SCC451 Coursework Report

Nhan Nguyen
School of Computing and Communications
SCC451 Machine Learning

I. Task 1: Basel Climate Dataset

The Basel Climate Dataset (ClimateDataBasel.csv) is a subset of publicly available weather data from Basel, Switzerland, obtained from [1]. It contains 1,763 records with 18 features collected during the summer and winter seasons from 2010 to 2019.

As the dataset lacks column headers, assigning appropriate names to each feature is necessary to ensure readability, consistency, and efficient data handling during analysis. The complete list of column names and their meanings is provided in Table II .

A. Preprocessing

Preprocessing is an important step ensure our data is high quality and reliable for analysis. Our process for this step is as follow.

1) Data cleaning: handling missing data, identifying and treating outliers, resolving data inconsistencies and removing duplicates.
2) Feature scaling.
3) Feature selection / extraction.

a) *Data Cleaning:*

First, import data file as DataFrame using pandas package. We named it basel_df.

Handling missing data By using isna() function of Pandas. We can see that the dataset has no missing value. Code for missing value detection is in Fig. 1 (Appendix A).

Identifying and treating outliers / noise Outliers are data points that deviate significantly from the overall pattern of the dataset. They can take unusually high or low values compared to the majority of observations, potentially indicating measurement errors, or rare but valid phenomena [2]. In the following step, we will detect potential anomalies across all features and decide on appropriate treatments, such as removal, transformation, or imputation, depending on their impact on the dataset.

There are some techniques to detect anomalies in data. We introduce here their advantages and disadvantages. a

TABLE I
Anomaly detection methods

Method	Advantage	Disadvantages
Z-score (Standard Deviation Method)	Simple and fast to compute. Works well for approximately normal distributions. Easy to interpret (values |z| > 3 often flagged).	Sensitive to outliers and non-Gaussian data. Not reliable for skewed or heavy-tailed features.
IQR Method	Simple and interpretable. Robust to outliers; no distributional assumptions. Suitable for skewed climate data (e.g., precipitation).	Univariate only; ignores multivariate interactions. May flag valid extremes in naturally variable data.
Isolation Forest, "Randomly partition data; anomalies have shorter average path length."	Model-free and scalable to large datasets. Handles non-linear, high-dimensional data. Works well for mixed seasonal patterns and irregular weather events.	Randomness may cause variability between runs. Requires hyperparameter tuning for small datasets. Less interpretable than statistical methods.

References

[1] "Weather Galgate." [Online]. Available: https://www.meteoblue.com/en/weather/week/galgate_united-kingdom_2648924
[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

Appendix A: Code

```
[5]
✓ 0s
# Handling missing data
columns = basel_df.columns

missing_summary = pd.DataFrame({
    "Missing Values": basel_df.isna().sum(),
    "Percentage": basel_df.isna().mean() * 100
}).sort_values(by="Missing Values")

missing_summary
```

	Missing Values	Percentage
temp_min_c	0	0.0
temp_max_c	0	0.0
temp_mean_c	0	0.0
rh_min_pct	0	0.0
rh_max_pct	0	0.0
rh_mean_pct	0	0.0
slp_min_hpa	0	0.0
slp_max_hpa	0	0.0
slp_mean_hpa	0	0.0
precip_mm	0	0.0
snow_cm	0	0.0
sunshine_min	0	0.0
gust_min_kmh	0	0.0
gust_max_kmh	0	0.0
gust_mean_kmh	0	0.0
wind_min_kmh	0	0.0
wind_max_kmh	0	0.0
wind_mean_kmh	0	0.0

Fig. 1. Code and result of detecting missing value.

Appendix B: Table

TABLE II
TABLE COLUMNS NAME AND THEIR MEANING

Column Name	Description	Unit
temp_min_c	Minimum daily temperature	°C
temp_max_c	Maximum daily temperature	°C
temp_mean_c	Mean daily temperature	°C
rh_min_pct	Minimum daily relative humidity	%
rh_max_pct	Maximum daily relative humidity	%
rh_mean_pct	Mean daily relative humidity	%
slp_min_hpa	Minimum daily sea level pressure	hPa
slp_max_hpa	Maximum daily sea level pressure	hPa
slp_mean_hpa	Mean daily sea level pressure	hPa
precip_mm	Total daily precipitation	mm
snow_cm	Total daily snowfall	cm
sunshine_min	Total sunshine duration per day	min
gust_min_kmh	Minimum daily wind gust speed	km/h
gust_max_kmh	Maximum daily wind gust speed	km/h
gust_mean_kmh	Mean daily wind gust speed	km/h
wind_min_kmh	Minimum daily wind speed	km/h
wind_max_kmh	Maximum daily wind speed	km/h
wind_mean_kmh	Mean daily wind speed	km/h