ПРОГНОЗИРОВАНИЕ РИСКА РАЗВИТИЯ СЕРДЕЧНО - СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Емельянов А.С.
Томский политехнический университет, ИШИТР, e-mail: andreiomsk02@gmail.com

Введение

В наше время всё чаще возникают задачи, требующие решения обработки и анализа больших массивов данных. Задачи собраны названием Big Data. Анализ данных может проводиться как вручную, так и с помощью методов машинного обучения, то есть с помощью алгоритмов нахождения закономерностей и связей на основе эмпирического и теоретического опытов. Одна из областей применения алгоритмов машинного обучения - медицина.

В рамках работы будут исследованы алгоритмы машинного обучения к анализам, текстам и иным медицинским документам для выявления ключевых параметров, влияющих на сердечно - сосудистую систему, что может быть использовано для более точного выявления заболеваний этой системы.

Целью данной работы ставлю поиск наиболее оптимального алгоритма машинного обучения для выявления ключевых параметров, влияющих на сердечно - сосудистую систему.

В исследовании обрабатываются данные результатов обследования пациентов, строятся метрики для выявления зависимых переменных.

Описание алгоритма

Для того, чтобы датасет подвергался наиболее эффективной обработке и метрики были более точными - нужно провести первоначальный и одномерный, двумерный и многомерный анализы. Далее будет описан алгоритм того, как нужно обрабатывать и анализировать данные.

1. Подгружаем датасет и заданные переменные. Для построения графиков и зависимостей используем библиотеки numpy, pandas, seaborn. Визуализировать данные будем с помощью библиотеки matplotlib. Модуль os обеспечивает работу функций с операционной системой.

2. С помощью метода head() по датасету видно, что имеется определённый целевой класс. В случае, когда он равен 0 пациент здоров, в случае значения равного 1 он имеет ССЗ. Цифры 2 и 3 говорят о превышении показателей выше и значительно выше нормы соответственно.

3. Методом describe() выявляем минимальное, максимальное, среднее и средне - квадратичное отклонение.

4. Определим как переменные распределены среди целевого класса на рисунке 1.

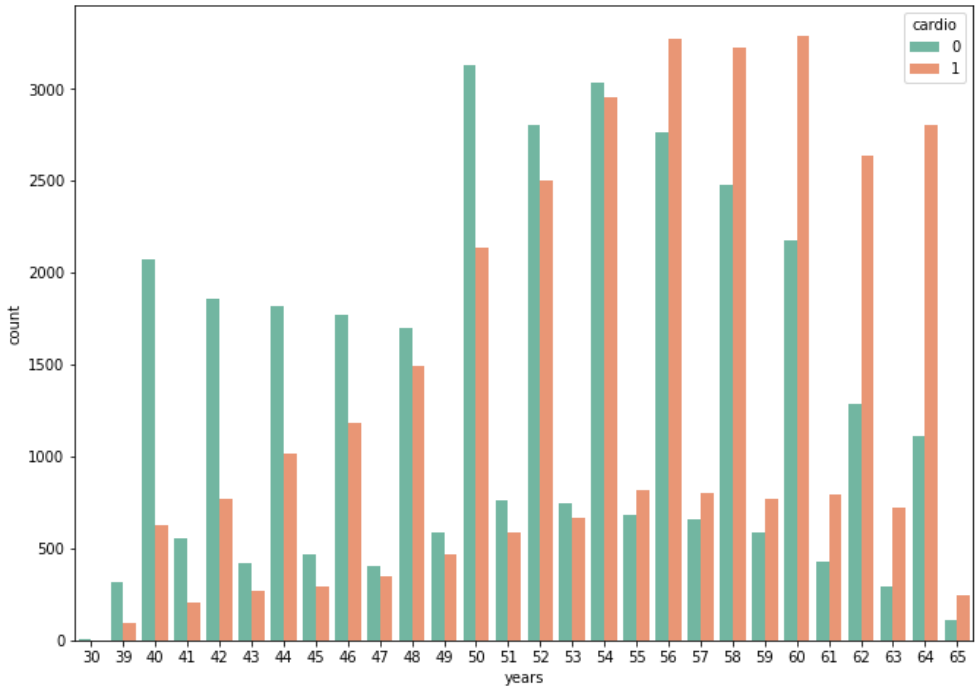

Рис. 1. График подсчёта людей с ССЗ и без ССЗ

5. На рисунке 2 рассмотрим распределение категориальных переменных.

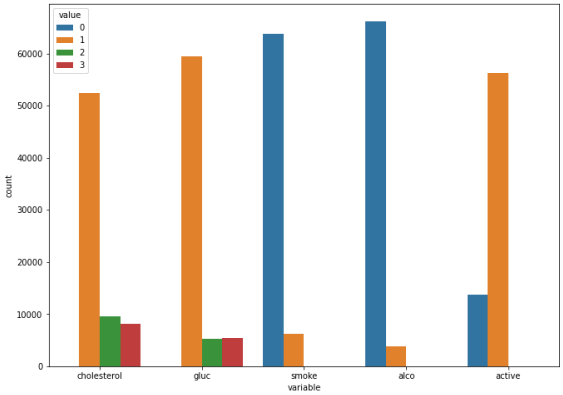

Рис. 2. График подсчёта людей по категориям

Таким образом был проведён первоначальный и первичный анализы. В первоначальном получена информация о датасете, типах данных и категориальных переменных. В одномерном построены графики подсчёта людей без ССЗ и с ССЗ, люди разделены на категории.

Далее проведём двумерный анализ.

