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

ОТЧЕТ

по лабораторной работе № 1
по дисциплине «Машинное обучение»
Тема: «Подготовка и нормализация данных»

Студент гр. 0308 ___ Сабурова Е.А.

Преподаватель ___ Гатауллин Р.И.

Санкт-Петербург
2023

Оглавление

1. Цель работы

Получить и закрепить навыки предобработки данных для дальнейшего применения методов машинного обучения для решения задач.

2. Задание

Выбор набора данных на сайте Kaggle.com.

В ходе самой лабораторной работы должны быть выполнены следующие этапы:

1. Визуализация значимых признаков (диаграммы рассеяния, ящик с усами, гистограммы)

2. Очистка данных (удаление пропусков, нормализация, удаление дубликатов)

3. Корреляция данных (матрица корреляций)

3. Краткая теоретическая информация по теме работы

Задача машинного обучения сводится к получению набора выборок данных и к попыткам предсказать свойства неизвестных данных. Если каждый набор данных — это не одиночное число, а например, многомерная сущность, то он должен иметь несколько признаков или фич.

Процесс обучения состоит в подгонке модели к обучающим данным (минимизация ошибок между предсказаниями модели и правильными ответами).

Выборка — это анализ подмножества данных с целью выявить значимую информацию в большем наборе данных.

Рисунок 1 – типы переменных

Мода — значение во множестве наблюдений, которое встречается наиболее

часто.

Медиана (от лат. mediāna «середина») или серединное значение набора чисел — число, которое находится в середине этого набора, если его упорядочить по возрастанию, то есть такое число, что половина из элементов набора не меньше него, а другая половина не больше.

Среднее значение (mean или среднее арифметическое) — это сумма всех значений в распределении деленное на их количество.

Размах (Range) - разность максимального и минимального значения.

Дисперсия - средний квадрат отклонений индивидуальных значений признака от их средней величины.

Квартили — три точки (значения признака), которые делят упорядоченное множество данных на 4 равные части.

Квартили распределения — это квантили, кратные 25%, то есть соответствующие 25%, 50% и 75%. Их ещё иногда называют соответственно «первый», «второй» и «третий» либо "нижний", "средний" и "верхний".

Второй квартиль является самостоятельной полезной статистической величиной, так как показывает, что 50% наблюдений в выборке лежит ниже данного числа, а остальные — соответственно выше, то есть он фактически

делит выборку пополам. Чаще в различной литературе его можно встретить под названием «медиана».

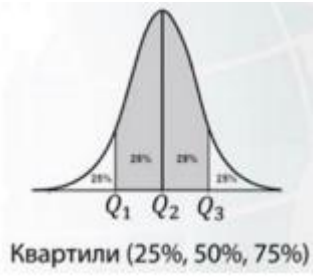

Рисунок 2 – квартили

Диаграмма рассеяния – математическая диаграмма, изображающая значения двух переменных в виде точек на декартовой плоскости.

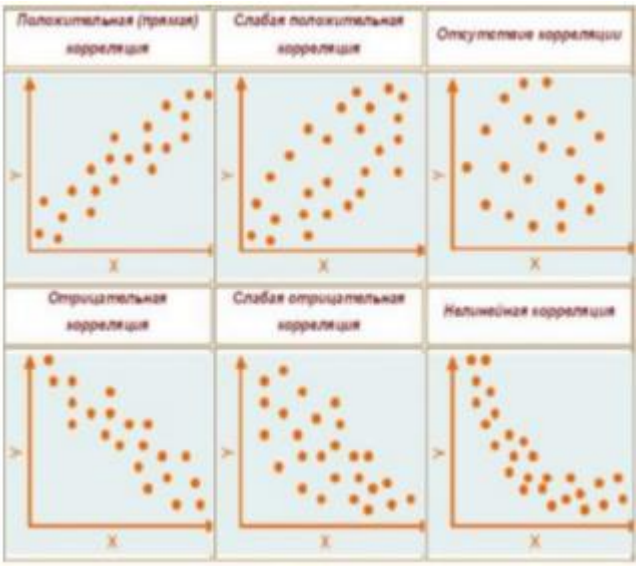

Рисунок 3 – диаграммы рассеяния различного вида и типы корреляций, которые они показывают

Ящик с усами — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Рисунок 4 – ящик с усами и плотность распределения, соответствующая ему

В широком смысле стандартизация/нормализация данных представляет собой этап их предобработки с целью приведения к определённому формату и представлению, которые обеспечивают их корректное применение в многомерном анализе, совместных исследованиях, сложных технологиях аналитической обработки.

