МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южный федеральный университет»
Институт высоких технологий и пьезотехники

Кафедра прикладной информатики и инноватики

Направление подготовки:
09.03.03 "Прикладная информатика"

Отчёт на тему:
“Классификация болезней сердца на основе медицинских данныха”

Выполнили студенты 3 курса _____ Чудиков Д.М.
подпись
_____ Пушкин В.А.
подпись

Ростов-на-Дону – 2024

Цель работы:

Перед нами стояла задача разработать модель, способную предсказывать вероятность наличия сердечных заболеваний на основе имеющихся у нас данных реальных пациентов. В данной работе мы так же должны будем применить изученный нами фреймворк машинного обучения PySpark для анализа данных и вывода результатов.

Ход работы:

Первый этап
Сбор данных

С сайта «Kaggle» нами был выбран соответствующий датасет со всеми необходимыми данными(**Heart Disease Dataset 2.0**). Все данные о сердечно-сосудистых заболеваниях, собранны из хранилища UCI в следующих местах: Кливленд, Венгрия, Швейцария и Лонг-Бич, Вирджиния. Датасет имеет более 300 000 записей. В полях датасета содержится следующая информация: возраст, пол, категория боли в груди, показания стенокардии, наличие или отсутствие дополнительных симптомов, показания давления, показания холестерина, показания превышает ли уровень сахара в крови норму или нет, Максимальная частота сердечных сокращений и конечный результат уже проведённых исследований – болен был этот человек или здоров.

Второй этап
Анализ данных и их обработка

Здесь мы проанализировали данные, содержащиеся в столбцах датасета. Нам было необходимо, чтобы данные, которые будут взаимодействовать друг с другом в дальнейшем, были представлены в одинаковых единицах измерения. Так же все данные о наличии того или иного заболевания или симптома были исключительно бинарные – то есть имели значение 0 или 1, где 0 обозначает отсутствие, а 1 – наличие. Подобные преобразования данных делались при помощи языка программирования Python.

Третий этап
Разработка и обучение модели

После того, как данные были полностью готовы к работе мы приняли решение использовать 3 разных алгоритма машинного обучения, а именно: метод линейной регрессии, метод опорных векторов (SVM) и метод случайного леса. Нам предстояло выяснить какой из этих методов окажется эффективнее для работы с этим датасетом.

Первым методом была протестирована линейная регрессия, мы разделили данные на тренировочные и тестируемые(те, на которых будет проверяться правильность предсказывания), обучили модель и запустили её. Для оценки точности предсказания был вызван специальный оценщик точности, объект класса BinaryClassificationEvalutor.Точность предсказания с этим методом составила примерно 0.999577

Далее по такой же логике были обучены модели для методов опорных векторов и случайного леса с небольшим дополнением. В методе опорных векторов есть такой параметр как «Количество итераций», а в методе случайных лесов – «количество деревьев». Этот параметр позволяет регулировать количество прогонов данных. Мы сделали несколько таких запусков для более точного предсказания. Лучшая точность составила примерно 0.71 и 0.77 соответственно.

Таким образом, было выявлено, что метод линейной регрессии показал самую высокую точность и следовательно оказался лучшим методом из представленных для работы с этим датасетом.

Четвертый этап
Статистка из данных

И напоследок мы решили провести небольшой анализ и выявить как различаются разные показатели у мужчин и женщин, исходя из данных датасета. В конце работы предоставлены результаты о : количестве мужчин и женщин с сердечными заболеваниями, среднем возрасте мужчин и женщин с сердечными заболеваниями, среднем уровне холестерина у мужчин и женщин с сердечными заболеваниями, количестве людей с ангиной среди мужчин и женщин с сердечными заболеваниями

Заключение

Цель проекта была достигнута, мы разработали модели для предсказания наличия сердечных заболеваний на основе медицинских данных пациентов.

Лучшим методом для предсказания метки класса (болезни сердца человека) оказался метод линейной регрессии с точностью 0.999657

Итого получаем, что в нашем случае, на данном датасете, мы с вероятностью почти 100% можем предсказать есть ли у человека болезнь сердца или нет при помощи метода линейной регрессии.