Московский Авиационный Институт

(Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

**Лабораторная работа №0 по курсу**

**«Машинное обучение»**

**Data Mining и исследование данных**

Студент: Моисеенков Илья Павлович

Группа: М80 – 308Б-19

Дата: 04.05.2022

Оценка: \_\_\_\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_

Москва, 2022

1. **Постановка задачи**

Найти набор данных и провести исследовательский анализ. Подготовить отчет с результатами исследования.

1. **Описание датасета**

Имеется датасет с информацией о пациентах США. Нужно выявить пациентов с высоким рисков сердечного приступа.

Имеем следующие сведения об опрошенных людях:

* **HeartDiseaseorAttack** - Таргет, был ли сердечный приступ у пациента
* *HighBP* - повышенное давление
* *HighChol* - повышенный холестерин
* *CholCheck* - была ли проверка на холестерин за последние 5 лет
* *BMI* - индекс массы тела
* *Smoker* - курит ли человек
* *Stroke* - был ли инсульт у человека
* *Diabetes* - есть ли диабет. Если есть, то какой степени
* *PhysActivity* - занимался ли человек физкультурой за последний месяц
* *Fruits* - ест ли человек фрукты каждый день
* *Veggies* - ест ли человек овощи каждый день
* *HvyAlcoholConsump* - потребляет ли человек много алкогольных напитков
* *AnyHealthcare* - имеет ли человек медицинскую страховку
* *NoDocbcCost* - был ли за последний год случай, когда нужно было попасть ко врачу, но не было денег на это
* *GenHlth* - субъективная оценка здоровья человека (1 = отличное, 2 = очень хорошее, 3 = хорошее, 4 = удовлетворительное, 5 = плохое)
* *MentHlth* - сколько раз за последний месяц наблюдались ментальные проблемы (депрессии, стресс итд)
* *PhysHlth* - сколько раз за последний месяц наблюдались проблемы после физических нагрузок (травмы)
* *DiffWalk* - есть ли проблемы с ходьбой
* *Sex* - пол (0 = женский, 1 = мужской)
* *Age* - возрастная группа

1 = 18-24

2 = 25-29

3 = 30-34

4 = 35-39

5 = 40-44

6 = 45-49

7 = 50-54

8 = 55-59

9 = 60-64

10 = 65-69

11 = 70-74

12 = 75-79

13 = 80+

* *Education* - уровень образования

1 = Never attended school or only kindergarten

2 = Grades 1 through 8 (elementary)

3 = Grades 9 through 11 (some high school)

4 = Grade 12 or GED (high school graduate)

5 = College 1-3 years (some college or technical school)

6 = College 4 years or more (college graduate)

* *Income* - годовой доход в долларах

1 = <10,000

2 = 10,000-14,999

3 = 15,000-19,999

4 = 20,000-24,999

5 = 25,000-34,999

6 = 35,000-49,999

7 = 50,000-74,999

8 = >75,000

Ссылка на датасет: <https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>

Датасет не содержит пропусков, можно приступать к анализу.

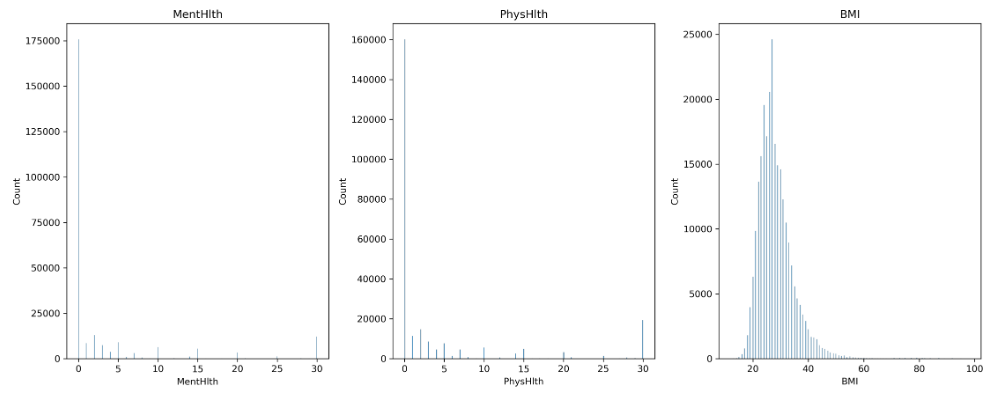
1. **Количественные признаки**

Имеем три количественных признака - количество ментальных и физических проблем и ИМТ.

