# МИНОБРНАУКИ РОССИИ

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

# «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

**Кафедра ИС**

# ОТЧЕТ

**по лабораторной работе №1**

# по дисциплине «Машинное обучение» Тема: Исследование набора данных

|  |  |
| --- | --- |
| Студент гр. 2373 | Гермаш А. В. |
| Преподаватель | Татчина Я.А. |

Санкт-Петербург 2024

Цель работы: исследовать алгоритмы классификации и кластеризации на выбранном наборе данных.

# Краткое описание набора данных

Данные о погоде в Лондоне с 1979 по 2021. Датасет был взят с сайта [www.kagle.com](http://www.kagle.com/) по [ссылке](https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data). Все данные реальны, они были загружены с веб-сайта European Climate Assessment and Dataset в 2022 году. Все данные в датасете - числовые. В датасете представлены следующие атрибуты:

* + date - дата измерения в формате ГГГГММДД.
  + cloud\_cover - облачность в октантах (0-8).
  + sunshine - время, когда светило солнце в часах (0-24).
  + global\_radiation - энергия радиации в Вт/м2.
  + max\_temp - максимальная температура за сутки в градусах Цельсия (°C).
  + mean\_temp - средняя температура за сутки в градусах Цельсия (°C).
  + min\_temp - минимальная температура за сутки в градусах Цельсия (°C).
  + precipitation - атмосферные осадки в мм.
  + pressure - давление в Паскалях.
  + snow\_depth - глубина снежного покрова в см.

# Определение параметров

1. Среднее значение, СКО

Среднее значение и СКО атрибутов были определены с помощью функций библиотеки numpy 'np.mean' и ‘np.std’ соответственно.

* + cloud\_cover: среднее = 5.27, СКО = 2.07
  + sunshine: среднее = 4.35, СКО = 4.03
  + global\_radiation: среднее = 118.76, СКО = 88.9
  + max\_temp: среднее = 15.39, СКО = 6.55
  + mean\_temp: среднее = 11.48, СКО = 5.73
  + min\_temp: среднее = 7.56, СКО = 5.33
  + precipitation: среднее = 1.67, СКО = 7.34
  + pressure: среднее = 101536.61, СКО = 1049.72
  + snow\_depth: среднее = 0.04, СКО = 0.55

1. Построить гистограммы распределения значений, определить есть ли выбросы

Поиск выбросов был осуществлен с помощью следующего алгоритма:

* + Найти Q1 и Q3 (1-ый и 3-ый квартиль)
  + Найти межквартильный размах IQR
  + Проверить, какие наблюдения вышли за границы [Q1 – 1.5\*IQR; Q3 + 1.5\*IQR]

Наличие выбросов:

* + cloud\_cover: нет
  + sunshine: нет
  + global\_radiation: нет
  + max\_temp: да
  + mean\_temp: да
  + min\_temp: да
  + precipitation: да
  + pressure: да
  + snow\_depth: да

Гистограммы распределений представлены на рис. 2.1 – 2.9

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Рисунок 2.1 – Гистограмма облачности



