**МИНОБРНАУКИ РОССИИ САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №2**

**по дисциплине «Машинное обучение»**

**Тема: «Метод кластеризации K-mean»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чесноков М. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Татчина Я. А.

Санкт-Петербург

2024

**Цель работы:** Познакомиться с методом кластеризации K-mean с помощью пакета sklearn.

**Краткое описание датасета**

Датасет о погоде. Датасет был взят с сайта [www.kaggle.com](http://www.kaggle.com/).

Почти все данные в нём числовые, только в weather содержатся строковые. В датасете представлены следующие атрибуты:

* data - дата измерений
* precipitation - выпавшие осадки в мм
* temp\_max - максимальная температура в этот день в градусах Цельсия
* temp\_min - минимальная температура в этот день в градусах Цельсия
* wind - скорость вертра в м/с
* weather - погодные условия. Принимает значения "drizzle" (морось), "rain" (дождь), "sun" (солнце), "snow" (снег), "fog" (туман)

**Добавление нового атрибута**

Добавим атрибут "Максимальная температура по Фаренгейту". Для этого значение атрибута temp\_max надо умножить на 1.8 и прибавить 32.

**"Причесать" датасет**

В полученном датасете отсутствуют пустые значения и дубликаты. Были обнаружены выбросы по атрибутам precipitation и wind. Удалим записи с выбросами по данным атрибутам. В итоге, было удалено 250 записей.

Для дальнейшей кластеризации нормализуем данные. Для этого воспользуемся функцией StandardScaler библиотеки sklearn.preprocessing.

**Построить графики зависимости одной переменной от другой**

Построим графики зависимостей атрибутов precipitation, temp\_max, temp\_min и wind друг от друга.

