Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа № 1

по дисциплине «Методы машинного обучения»

Создание «истории о данных»

ИСПОЛНИТЕЛЬ:

студент ИУ5-25М
Мацнев А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

"__"_____2023 г.

Москва, 2023

Задание

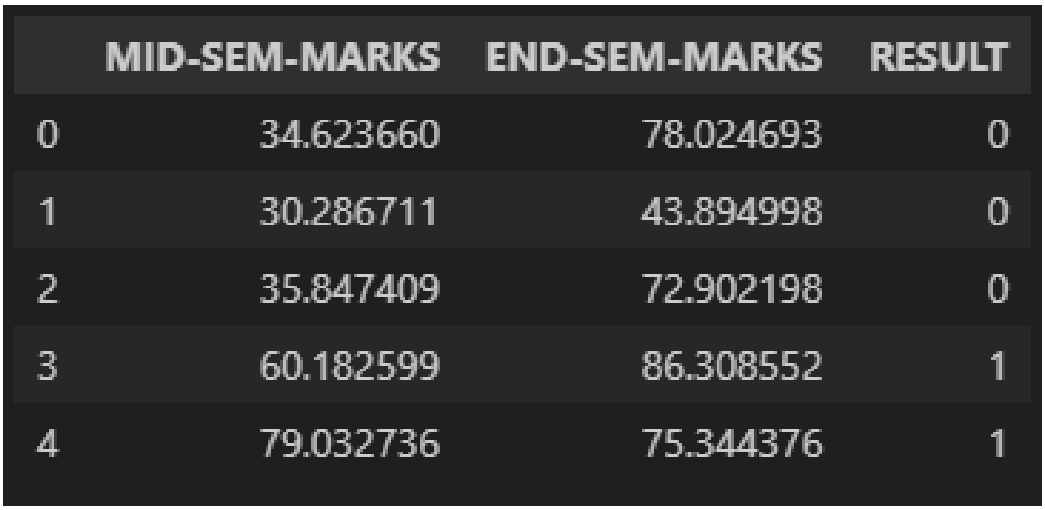
1. Выбрать набор данных (датасет)
2. Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

- История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
- На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
- Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
- Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
- История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Выполнение задания

1. Выбрать набор данных (датасет)

В качестве набора данных были выбраны данные об успеваемости 100 студентов на протяжении семестра и результат зачёта. Данные содержат результат двух тестов (от 0 до 100) и результат зачёта (0 или 1). Пример данных, содержащихся в наборе данных, представлен на рисунке 1.



	MID-SEM-MARKS	END-SEM-MARKS	RESULT
0	34.623660	78.024693	0
1	30.286711	43.894998	0
2	35.847409	72.902198	0
3	60.182599	86.308552	1
4	79.032736	75.344376	1

Рис. 1. Пример изучаемых данных.

2. Создать "историю о данных" в виде юпитер-ноутбука

В начале изучения набора данных была сформирована таблица, в которой отражены основные характеристики набора. Результаты представлены в таблице 1.

Таблица 1. Характеристика изучаемого набора данных

	MID-SEM-MARKS	**END-SEM-MARKS**	**RESULT**
count	100.000000	100.000000	100.000000
mean	65.644274	66.221998	0.600000
std	19.458222	18.582783	0.492366
min	30.058822	30.603263	0.000000
25%	50.919511	48.179205	0.000000
50%	67.032988	67.682381	1.000000
75%	80.212529	79.360605	1.000000
max	99.827858	98.869436	1.000000

Далее, в качестве эксперимента была создана тепловая карта распределения данных, она представлена на рисунке 2.

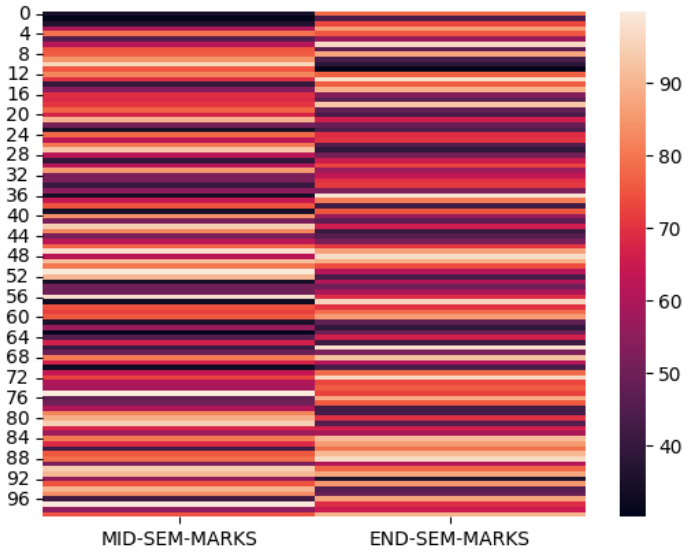


Рис. 2. Тепловая карта.

