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: Якушкина Мария Константиновна

Группа: М80 – 407Б-19

Дата: 28.10.2022

Оценка: \_\_\_\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_

Москва, 2022

* **Постановка задачи**

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

* **Описание датасета**

Имеется датасет с информацией о пациентах США. Нужно выявить пациентов подверженных инсульту

Имеем следующие сведения об опрошенных людях:

gender - пол

age - возраст

hypertension - наличие гипертонии

heart\_disease - наличие заболеваний сердца

ever\_married - состоял ли в браке

work\_type - тип работы

Residence\_type - местность проживания

avg\_glucose\_level - средний уровень глюкозы

bmi - индекс массы тела

smoking\_status - употребление табака

**stroke -** Таргет, был ли инсульт у пациента

Ссылка на датасет: <https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset>

Датасет не содержит пропусков, можно приступать к анализу.

* **Количественные признаки**



Количественные признаки: возраст, среднее кровяное давление, ИМТ

Из графиков видно, что шанс получить инсульт зависит от возраста: до 40 лет инсульт пациентов является единичныи случаями

От среднего же уровня глюкозы и индекса массы тела закономерностей не видно

* **Категориальные признаки**

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



Из полученных графиков видно, что:

гипертонией страдают 10% отпрошенных

заболеваниями сердца 94%

в браке состояло 2/3

больше половины работают в частном бизнесе

распределение жилья в городской и сельской местности отличаются на 2%

* **Таргет**

Посмотрим на распределение таргета.



Классы очень несбалансированные. Имеем соотношение 95/15. Это нужно будет учитывать в дальнейшем.

Посмотрим на корреляционную матрицу.



Возраст сильно коррелирует с остальными параметрами, что логично. Из этих параметров больше всего коррелирует ИМТ

Таргет тоже довольно сильно коррелирует с возрастом, но не с ИМТ

Таргет также коррелирует от гипортенией и заболеваниями сердца. Гипортения и заболевания сердца тоже коррелируют между собой, что означает, что болеющим одной из болезней хорошо бы провериться на остальные

* **Вывод**

В данной лабораторной работе я провела полное исследование медицинского датасета. Также изучила все имеющиеся признаки, чтобы понять, с чем вообще имею дело. Попутно я находила различные инсайты в данных, которые, возможно, пригодятся в дальнейшем. Я попрактиковалась в работе с таблицами и в визуализации.

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