

Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

**Лабораторная работа №1 по курсу
«Искусственный интеллект (Машинное обучение)»**

Студент: Суханов Е. А.

Группа: М8О-406Б-19

Преподаватель: Ахмед Самир Халид

Дата:

Оценка:

Подпись:

Москва, 2022

Описание

Задача:

- 1) реализовать следующие алгоритмы машинного обучения: Linear/ Logistic Regression, SVM, KNN, Naive Bayes в отдельных классах
- 2) Данные классы должны наследоваться от BaseEstimator и ClassifierMixin, иметь методы fit и predict
- 3) Вы должны организовать весь процесс предобработки, обучения и тестирования с помощью Pipeline
- 4) Вы должны настроить гиперпараметры моделей с помощью кросс валидации, вывести и сохранить эти гиперпараметры в файл, вместе с обученными моделями
- 5) Прodelать аналогично с коробочными решениями
- 6) Для каждой модели получить оценки метрик: Confusion Matrix, Accuracy, Recall, Precision, ROC_AUC curve
- 7) Проанализировать полученные результаты и сделать выводы о применимости моделей
- 8) Загрузить полученные гиперпараметры модели и обученные модели в формате pickle на гит вместе с jupyter notebook ваших экспериментов

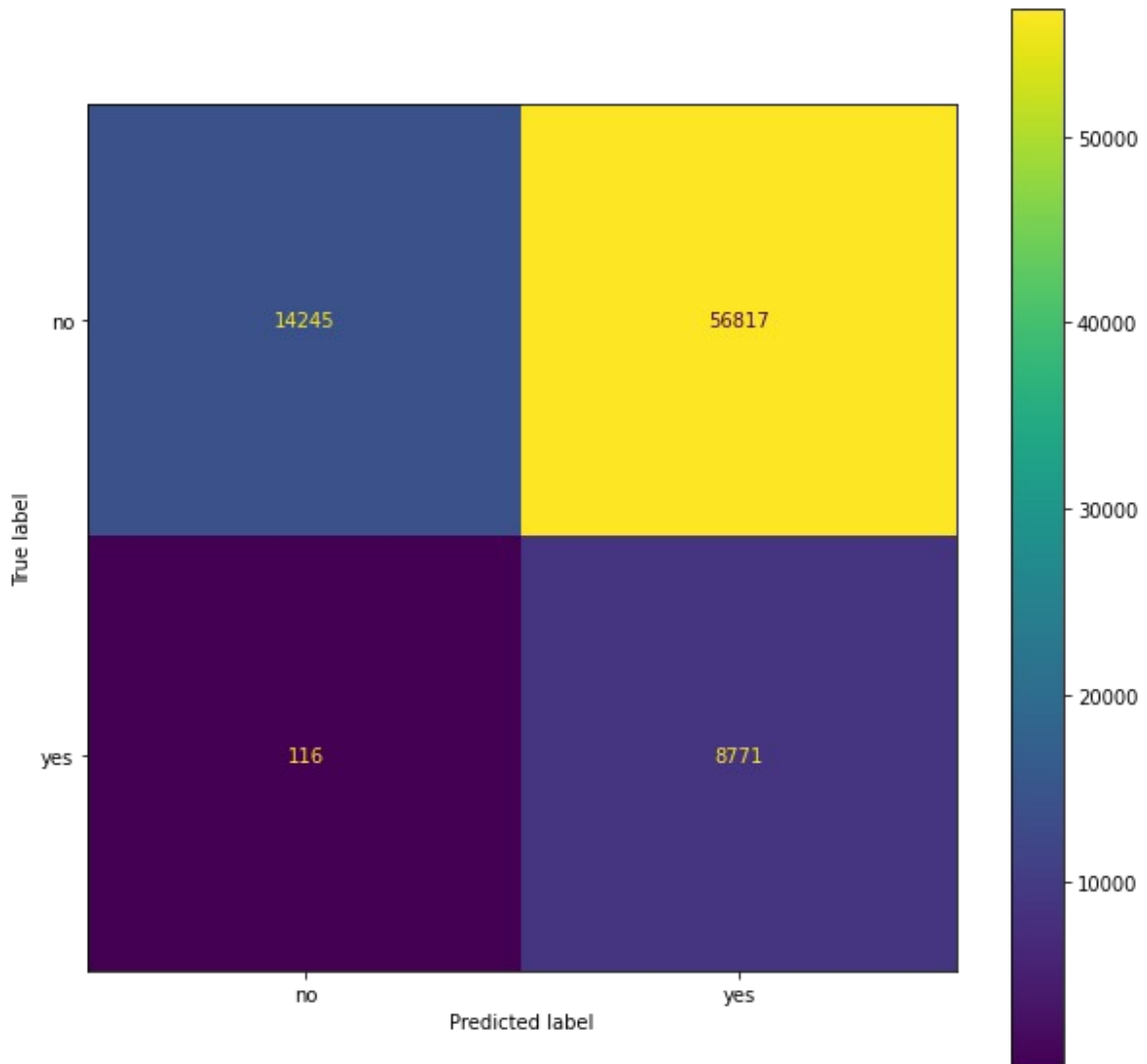
Я взял датасет «Personal Key Indicators of Heart Disease» (<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>). В нем содержится 18 полей и 400 тысяч записей. Я думаю, что для моих целей его хватит с лихвой. Поля могут быть как и числовые, так и символьные (с ограниченным числом вариантов).

