Olufemi Ige (01/21/2026)

Achievement 2: Unsupervised Learning Algorithms: ClimateWins Weather Analysis

Summary Report - OSLO vs VALENTIA (2020)

Overview

ClimateWins wants to use unsupervised learning (hierarchical clustering) to see whether weather stations form meaningful groups (i.e., whether any station-to-station connections" stand out) and to explore patterns in the 2020 weather data.

Methods

1. Data Preparation:

I used the 2020 slice of the climate dataset and applied feature scaling (StandardScaler) so that different weather measurements contribute fairly to distance calculations in clustering. The full climate dataset shown in the notebook is (22,950 rows × 170 columns), so narrowing to 2020 and scaling is a practical step before clustering.

2. Clustering Algorithm

I compared hierarchical clustering dendrograms across multiple linkage strategies: Single linkage (tends to "chain" points), Complete linkage, Average linkage, and Ward linkage (minimizes within-cluster variance; often produces compact clusters).

3. Dimensionality Reduction (PCA)

PCA (Dimensionality Reduction): Because the dataset is large, and computing power can be limited, PCA is used to reduce the number of features while keeping most of the important variation in the data. PCA is used specifically to shrink the dataset before clustering, and 10 components were selected based on the elbow point in the cumulative explained variance curve as shown in the figure below:

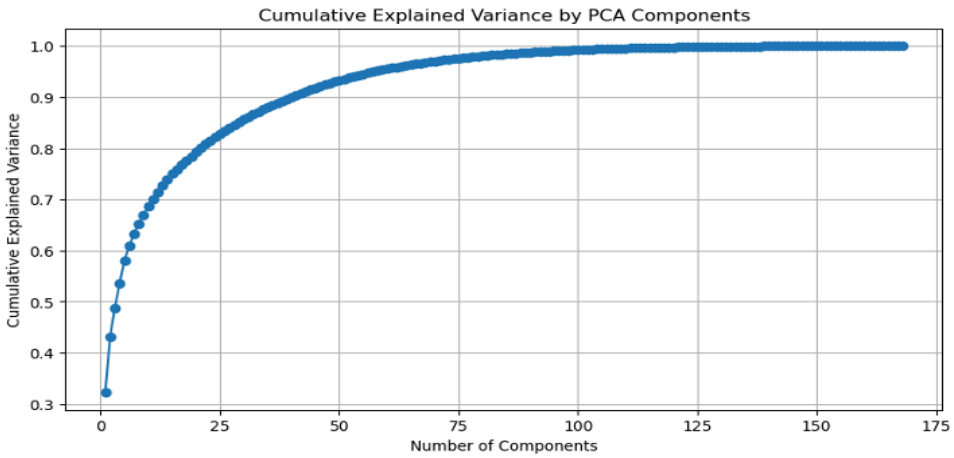

Figure 1. Cumulative explained variance (PCA) used to select 10 components.

Dendrograms- focusing on two weather stations: OSLO and VALENTIA 2020

What was learned from the dendrograms (OSLO & VALENTIA - 2020):

Single Method: Forms a long “chain,” which makes it hard to see meaningful structure or clear station relationships.