Рисунок 2.2 – Гистограмма времени свечения солнца

Изображение выглядит как График, линия, диаграмма, снимок экрана

Автоматически созданное описание

Рисунок 2.3 – Гистограмма энергии радиации

Изображение выглядит как диаграмма, График, снимок экрана, линия

Автоматически созданное описание

Рисунок 2.4 – Гистограмма максимальной температуры

Изображение выглядит как График, диаграмма, снимок экрана, линия

Автоматически созданное описание

Рисунок 2.5 – Гистограмма средней температуры

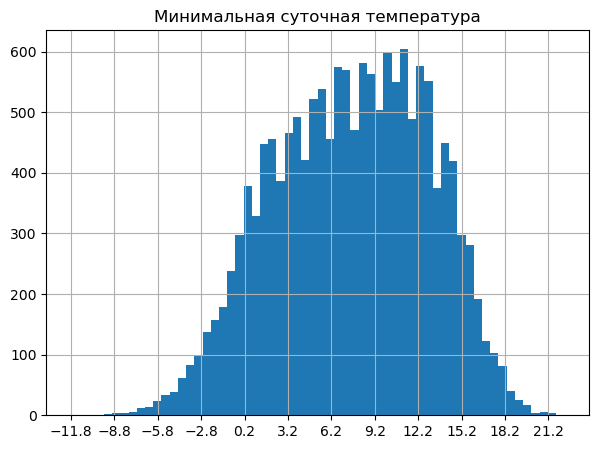


Рисунок 2.6 – Гистограмма минимальной температуры

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 2.7 – Гистограмма атмосферных осадков

Изображение выглядит как диаграмма, График, снимок экрана, линия

Автоматически созданное описание

Рисунок 2.8 – Гистограмма атмосферного давления

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Рисунок 2.9 – Гистограмма глубины снежного покрова

1. Определить, есть ли пропущенные значения

Наличие пропущенные значений определялось с помощью функции df[‘Название’].isna().sum()

Количество пропущенных значений:

* + cloud\_cover: 19
  + sunshine: нет
  + global\_radiation: 19
  + max\_temp: 6
  + mean\_temp: 36
  + min\_temp: 2
  + precipitation: 6
  + pressure: 4
  + snow\_depth: 1441

1. Предложить вариант обработки пропущенных значений
   * cloud\_cover: между cloud\_cover и precipitation есть корреляция. Для пропущенных значений возьмём precipitation / 4, но не более 8
   * sunshine: нет
   * global\_radiation: видна сильная корреляция со значениями sunshine. Для пропущенных значений возьмём sunshine \* 25
   * max\_temp: для пропущенных значений вычислим по формуле: 2 \* mean\_temp - min\_temp
   * mean\_temp: для пропущенных вычислим по формуле: (max\_temp + min\_temp) / 2
   * min\_temp: для пропущенных значений вычислим по формуле: 2 \* mean\_temp - max\_temp
   * precipitation: явных зависимостей нет, лучше взять среднее значение двух ближайших дней
   * pressure: с увеличением температуры и солнечности повышается и давление, однако нет сильной корреляции. Лучшим способом будет взять среднее значение двух соседних дней.
   * snow\_depth: Если значения пропущены в летнеий период (условно с апреля по октябрь), то можем принять глубину снега равной 0. Если значения пропущены в зимний период, то стоит взять среднее значение соседних измерений. Если же значения пропущены в зимний период и пропущено много значений подряд (более 10), то данные не получится обработать

# Определение корреляции

# Матрица корреляции построена при помощи функции scatter\_matrix из библиотеки pandas.plotting

Изображение выглядит как текст, шаблон, линия, снимок экрана

Автоматически созданное описание

Рис. 3.1 – матрица графиков корреляции

Характер корреляции значений:

* + cloud\_cover: прослеживается только линейная корреляция с precipitation
  + sunshine: линейная сильная корреляция с global\_radiation, нелинейная корреляция с precipitation. Очень слабая корреляция с min\_temp, mean\_temp, max\_temp
  + global\_radiation: видна сильная корреляция со значениями sunshine, нелинейная корреляция с precipitation, немного прослеживается корреляция с min\_temp, max\_temp, mean\_temp
  + max\_temp: сильная линейная корреляция с min\_temp и mean\_temp, немного прослеживается корреляция с global\_radiation
  + mean\_temp: сильная линейная корреляция с min\_temp и max\_temp, немного прослеживается корреляция с global\_radiation
  + min\_temp: сильная линейная корреляция с max\_temp и mean\_temp, немного прослеживается корреляция с global\_radiation
  + precipitation: нелинейная корреляция с sunshine, snow\_depth и global\_radiation
  + pressure: нет явных корреляций
  + snow\_depth: нелинейно коррелирует с precipitation

Таким образом, в большинстве атрибутов можно увидеть корреляцию с температурой воздуха.