На основе подготовленного датасета, взятого из прошлой лр, Я хочу научиться определять, с какой вероятностью у опрашиваемого человека есть проблемы с сердцем.

Ход выполнения лабораторной работы

Во-первых я написал алгоритмы Logistic Regression, SVM, KNN и Naïve Bayes в отдельных классах. А так же обучил модели и сравнил их с библиотечными реализациями.

Linear Regression



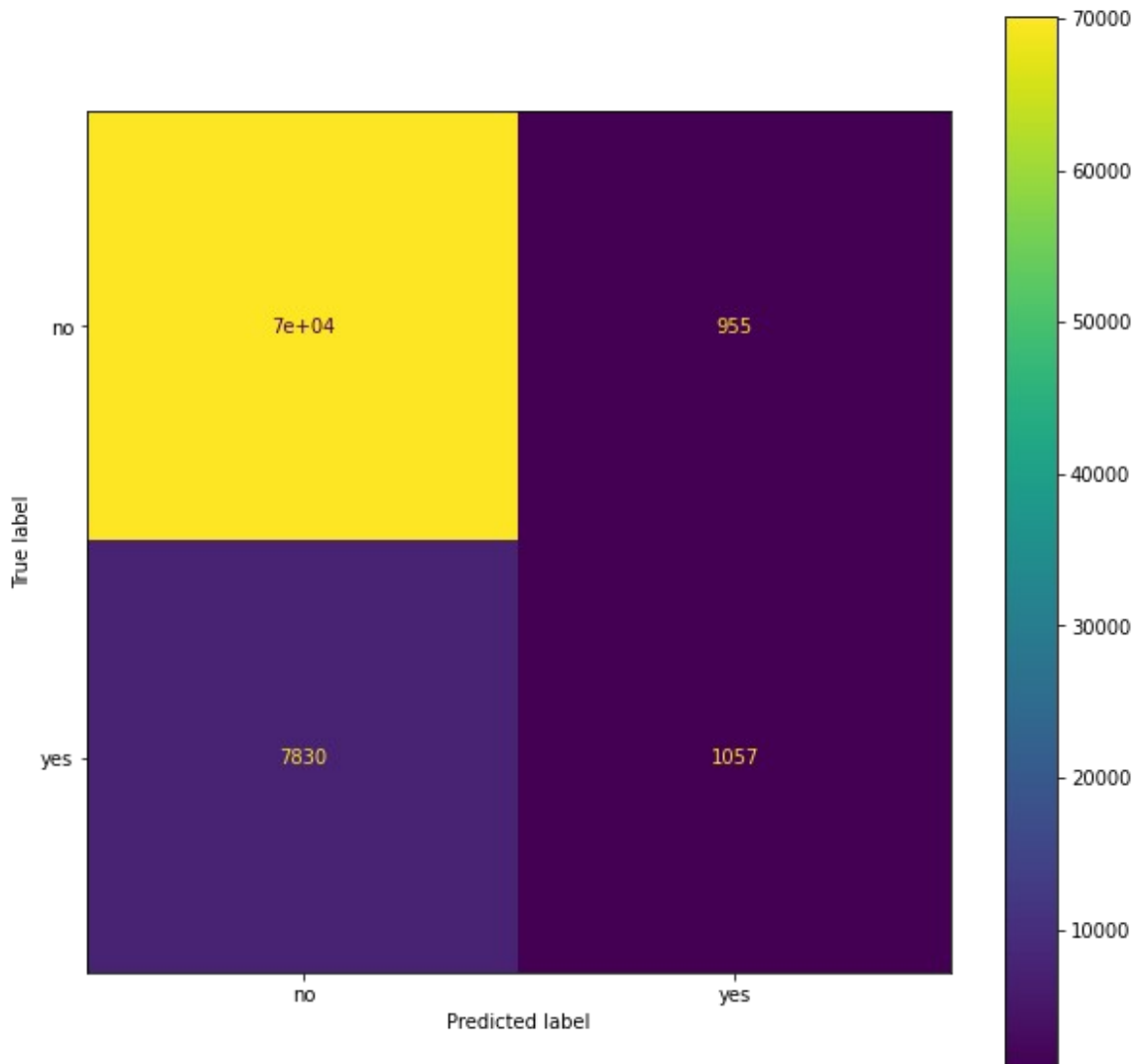
Линейная регрессия плохо подходит для классификации. Так как результат предсказания модели нужно пропустить через функцию, которая будет давать окончательный ответ.

