МИНОБРНАУКИ РОССИИ
Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
Кафедра Вычислительной техники

отчет

по лабораторной работе № 2

по дисциплине «Основы машинного обучения»

Тема: **«Применение методов машинного обучения и их оценка для задач классификации»**

Студенты гр. 3311 Аршин А. Д
Баймухамедов Р. Р.
Пасечный Л. В.

Преподаватель Петруша П. Г.

Санкт-Петербург

2025

Цель работы

Получение и закрепление навыков в применении методов машинного обучениях и их оценки для задач классификации.

Задание работы

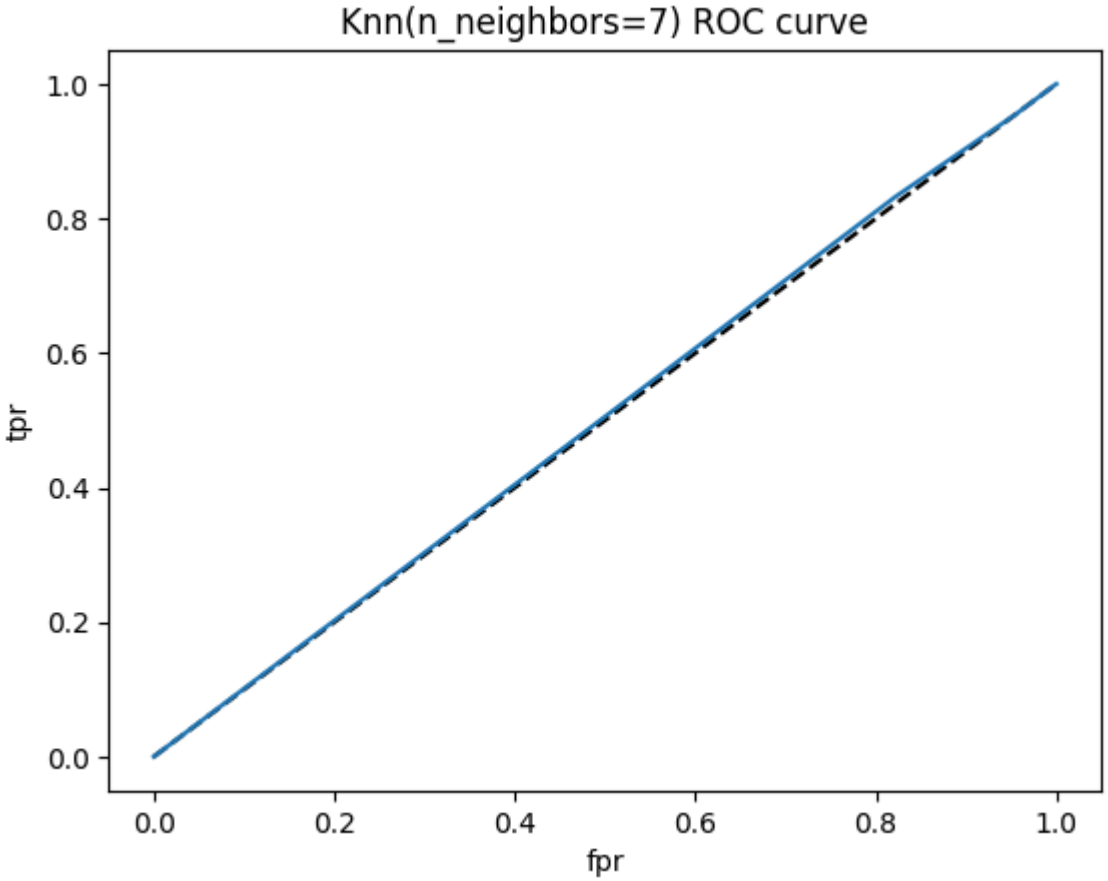
На самостоятельно выбранном в предыдущей лабораторной работе наборе данных обучить модель для решения задач классификации следующими методами:

- K-nn (Метод ближайших соседей)
- SVM (Метод опорных векторов)
- Random Forest (Случайный лес)