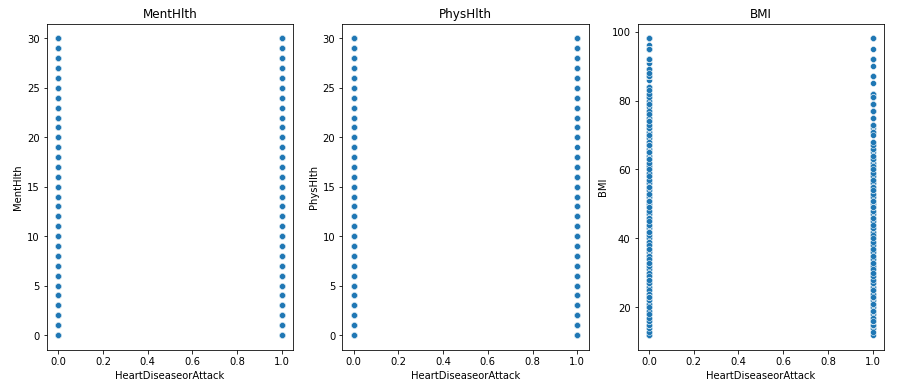
Посмотрим на статистическое описание признаков



Посмотрим на распределение этих величин



Также построим точечный график зависимости этих величин от таргета



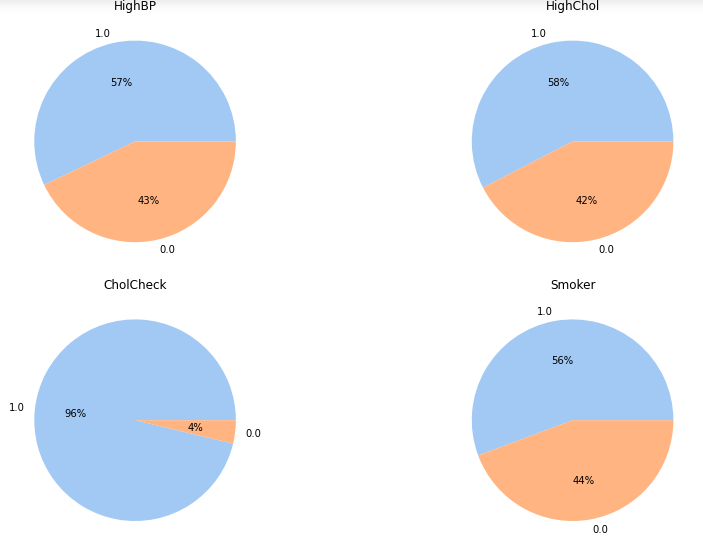
Видим, что 75-ый перцентиль у фичей с количеством ментальных и физических проблем равен 2 и 3 соответственно. Это значит, что только у 25% опрошенных наблюдалось большее количество проблем. Большая же часть людей сталкивалась с ними не часто.

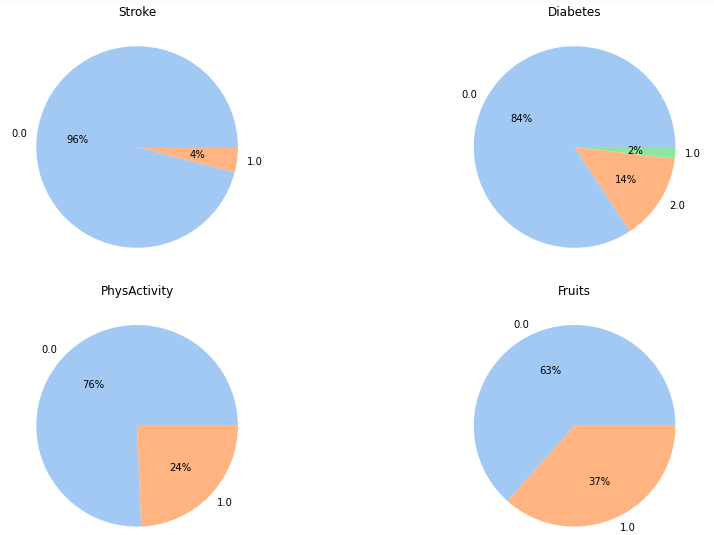
По ИМТ видим, что здесь, наоборот, больше часть людей (примерно 70%) имеет ИМТ ниже среднего. Высокие значения индекса (больше 30) встречаются редко - примерно в 25% случаев. Это хорошо видно по точечному графику.

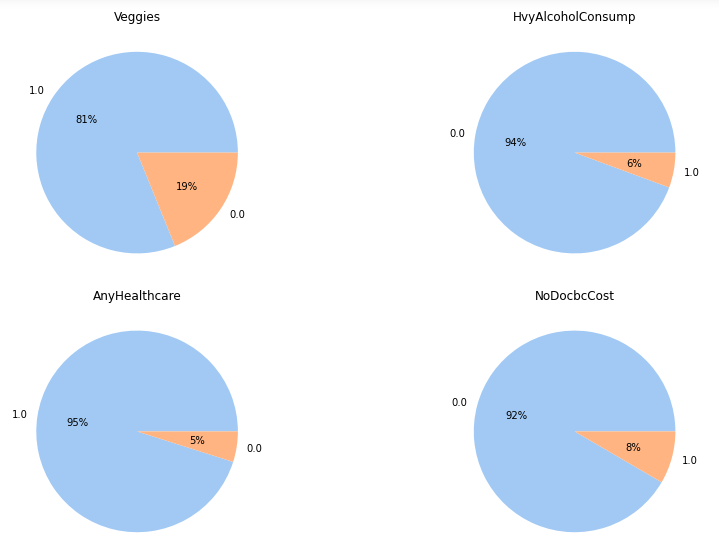
Распределение ИМТ похоже на нормальное.

