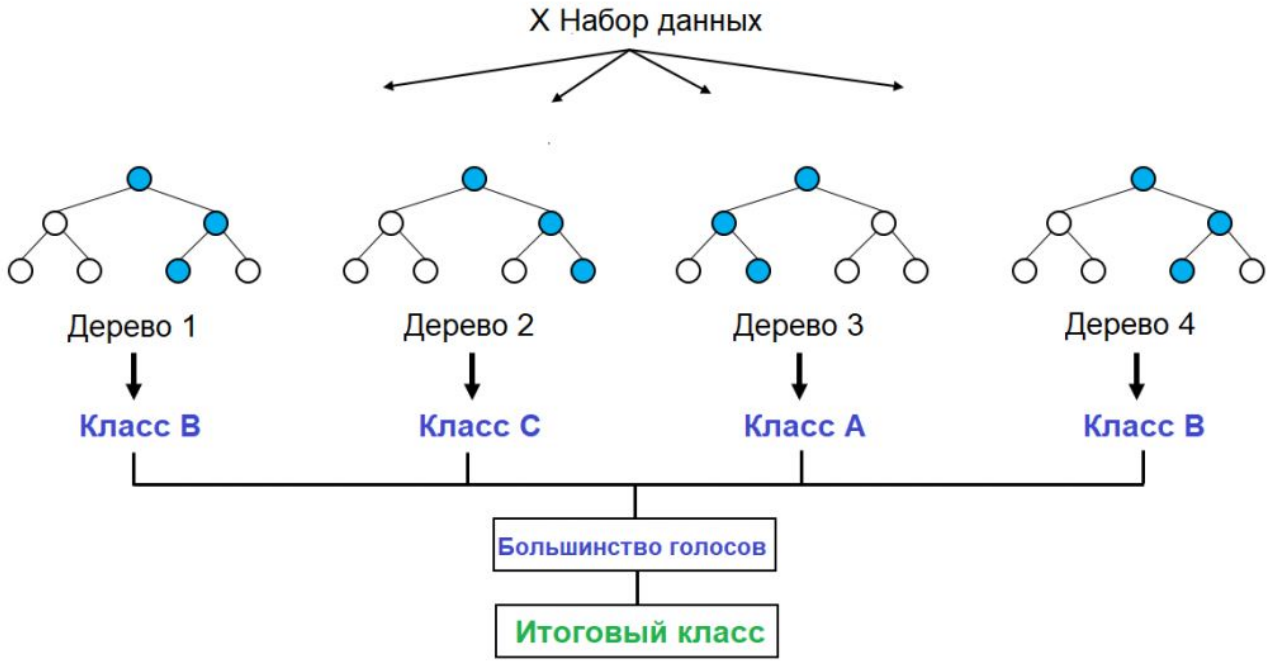
- Минимальное количество выборок для расщепления узла = 2
- Минимальное количество выборок для конечного узла = 1

Наилучшие параметры модели:

- Минимальное количество выборок для расщепления узла = 3
- Минимальное количество выборок для конечного узла = 6
- Глубина = 3

Случайный лес

Случайный лес состоит из большого количества отдельных деревьев решений. Каждое отдельное дерево в случайном лесу выдает прогноз класса, и класс с наибольшим количеством голосов становится прогнозом модели.



Базовые параметры модели:

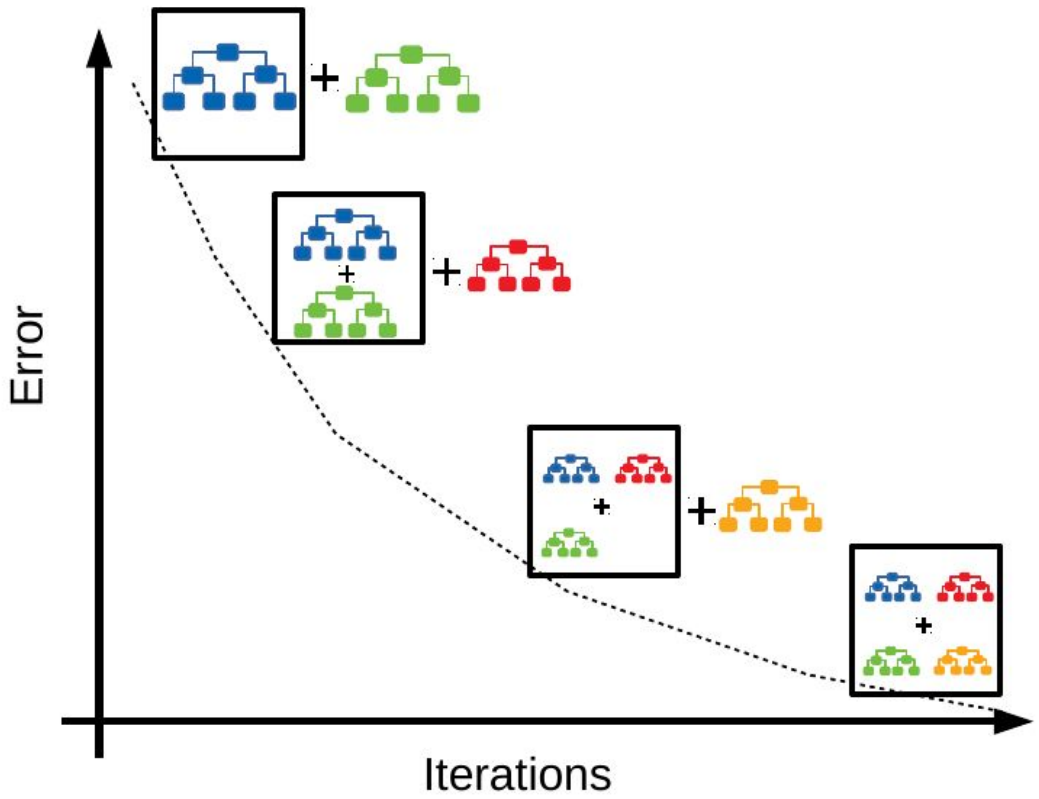
- Минимальное количество выборок для расщепления узла = 2
- Минимальное количество выборок для конечного узла = 1
- Количество деревьев = 100

Наилучшие параметры модели:

- Минимальное количество выборок для расщепления узла = 7
- Минимальное количество выборок для конечного узла = 8
- Количество деревьев = 350

Градиентный бустинг

Это алгоритм, который минимизирует функцию потерь, путем последовательного добавления деревьев по одному шагу за раз. После каждой итерации нам нужно быть ближе к нашей окончательной модели. Каждая итерация должна уменьшать значение нашей функции потерь.



Базовые параметры модели:

- Скорость обучения = 0.1
- Число деревьев = 100
- Глубина = 3

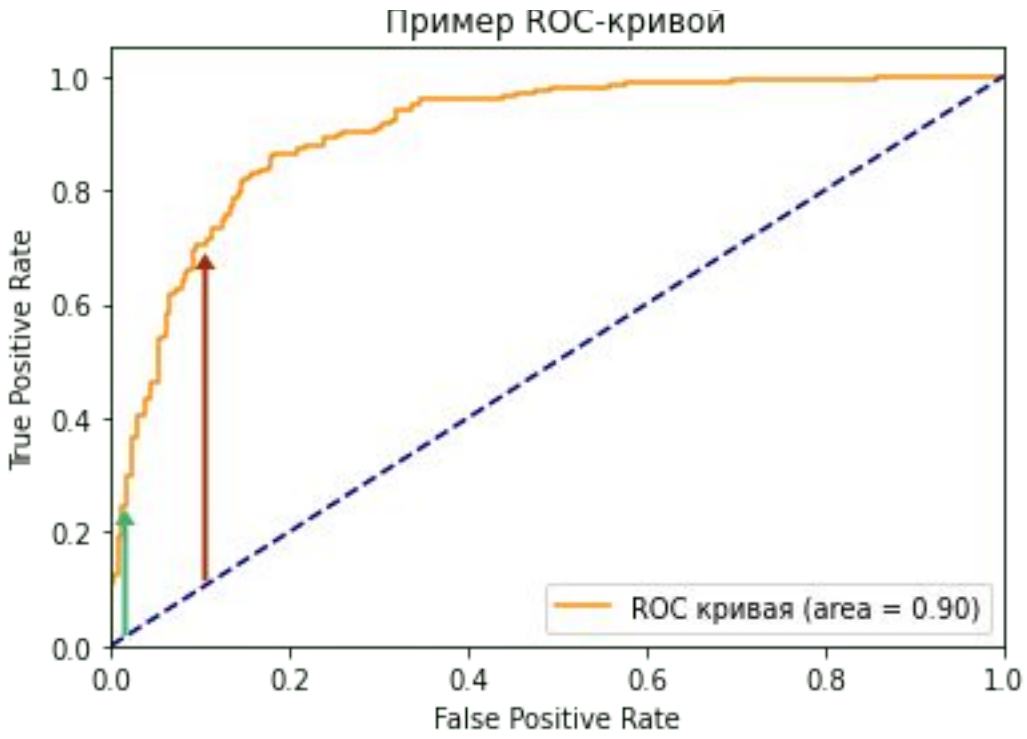
Наилучшие параметры модели:

- Скорость обучения = 0.9
- Число деревьев = 10
- Глубина = 3

МЕТРИКА ОЦЕНКИ КАЧЕСТВА

$$TPR = \frac{TP}{TP + FN} \cdot 100\,\%$$

$$FPR = \frac{FP}{TN + FP} \cdot 100\,\%$$



ROC-анализ — аппарат для анализа качества моделей.

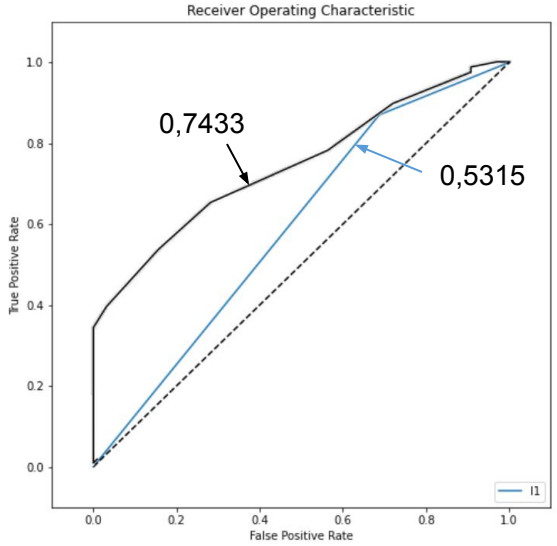
ROC кривая показывает отношение TPR к FPR.

Где, TPR – показывает, какой процент среди всех positive предсказан верно, а FPR – какой процент среди всех negative предсказан неверно.

Чем больше площадь под кривой (AUC), тем лучше классификация.

Модель	**Фактическое**	
	Положительно	Отрицательно
Положительно	TP	FP
Отрицательно	FN	TN

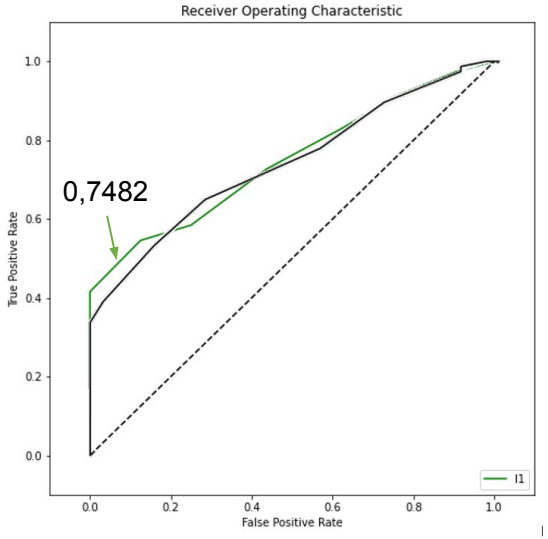
РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ

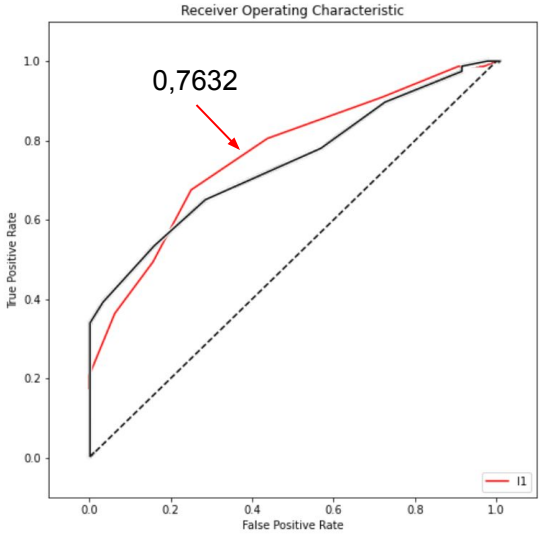


Синим – график для модели с базовыми параметрами

Черный – график с наилучшими параметрами

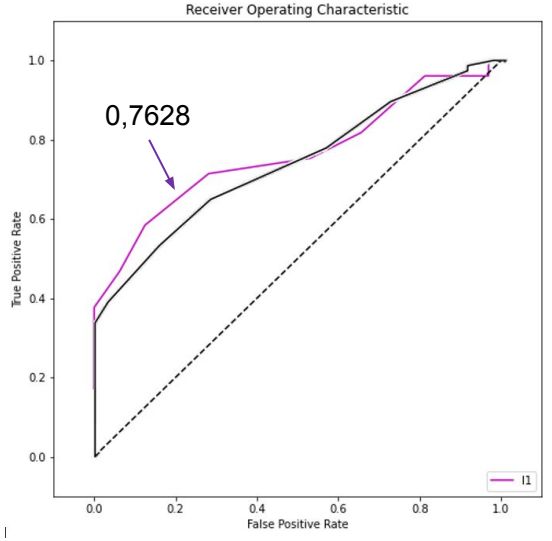
Зеленый – график без признака «пол»





Красный – график без признака «пол» и признака «отношение альбумина к глобулину»

Фиолетовый – график без признаков «пол», «отношение альбумина к глобулину» и «щелочная фосфатаза»



ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИДЕНТИФИКАЦИИ ЗАБОЛЕВАНИЙ ПЕЧЕНИ

РЕЗУЛЬТАТЫ

Методы классификации	Базовая модель	Лучшие параметры	После удаления признака «пол»	Удаление признака «пол» и данных отношения альбумина к глобулину	Удалении признака «пол», данных отношения альбумина к глобулину и данных щеточной фосфатазы
Логистическая регрессия	0,5769	0,8149	0,7431	-	-
Случайный лес	0,6134	0,7508	0,7427	-	-
KNN	0,5315	0,7433	0,7482	0,7631	0,7628
Дерево решений	0,5848	0,6648	0,6899	0,6694	-
SVM	0,5000	0,6356	0,5522	-	-
Градиентный бустинг	0,5418	0,7449	0,7149	-	-

ВЫВОДЫ

В результате работы были выполнены следующие задачи:

- Проанализированы наиболее распространенные заболевания печени
- Рассмотрены существующие методы классификации в машинном обучении
- Применение методов классификации в машинном обучении к выбранному набору данных

СПАСИБО ЗА ВНИМАНИЕ!