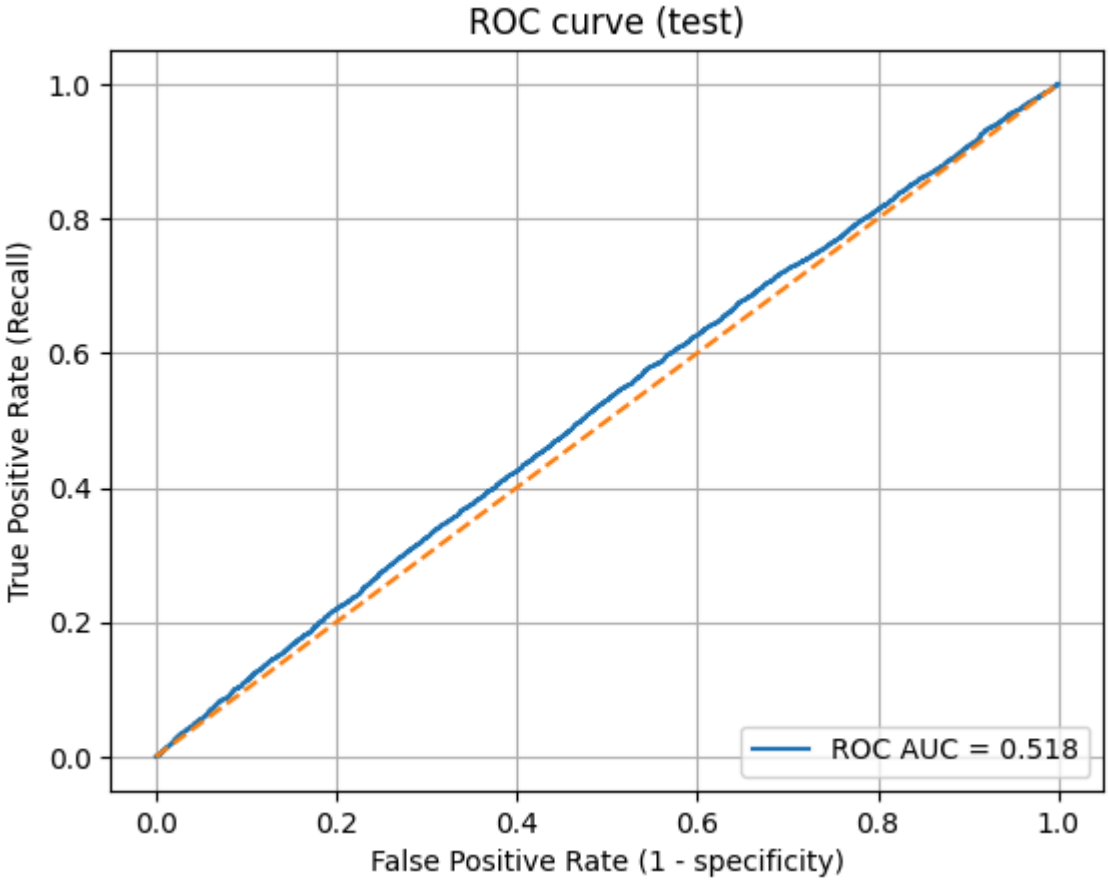
Основные шаги для обучения модели разными методами описаны в Google Colab. Здесь же приведем результат обучения

Лабораторная работа в среде Google Colab - https://colab.research.google.com/drive/1gVbkpqhRaBVxy9vG7_4ODncTRRoH5k_Z?usp=sharing