Несмотря на низкую точность, можно заметить, что моя модель имеет высокий recall. Это означает, что когда человек действительно болен ССЗ (сердечно-сосудистым заболеванием), то модель с высокой вероятностью это обнаруживает.

Однако у моделей много ложных срабатываний, о чем нам говорит precision.

Тем не менее, в нашей задаче важно иметь высокий recall.

Logistic Regression

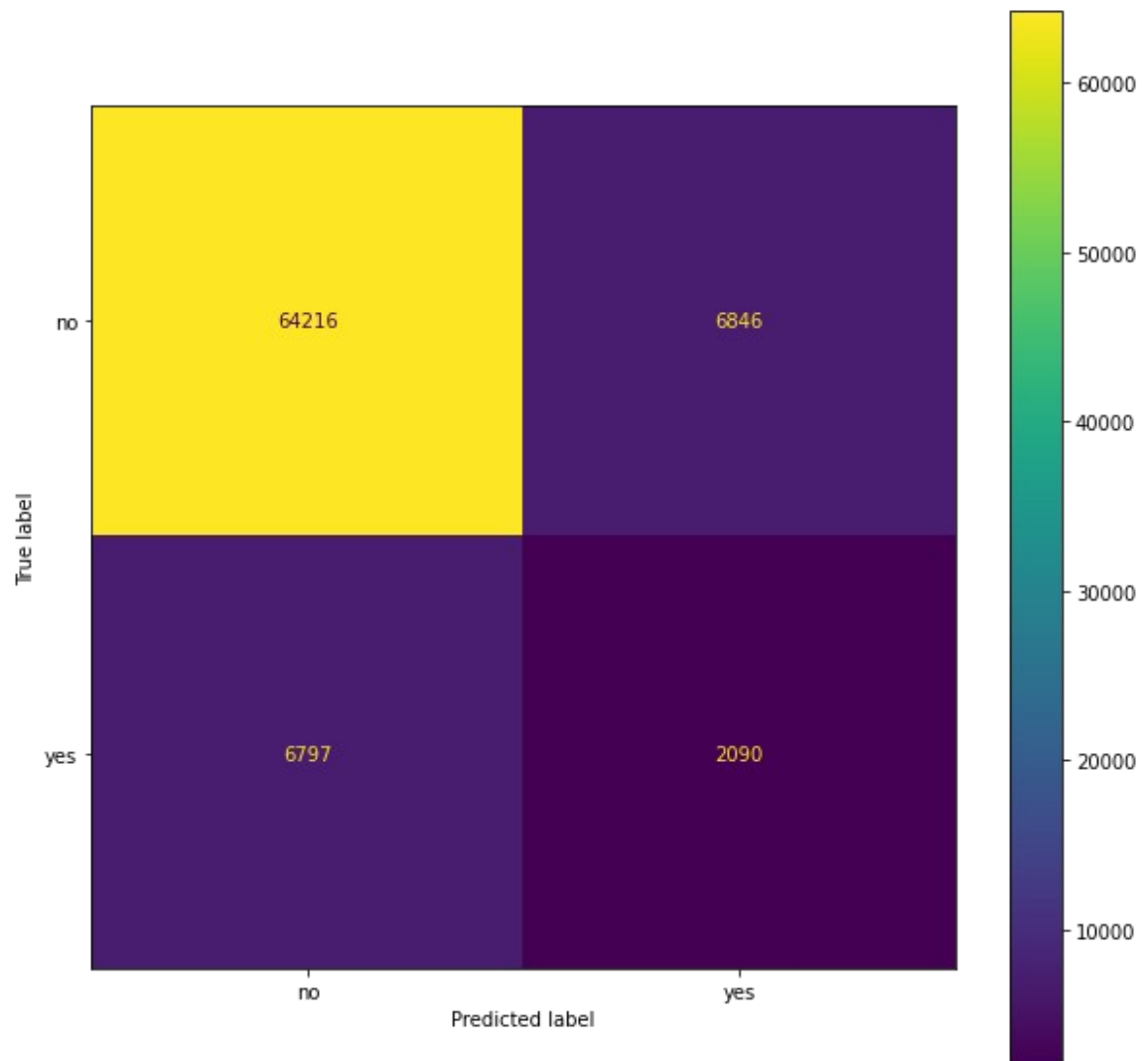


Логистическая регрессия больше подходит для классификации, так как она выдает значения в пределах четких границ, которые просто интерпретировать.