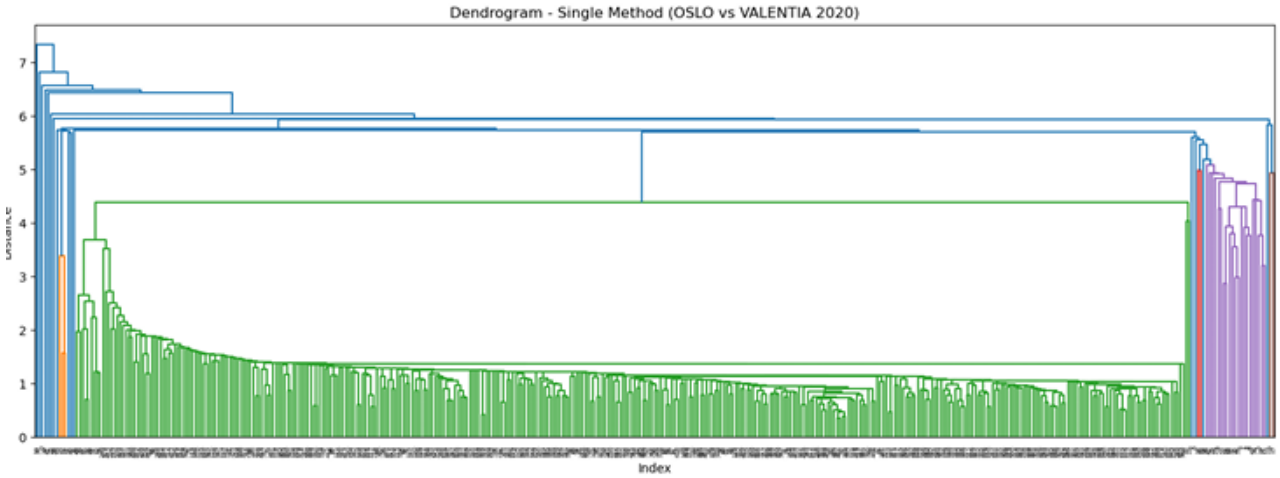

Figure 2. Dendrogram - Single Method (OSLO vs VALENTIA 2020).

Complete Method: Produces clearer separation than single linkage and is better at highlighting differences between OSLO and VALENTIA, but clusters can still feel broad.

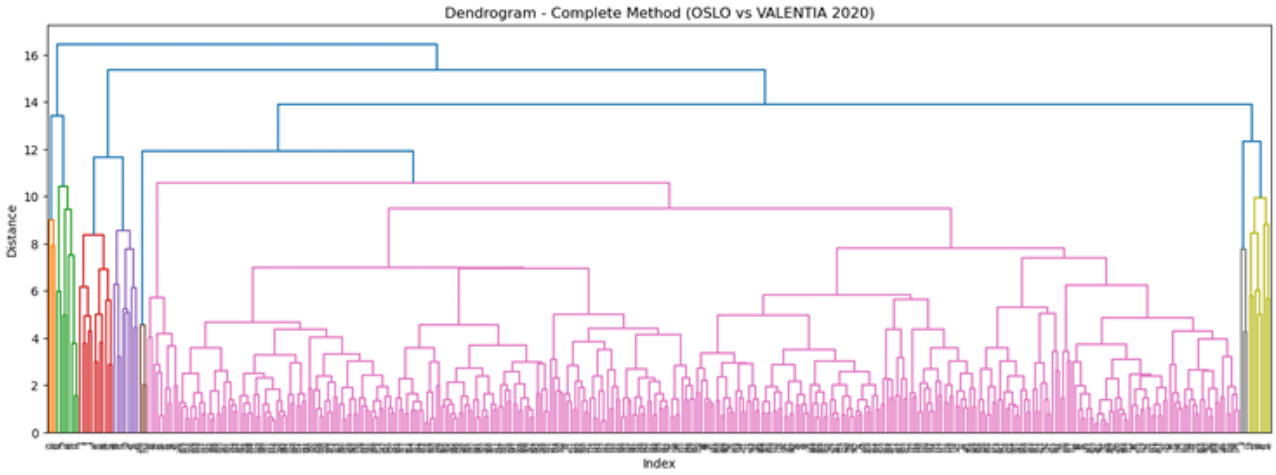

Figure 3. Dendrogram - Complete Method (OSLO vs VALENTIA 2020).

Average Method: Creates a few moderate groupings, but boundaries are not as crisp, so it's less decisive for spotting standout station connections.