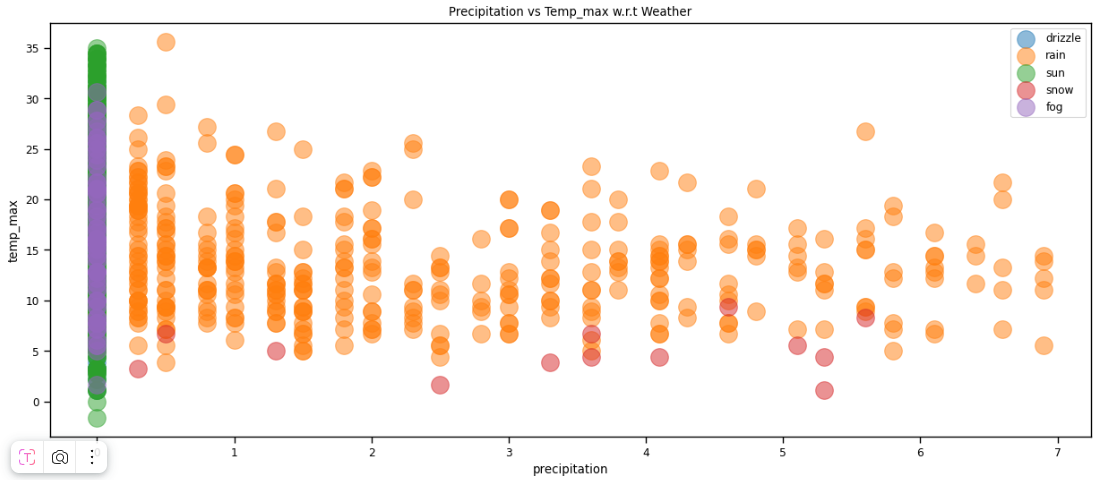


Рис. 1: График зависимости precipitation от temp\_max

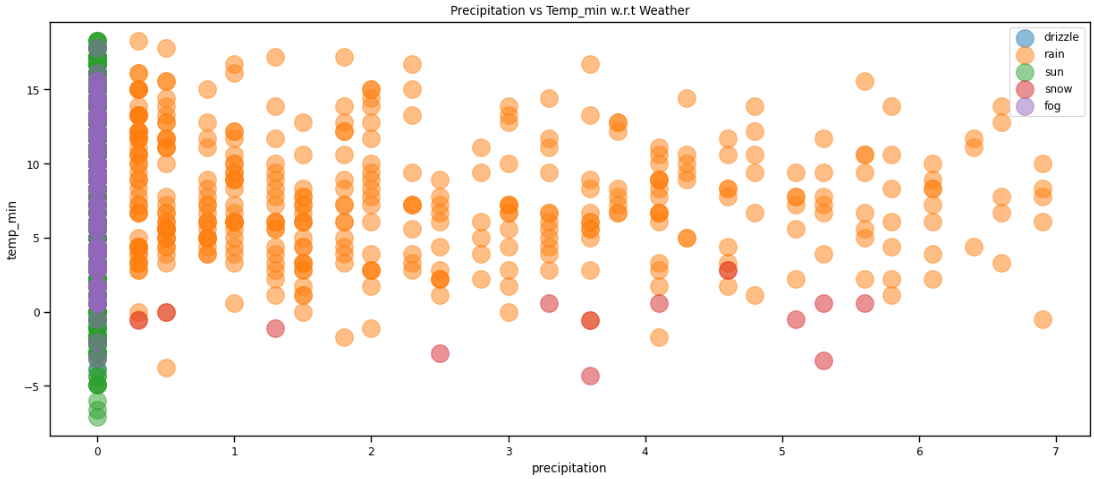


Рис. 2: График зависимости precipitation от temp\_min

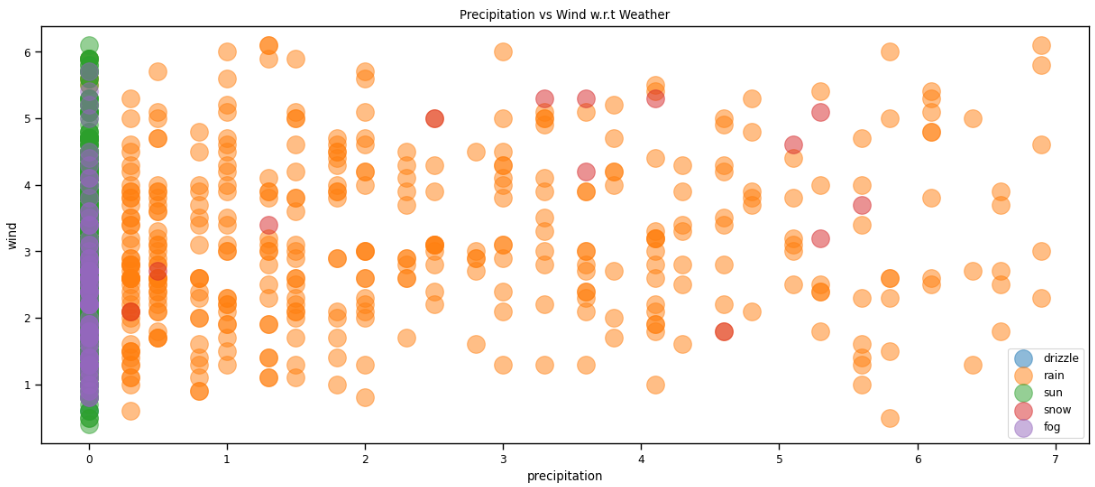


Рис. 3: График зависимости precipitation от wind

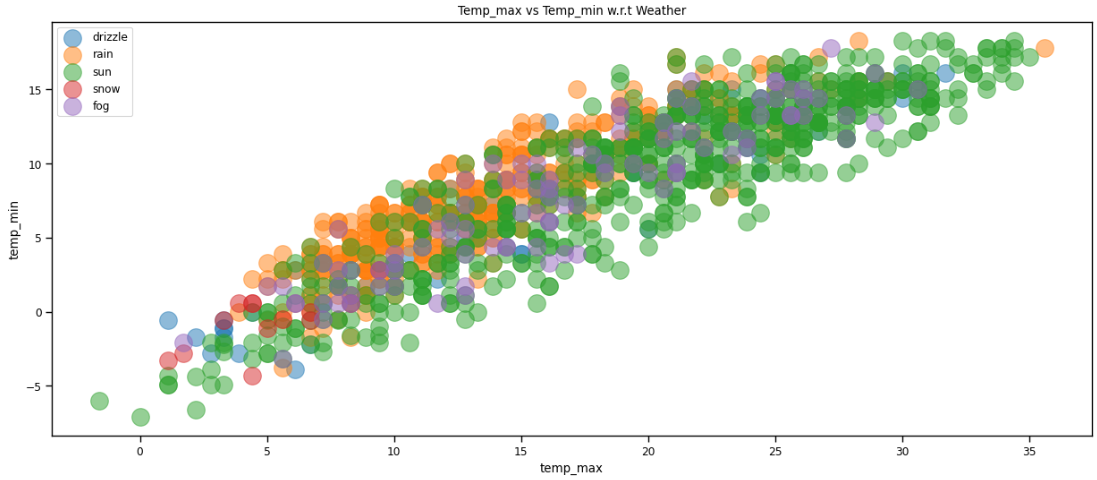


Рис. 4: График зависимости temp\_max от temp\_min

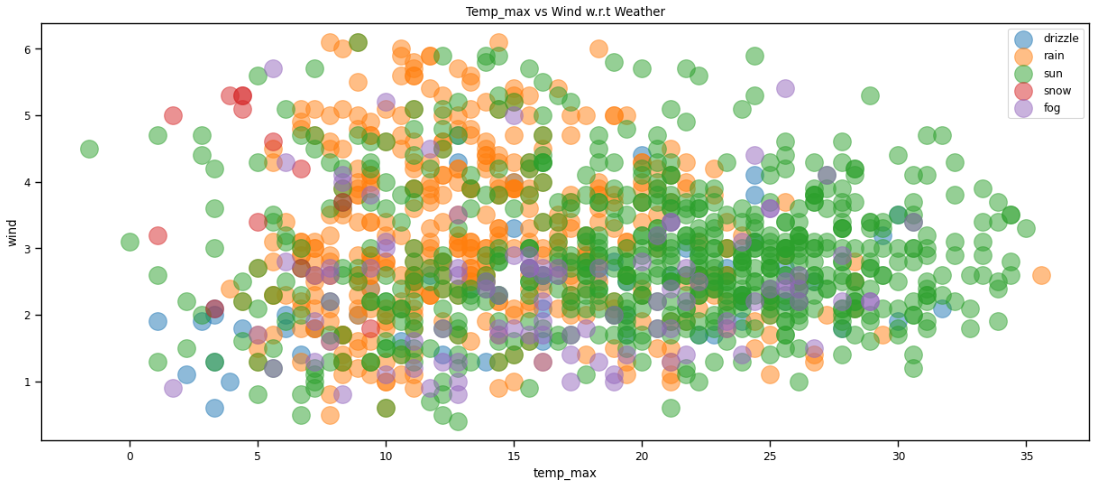


Рис. 5: График зависимости temp\_max от wind

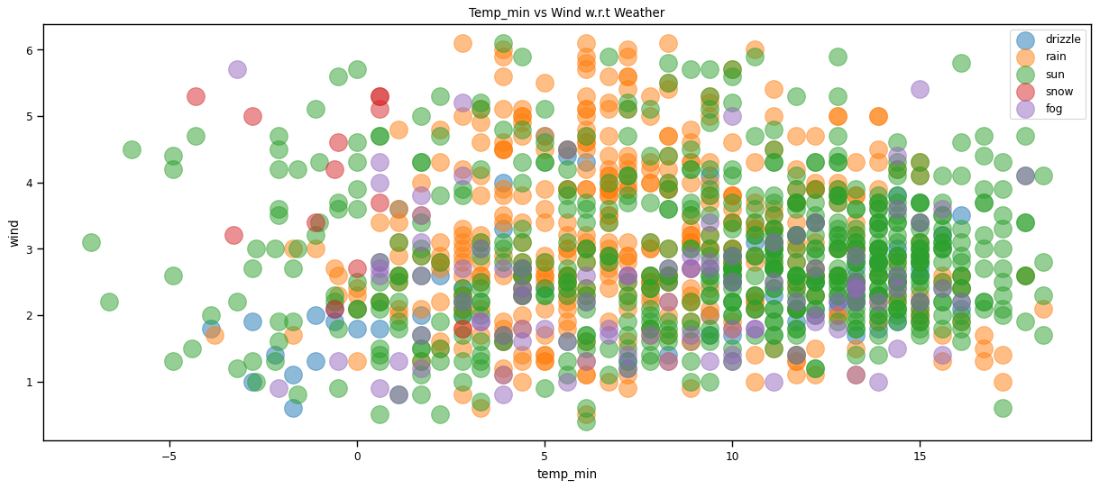


Рис. 6: График зависимости temp\_min от wind

Можно попробовать разделить набор данных по кластерам по погодным условиям. По графикам видно, что дождь чаще всего был при высоком значении выпавших осадков, а солнце при высоких значениях минимальной и максимальной температуры.

**Применить метод KMeans**

Алгоритм KMeans группирует данные, пытаясь разделить выборки на n групп с одинаковой дисперсией, минимизируя критерий, известный как инерция или сумма квадратов внутри кластера (см. ниже). Этот алгоритм требует указания количества кластеров. Он хорошо масштабируется для большого количества данных и используется в самых разных областях применения.

Для определения количества кластеров воспользуемся методом локтя. Для этого посчитаем значения инерции алгоритма при различных количествах кластеров и выберем значение инерции, после которого инерция перестаёт существенно уменьшаться.

Алгоритм KMeans был применён ко всем парам атрибутов и выведены графики зависимостей атрибутов с разбиением на кластеры (представлены в файле с программой).

**Вывод**

В ходе данной работы я добавил новый атрибут в датасет, провёл первичную обработку: удалил выбросы по атрибутам precipitation и wind, пропущенных значений и дубликатов не было обнаружено; и нормализировал данные в датасете.

Также я построил графики зависимостей атрибутов от погодных условий, по которым видно, что дождь чаще всего был при высоком значении выпавших осадков, а солнце при высоких значениях минимальной и максимальной температуры.

Я изучил метод кластеризации KMeans, с помощью которого попробовал выделить кластеры для данного датасета.