У моделей высокий Accuracy, но это не означает, что они хорошо работают. У них очень много False-Positive срабатываний. О чем нам говорит Recall.

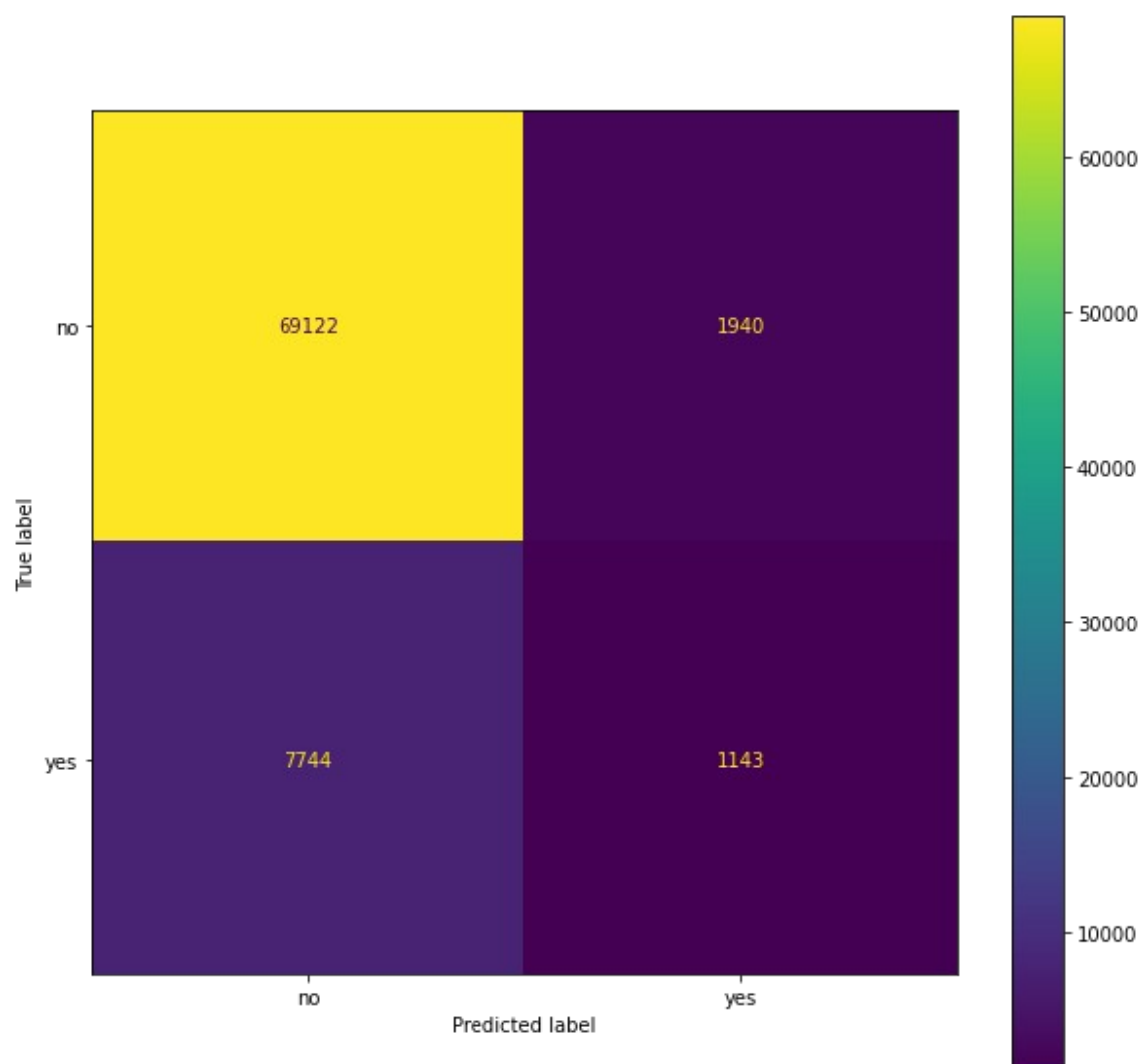
Поэтому такую модель нельзя использовать в нашем случае.