К сожалению, такой вид отображения данных не показал никаких зависимостей так как такой он не предназначен для анализа данных в таком виде.

Далее, была создана гистограмма, показывающая количество значений, попадающих в различные диапазоны шириной 10. Гистограмма представлена на рисунке 3.

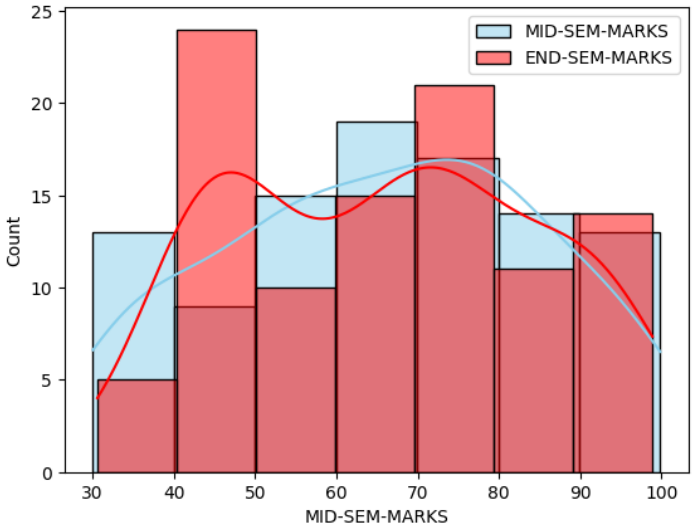


Рис. 3. Гистограмма распределения оценок.

На основании этой гистограммы можно сделать выводы о распределении оценок. Например, видно, что в середине семестра оценку в диапазоне от 60

до 70 получило большее число студентов. А в конце семестра студенты в основном получали оценку либо в диапазоне от 40 до 50, либо от 70 до 80.

Также, были созданы диаграммы типа «ящик с усами», данные диаграммы для двух величин представлены на рисунке 4.

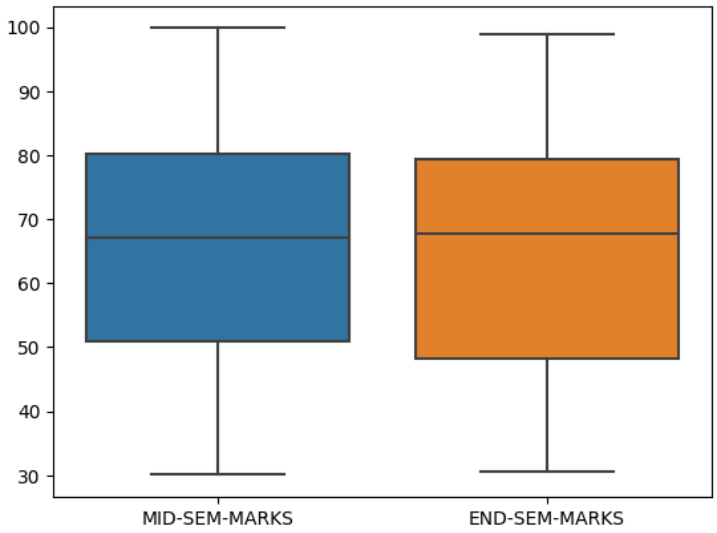


Рис. 4. Диаграмма «ящик с усами».

На этой диаграмме можно заметить, что с в течение семестра медиана почти не изменилась, однако разброс данных увеличился, хоть и незначительно. Это можно сопоставить с данными, отображёнными на предыдущей гистограмме и связать это с выделением двух групп студентов.

Наконец, была построена скрипичная диаграмма для двух величин и двух исходов зачёта. Набор этих диаграмм представлен на рисунках 5.1 и 5.2.