- Стандартизация: среднее значение 0, стандартное отклонение 1, нет верхней/нижней границы для максимальных/минимальных значений // чтобы узнать, как значения в выборке отклоняются от среднего •
- Нормализация: все значения от 0 до 1 // чтобы построить диапазон для

различных переменных, имеющих разные масштабы

Гистограмма – это способ представления табличных данных в графическом виде – в виде столбчатой диаграммы.

Корреляция или корреляционная зависимость – это статистическая взаимосвязь двух или более случайных величин, при этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин

4. Ход работы

Для применения различных инструментов, которые помогают работать с данными, подключим необходимые библиотеки с помощью нескольких import'ов, которые представлены на рисунке 5.

```
# Read the dataset
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
```

Рисунок 5 – библиотеки, подключенные для работы с данными

Чтение выбранного датасета осуществляется с помощью команды pandas.read_csv().

```
df = pd.read_csv('/content/IndianWeatherRepository.csv')
```

Рисунок 6 – чтение датасета

Выведем основную информацию о датасете, включая имена столбцов, количество ненулевых записей и типов данных в столбцах.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18055 entries, 0 to 18054
Data columns (total 19 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   location_name                 17960 non-null  object 
 1   region                        17960 non-null  object 
 2   latitude                      17955 non-null  float64
 3   longitude                     17973 non-null  float64
 4   temperature_celsius           17968 non-null  float64
 5   condition_text                17962 non-null  object 
 6   wind_kph                      17956 non-null  float64
 7   wind_degree                   17944 non-null  float64
 8   pressure_mb                   17955 non-null  float64
 9   precip_mm                     17969 non-null  float64
 10  humidity                      17958 non-null  float64
 11  cloud                         17948 non-null  float64
 12  feels_like_celsius            17968 non-null  float64
 13  visibility_km                 17964 non-null  float64
 14  uv_index                      17969 non-null  float64
 15  gust_kph                      17960 non-null  float64
 16  air_quality_Carbon_Monoxide   17972 non-null  float64
 17  air_quality_Ozone             17965 non-null  float64
 18  moon_illumination             17943 non-null  float64
dtypes: float64(16), object(3)
memory usage: 2.6+ MB
```

Рисунок 7 – результат работы функции info()

Удалим строки, в которых есть пустые ячейки с помощью функции dropna()

Удаление строк, где есть пустые поля

```
[34] df = df.dropna(axis=0)
```

Проверка на наличие пустых полей после их удаления

```
df.isnull().sum()
```

```
location_name                 0
region                        0
latitude                      0
longitude                     0
temperature_celsius           0
condition_text                0
wind_kph                      0
wind_degree                   0
pressure_mb                   0
precip_mm                     0
humidity                      0
cloud                         0
feels_like_celsius            0
visibility_km                 0
uv_index                      0
gust_kph                      0
air_quality_Carbon_Monoxide   0
air_quality_Ozone             0
moon_illumination             0
dtype: int64
```

Рисунок 8 – проверка работы функции dropna(), удаляющей строки, в которых есть нулевые значения

Для построения диаграмм рассеяния и гистограмм была использована функция scatter_matrix(). Чтобы на главной диагонали наблюдать гистограммы, в функции необходимо ввести параметр diagonal='hist'. Если задать значение 'kde', на главной диагонали будут представлены графики плотности распределения.