SVM



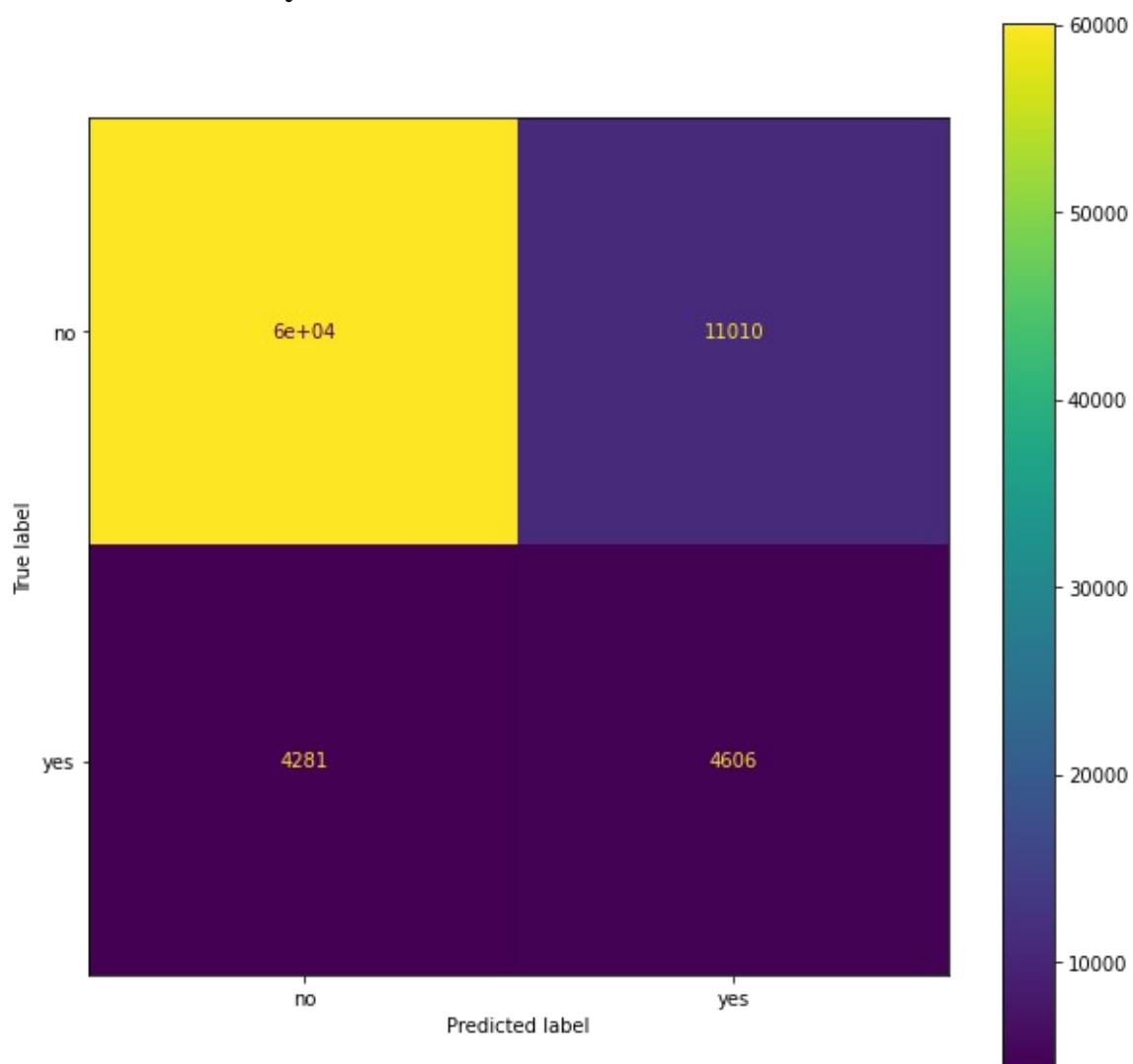
SVN дает более лучший результат, чем логистическая регрессия. Тем не менее, recall все равно низкий.

KNN



Результаты сравнимы с логистической регрессией.

Naive Bayes



Данная модель дает хороший recall, и, в отличие от линейной регрессии, дает меньше False-Negative ответов.

Выводы

С моей задачей лучше всего справляется линейная регрессия, если нам очень важно находить людей с ССЗ. Так как намного важнее узнать, есть ли у человека ССЗ, чем узнать, что у человека нет ССЗ.

Думаю, что я мог бы получить более хорошие результаты, если бы лучше обработал данные.