Университет ИТМО

Практическая работа №3

по дисциплине «Визуализация и моделирование»

Автор: Никулина Ксения Григорьевна
Поток: ВИМ 1.1
Группа: К3221
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

1 Heart Failure Prediction

Сердечно-сосудистые заболевания являются причиной смерти номер 1 во всем мире, забирая в среднем 17.9 миллионов жизней в год, что составляет 31% от всех смертей по всему миру. Остановка сердца - результат сердечно-сосудистых заболеваний и данный датасет содержит 12 показателей, которые можно использовать для предсказания смертности от остановки сердца.

Название столбца	Описание данных	Тип данных	Шкала
Возраст	Возраст пациента	Integer	Интерв
Анемия	Снижение гемоглобина	Binary	Номин
Креатинкиназа	Уровень фермента	Integer	Относит
Диабет	Наличие диабета	Binary	Номин
Фракция выброса	% крови при сокращении	Integer	Относит
Высокое кровяное давление	Гипертония	Binary	Номин
Тромбоциты	Тромбоциты в крови	Float	Номин
Креатинин	Уровень креатинина в крови	Float	Номин
Натрий	Уровень натрия в крови	Integer	Относит
Пол	Пол пациента	Binary	Номин
Курение	Курит ли пациент	Binary	Номин
Время	Кол-во дней наблюдения	Integer	Относит
Смерть	Умер ли пациент	Binary	Номин

Все данные таблицы представленны в числовом виде, значит такой проблемы как "текст неудобно использовать при построении не возникнет". Посмотрим на единственный столбец, который возможно потребует обработки - это столбец "возраст"

```
age_stat = {"min": df["age"].min(),
            "max": df["age"].max(),
            "mean": df["age"].mean(),
            "median": df["age"].median()
            }
age_stat
```

```
{'min': 40.0, 'max': 95.0, 'mean': 60.83389297658862, 'median': 60.0}
```

Среднее значение - шестьдесят лет является достаточной приемлимым значением при минимальном - 40 и максимальном - 95. Существует эвристика, по которой выбросами считаются значения, входящие в 5% крайних с обеих сторон процентов ранжированной выборки. Т.е. первые 5% и последние 5%. Посмотрим, какие значения попадают в эти диапазоны в этом датасете. Оставим значения только в этом промежутке, удалив остальные Т.к. пустых значений в датасете нет, столбцы содержат адекватные данные, опечаток не наблюдается, то можно считать на данный момент предобработку выполненной.

Гипотезы:

1. Количество смертей в зависимости от возраста растет линейно.
2. Чем моложе человек, тем меньше у него шанс умерет от сердечной недостаточности (рассеивающая)
3. Количество женщин, в более старшем возрасте (60+) под наблюдением превосходит мужчин (столбчатая)
4. Пациенты с наличием гипертонии умирают чаще от сердечной недостаточности, чем те, у которых ее нет
5. Наличие курения у людей 55+ лет повышает риск ссз