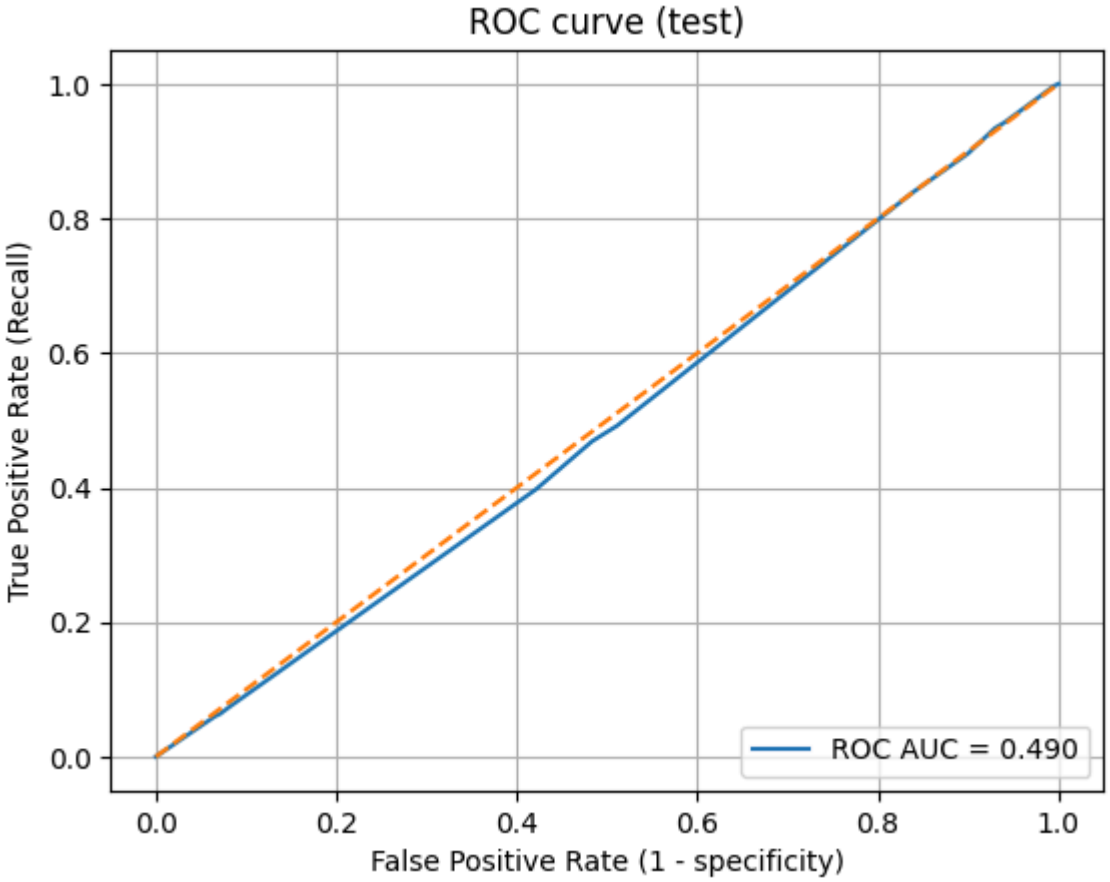
Альтернативный вариант - GitHub https://github.com/brick1ng5654/course-3/tree/RafaelB/boml/lab_02



ROC-кривая при использовании метода ближайших соседей.
ROC AUC ≈ 0,5



ROC-кривая при использовании метода опорных векторов.
ROC AUC ≈ 0,5



ROC-кривая при использовании метода случайного леса.
ROC AUC ≈ 0,5

Вывод

Что же как мы видим результат плох, значение ROC AUC примерно равно 0.5 на каждой модели, а значит модели работают словно бросок монеты. Ни обычный результат, ни его инвертированная версия не даст хорошего результата.

В ходе экспериментов было предложена версия задать меньший вес большинству признаков, которые аналитически меньше всего связаны с болезнями. Больший же весь будут иметь главные признаки как индекс массы тела, инсулин, холестерин и так далее. Но поскольку это была экспериментальная идея, то проще всего было реализовать это таким образом, при котором мы просто отбросим признаки с маленьким весом. Если предположение о шуме признаков верно, то мы бы заметили разницу в ROC AUC, однако этого не произошло. Как нам кажется это возможно связано с вероятным искусственным происхождением данного датасета, поскольку целевой признак достаточно размыт. Ввиду этого получается,

что основываясь на выбранных признаках невозможно получить уверенную классификацию объекта.

Дополнительно были изучены работы с выбранным датасетом других пользователей на сайте Kaggle. В ходе исследования удалось собрать следующие данные:

Из 8 работ:

- 5 работ сразу не подходят или отсутствия содержательной базы, вычисления не того целевого признака или имеют другой подход к задаче (безуспешная кластеризация)
- 2 работы, проверяющий много методов машинного обучения и получающий в результате аналогичный в нашей работе результат (ROC AUC = `0.5`)
- 1 работа, получающий ROC AUC = `1.0`, что звучит неправдоподобно и скорее всего автор допустил ошибки