```
pd.plotting.scatter_matrix(df1, alpha = 0.2, figsize=(15, 10), diagonal='hist')
```

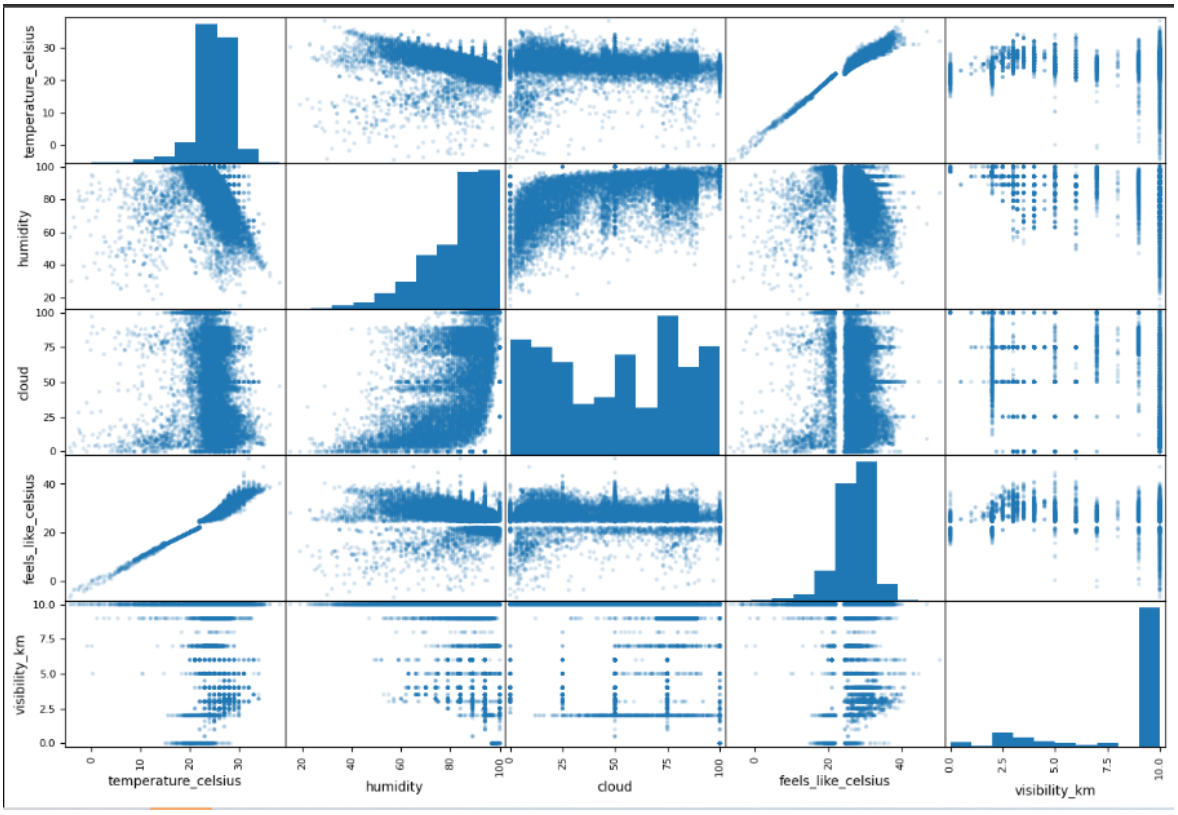

Рисунок 9 – матрица с диаграммами рассеивания и гистограммами

Построим ящики с усами для оценки скорости ветра в км/ч с помощью функции boxplot()