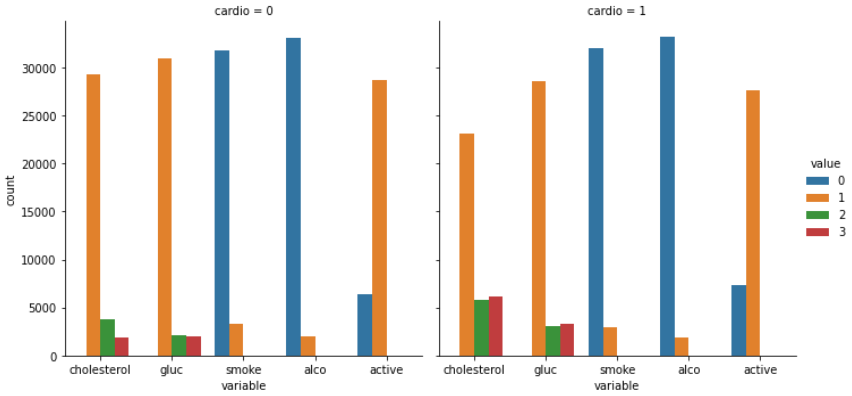

Рис. 3. График подсчёта людей по целевому классу

6. Из рисунка 3 отчётливо видно, что пациенты с ССЗ малоподвижны, имеют высокие показатели холестерина и глюкозы в крови.

7. Затем я группирую датасет по гендеру и выясняем, что большое количество курильщиков среди мужчин.

8. Очищаю датасет от выбросов и пропущенных значений.

9. Для того, чтобы выяснить очистились ли давления систолическое и диастолическое от выбросов и узнать распределение вероятностей между ними, построим ящик с усами.

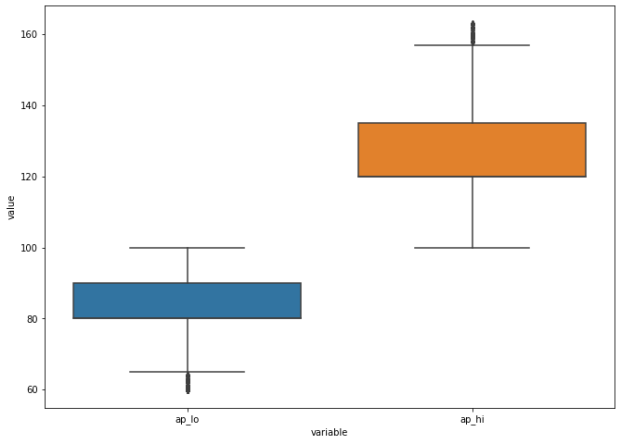

Рис. 4. Ящик с усами для давлений

10. Построим корреляционную матрицу для наглядности зависимостей атрибутов друг от друга.

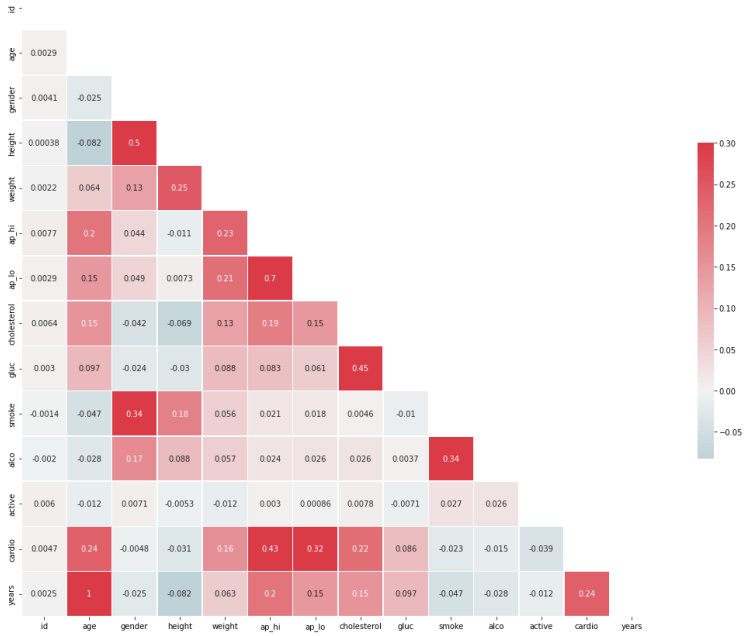

Рис. 5. Корреляционная матрица

На рисунке 5 видно какие коэффициенты значительно влияют на риск развития ССЗ. Если внимательно присмотреться, то по итогу сильное воздействие оказывают такие показатели как, возраст и уровень холестерина. Однако, стоит отметить, что они не особо коррелируют с целевым классом.

Заключение

В рамках данной работы был подготовлен большой массив данных (датасет), среди как здоровых пациентов, так и людей с ССЗ. Были обработаны и очищены от пропусков и выбросов данные результатов пациентов. Построены зависимости и метрики, показывающие переменные, которые в значительной степени влияют друг на друга. В дальнейшем будет произведён анализ работы и проведётся расширение набора данных для разработки точного алгоритма машинного обучения.

Список использованных источников

1. Губин Е. И. Методология подготовки больших данных. – URL: https://portal.tpu.ru/SHARED/g/GUBINE/academics/Tab/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F%20%D0%BF%D0%BE%D0%B4%D0%B3%D0%BE%D1%82%D0%BE%D0%B2%D0%BA%D0%B8%20%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B8.pdf (дата обращения 27.02.2023).
2. Представление данных корреляционного анализа. – URL: https://allasamsonova.ru/statistika/predstavlenie-dannyh-korreljacionnogo-analiza/ (дата обращения 27.02.2023).
3. Построение графиков в Python при помощи Matplotlib. – URL: https://python-scripts.com/matplotlib (дата обращения 27.02.2023).
4. Введение в многомерный анализ. – URL: https://habr.com/ru/post/126810/ (дата обращения 27.02.2023).