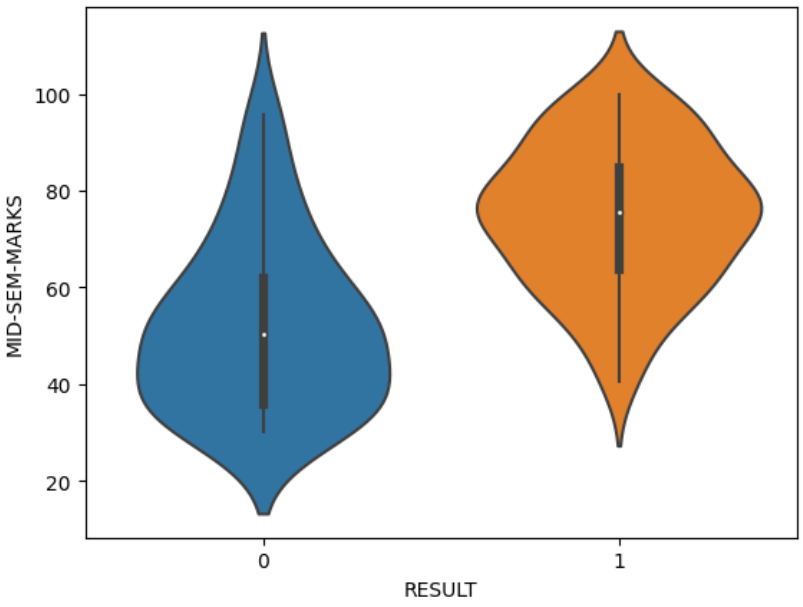


Рис. 5.1. Скрипичная диаграмма для оценок студентов в середине семестра

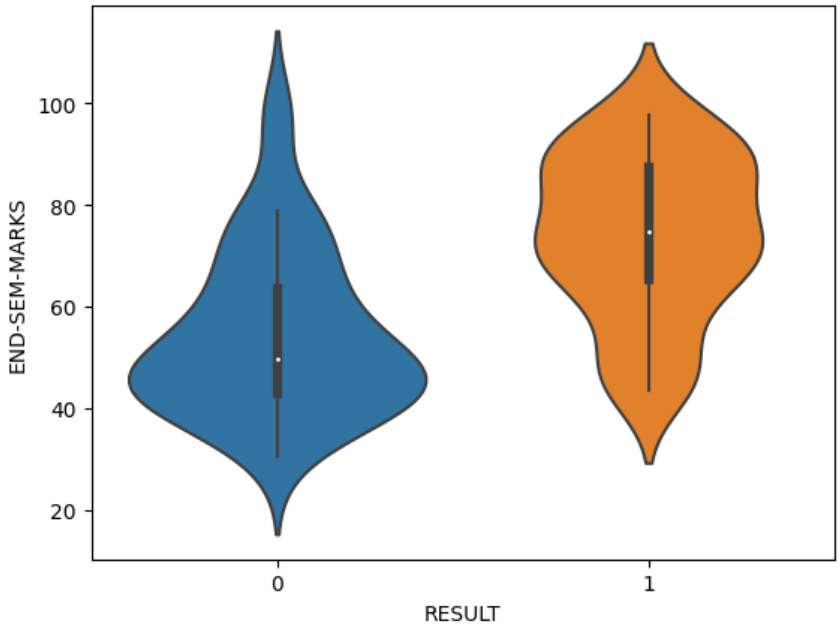


Рис. 5.2. Скрипичная диаграмма для оценок студентов в конце семестра

Данный вид диаграммы оказался самым информативным для анализа набора данных. На диаграмме видно, что тест по результатам теста в конце семестра большинство студентов, не сдавших впоследствии экзамен, получили около 50 баллов. При этом, студенты, сдавшие экзамен, получили около 70 баллов.

Кроме того, можно заметить, как изменялись оценки у двух групп студентов в течение семестра. Так, студенты, успешно сдавшие экзамен, улучшили результаты тестов, в то время как провалившие экзамен студенты, показали результаты хуже, чем в начале семестра.

Таким образом, можно предположить, что успешная сдача экзамена связана с успешным и своевременным освоением учебной программы. Хотя этот вывод и очевиден, он был получен из результатов анализа данных, что и являлось целью работы.

Список литературы

1. Датасеты для скачивания [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/datasets. – Дата доступа: 10.05.2023.
2. Сайт с методологией визуализации - data-to-viz [Электронный ресурс]. – Режим доступа: https://www.data-to-viz.com. – Дата доступа: 10.05.2023.
3. Частые ошибки при визуализации данных [Электронный ресурс]. – Режим доступа: https://www.data-to-viz.com/caveats.html. – Дата доступа: 10.05.2023.