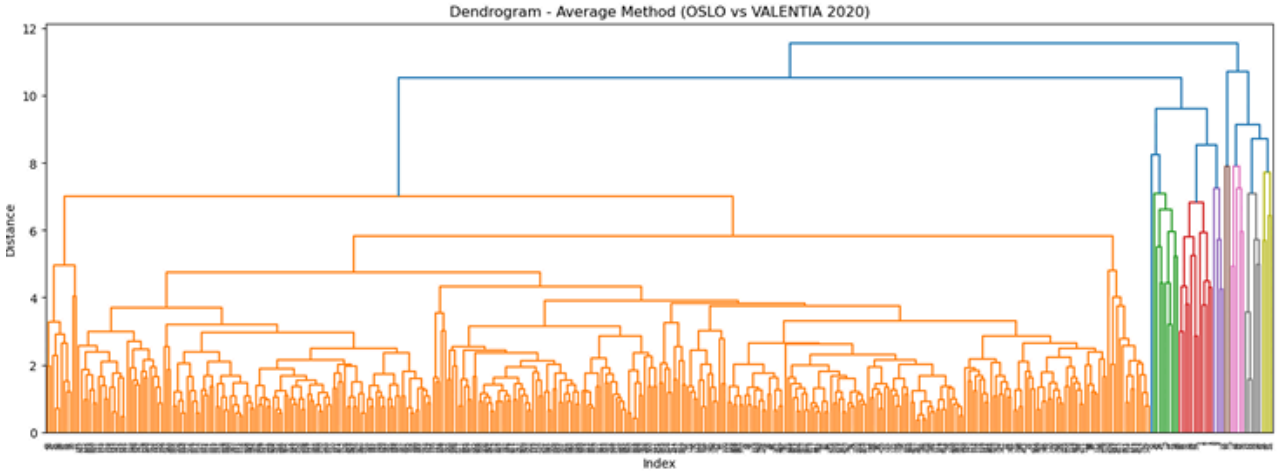

Figure 4. Dendrogram - Average Method (OSLO vs VALENTIA 2020).

Ward Method: Produces the clearest and most interpretable structure, giving the most useful separation between OSLO and VALENTIA patterns.

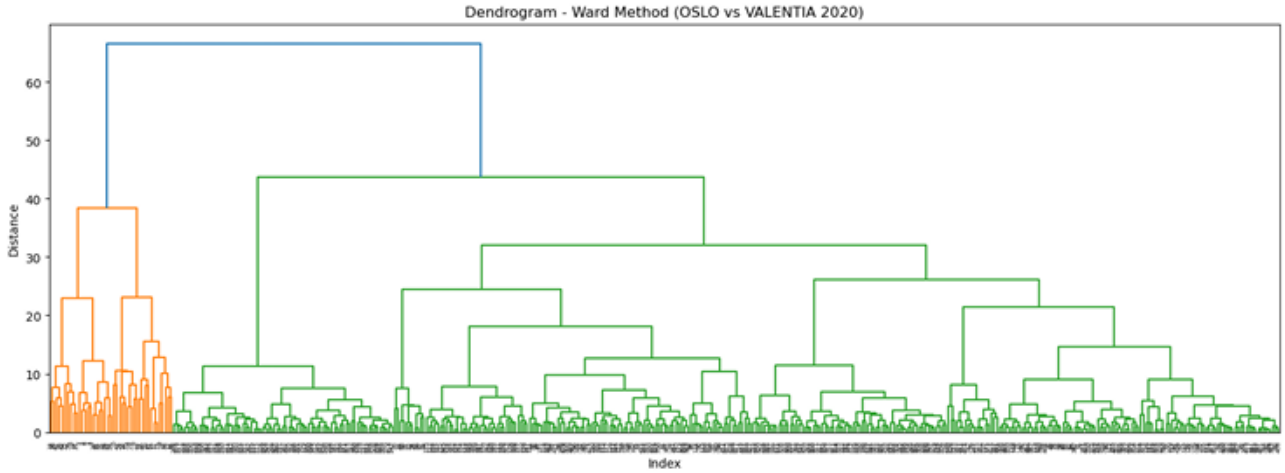

Figure 5. Dendrogram - Ward Method (OSLO vs VALENTIA 2020).

Key Takeaway

Overall takeaway (OSLO vs VALENTIA): Ward's method gives the most usable clustering for understanding whether OSLO and VALENTIA show a meaningful relationship (or a clear separation) in 2020

Pleasant Weather Validation Step

Pleasant weather data: The workflow also imports the pleasant weather dataset, which sets up a validation step—checking whether discovered clusters align with “pleasant” labels (not just station similarity).

Recommendations

Use Ward linkage on scaled data as the default for identifying connections that stand out, because it produces the most interpretable separation for OSLO vs VALENTIA. After clustering, compare cluster membership against the pleasant-weather labels to test whether clusters map to “pleasant” conditions rather than just station similarity