```
sns.boxplot(y = df['wind_kph'])
```

```
<Axes: ylabel='wind_kph'>
```

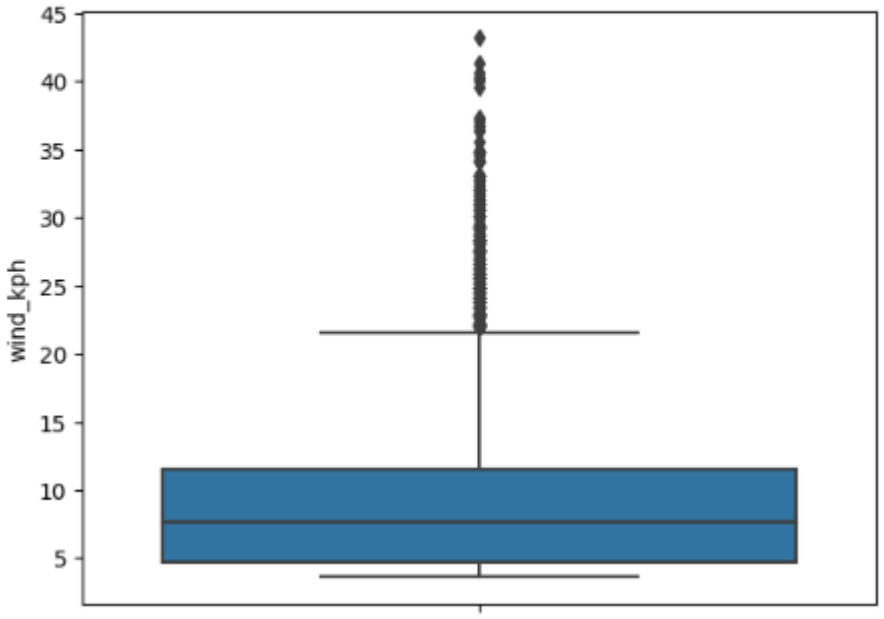

Рисунок 10 – ящик с усами, оценивающий скорость ветра в км/ч

На рисунке 10 можно увидеть выбросы, медианы, максимумы, минимумы, а также процентили.

Была произведена нормализация данных

```
df_n = df.drop(columns = ['location_name', 'region', 'condition_text'])
#normalize values in every column
df_norm = (df_n-df_n.min ())/ (df_n.max () - df_n.min ())

#view normalized DataFrame
df_norm
```

Рисунок 11 – код для нормализации данных

	latitude	longitude	temperature_celsius	wind_kph	wind_degree	pressure_mb	precip_mm	humidity	cloud	feels_like_celsius	visibility_km	uv_index	gust_kph
0	0.622499	0.326127	0.747073	0.426768	0.779944	0.458333	0.000000	0.611765	0.26	0.636998	1.00	0.750	0.332402
1	0.575689	0.329109	0.747073	0.300505	0.796657	0.458333	0.000000	0.647059	0.19	0.642234	1.00	0.750	0.251397
2	0.528124	0.371226	0.718970	0.373737	0.880223	0.500000	0.000000	0.647059	0.51	0.610820	1.00	0.750	0.296089
3	0.520196	0.333955	0.702576	0.335859	0.824513	0.500000	0.000000	0.717647	0.65	0.600349	1.00	0.625	0.291899
4	0.553794	0.326127	0.740047	0.318182	0.760446	0.500000	0.000000	0.694118	0.82	0.640489	1.00	0.625	0.261173
...	...	...	...	...	...	...	...	...	...	...	...	...	...
18050	0.785202	0.319046	0.735363	0.063131	1.000000	0.416667	0.000000	0.694118	0.00	0.605585	0.32	0.000	0.053073
18051	0.599849	0.879985	0.669789	0.000000	0.075209	0.416667	0.000000	0.988235	1.00	0.593368	0.00	0.000	0.029330
18052	0.183088	0.386135	0.690867	0.227273	0.768802	0.416667	0.000000	0.600000	0.31	0.582897	1.00	0.000	0.301676
18053	0.033598	0.310473	0.620609	0.191919	0.738162	0.583333	0.001142	0.905882	0.84	0.547993	1.00	0.000	0.248603
18054	0.522839	0.193067	0.672131	0.017677	0.083565	0.375000	0.000000	0.517647	0.27	0.567190	1.00	0.000	0.127095

Рисунок 12 – результат нормализации данных

Можно заметить, что все значения после нормализации лежат в диапазоне [0; 1].

Для создания матрицы корреляций воспользуемся функцией

```
correlations = df.corr()

plt.figure(figsize=(12, 8))
sns.set(font_scale=0.8)
sns.heatmap(correlations, cmap='coolwarm', square=True, annot=True)
plt.title('Correlation Matrix', fontsize=15)

