Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Лабораторная работа №0 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: вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще.

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).

Я хочу на основе этих данных научиться определять, с какой вероятностью у опрашиваемого человека есть проблемы с сердцем. Если быть более точным, то я хочу соединить поле HeartDisease и Stroke в одно, и научиться предсказывать по стальным данным его значение. То есть перед нами стоит задача бинарной классификации.

Описание свойств:

- HeartDisease: Была ли ишемическая болезнь сердца (ИБС) или инфаркта миокарда (ИМ);
- BMI: Индекс массы тела (ИМТ);
- Smoking: Выкурил ли человек за всю свою жизнь хотя бы 100 (5 пачек) сигарет?;
- AlcoholDrinking: Сильно пьющие (взрослые мужчины, выпивающие более 14 напитков в неделю, и взрослые женщины, выпивающие более 7 напитков в неделю;
- Stroke: Был ли инсульт?;
- PhysicalHealth: Сколько дней за последний месяц вы себя чувствовали физически плохо (травмы, физические заболевания)?;
- MentalHealth: Сколько дней в течение последних 30 дней ваше психическое здоровье было плохим?;
- DiffWalking: Испытываете ли вы серьезные трудности при ходьбе или подъеме по лестнице?;
- Sex: Вы мужчина или женщина?;
- AgeCategory: Возрастная категория четырнадцатого уровня;
- Race: Раса;
- Diabetic: У вас есть диабет?;
- PhysicalActivity: Взрослые, которые сообщили, что занимались физической активностью или физическими упражнениями в течение последних 30 дней, помимо своей обычной работы;

- GenHealth: Могли бы вы сказать, что в целом ваше здоровье – (варианты ответа);
- SleepTime: В среднем, сколько часов сна вы получаете за 24-часовой период?;
- Asthma: У вас есть астма?;
- KidneyDisease: Не считая камней в почках, инфекции мочевого пузыря или недержания мочи, вам когда-нибудь говорили, что у вас заболевание почек?;
- SkinCancer: У вас есть рак кожи?.

Ход выполнения лабораторной работы

Как видно, здесь есть категориальные и числовые признаки. Давайте проанализируем данный датасет, что бы сделать выводы о них. И затем определиться, какие поля нам оставить, а какие убрать.

Мы имеем только 4 числовых признака. И с ними все понятно, но как нам работать с категориальными признаками? Оказывается, для таких признаков можно использовать one-hot и/или label encoding.

Затем, с помощью матрицы корреляции мы убрали признаки, которые дают маленький вклад.

Я решил не нормировать данные, так как большинство из них имеют несколько возможных значений, а численные свойства не имеют четких границ.

Выводы

Я считаю, что нашел датасет на довольно неплохую тему. Я Научился выделять нужные и не нужные признаки и подготавливать данные для последующего использования в алгоритмах машинного обучения.