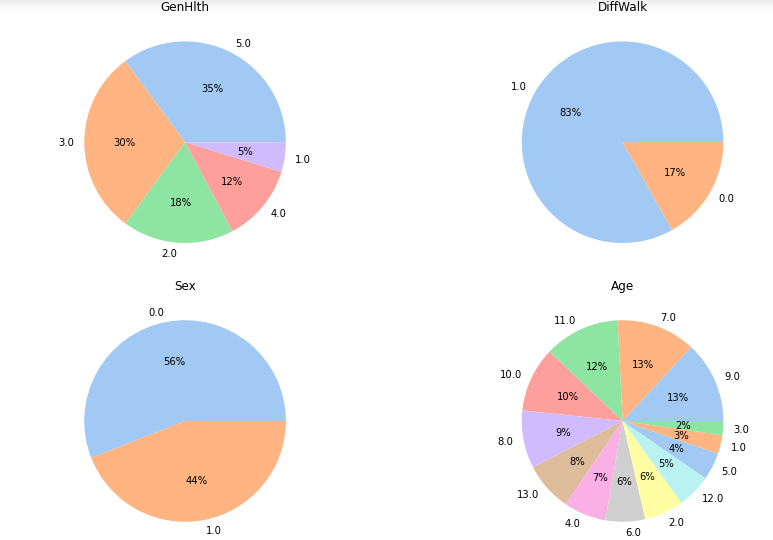
1. **Категориальные признаки**

Большая часть имеющихся признаков - категориальные. Посмотрим на их распределения.







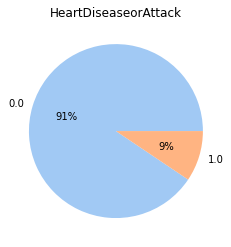


Можем заметить следующие интересные факты:

* 96% опрошенных следят за холестерином
* Люди едят овощи чаще, чем фрукты
* У большинства людей есть медицинская страховка и деньги на врача
* 48% пациентов оценивают свое состояние как хорошее или очень хорошее
* Среди опрошенных много людей с высшим образованием и с высоким доходом
* Большинство пациентов пожилые (50+ лет)

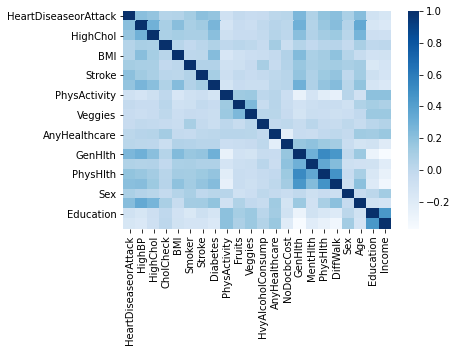
1. **Таргет**

Посмотрим на распределение таргета.



Классы очень несбалансированные. Имеем соотношение 90/10. Это нужно будет учитывать в дальнейшем.

Посмотрим на корреляционную матрицу. Матрица с численными значениями слишком большая, чтобы вставлять ее сюда. При необходимости ее можно посмотреть в ноутбуке.



* Количество ментальных проблем коррелирует с количеством физических проблем и с наличием трудностей при ходьбе (о\_о)
* Таргет не коррелирует с ИМТ (хотя у меня была гипотеза, что люди с высоким ИМТ более склонны к сердечным проблемам)
* Таргет коррелирует с повышенным давлением, наличием диабета и инсультов, что ожидаемо
* Количество ментальных и физических проблем имеет отрицательную корреляцию с доходом и уровнем образования. Богатые и образованные реже замечают у себя проблемы со здоровьем.
* Субъективная оценка здоровья коррелирует с наличием различных заболеваний. Чем хуже оценивает свое здоровье человек, тем более вероятно, что у него есть какие-либо заболевания. Это говорит о том, что люди оценивают свое общее состояние здоровья довольно правдиво.

1. **Вывод**

В данной лабораторной работе я провел полное исследование медицинского датасета. Я изучил все имеющиеся признаки, чтобы понять, с чем вообще имею дело. Попутно я находил различные инсайты в данных, которые, возможно, пригодятся в дальнейшем. Я попрактиковался в работе с таблицами и в визуализации.

Я убедился, что целевая переменная зависит от имеющихся признаков. Следовательно, у нас есть все шансы получить хорошую модель.