plt.show()
```

Рисунок 13 – код для матрицы корреляций

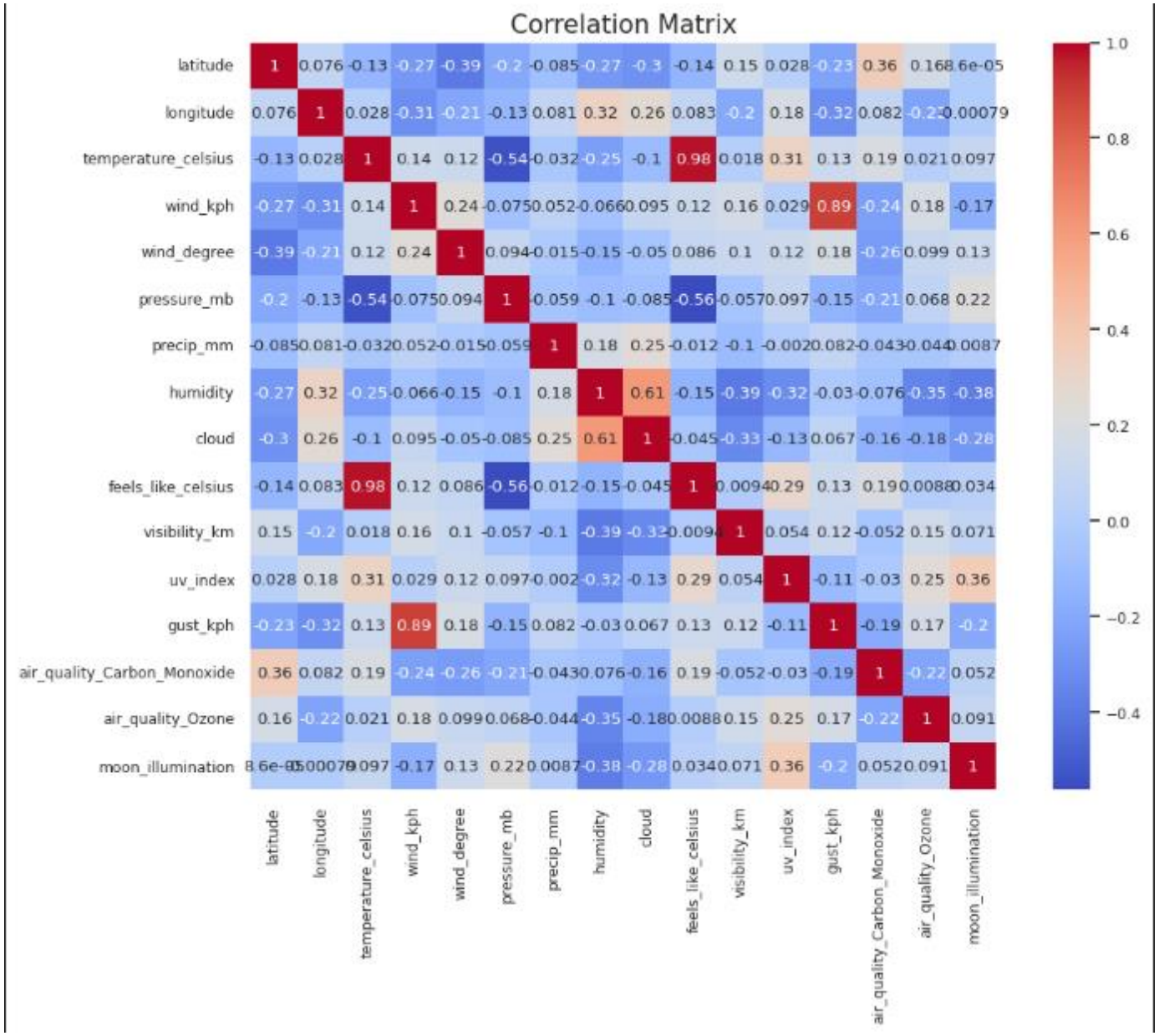

Рисунок 14 – матрица корреляций

По матрице корреляций видно, что линейной зависимости разных столбцов друг от друга нет, следовательно, нет необходимости ничего удалять.

5. Выводы

В ходе лабораторной работы были получены навыки предобработки данных для дальнейшего применения методов машинного обучения для решения задач. Были построены диаграммы рассеяния, ящик с усами и гистограммы для каждого значимого столбца.

Также была произведена очистка данных, которая включала в себя удаление пропусков и нормализацию. Была построена матрица корреляций для определения связей между значениями столбцов датасета.