RV UNIVERSITY

Go, change the world

an initiative of RV EDUCATIONAL INSTITUTIONS

Applied Data Mining Using Orange: A Study on Classification, Clustering, and Time Series Forecasting

Introduction

In the current data age, extracting insightful knowledge from raw data is essential in every line of business. Software such as Orange Data Mining simplifies this by making strong data analysis, classification, clustering, and forecasting possible with little programming expertise. This project demonstrates an end-to-end data science process in Orange beginning from data selection and cleaning through implementation of multiple machine learning methods for pattern discovery, classification, and time series trend analysis for forecasting purposes.

We utilized two datasets from the real world to demonstrate Orange's potential: a high-resolution weather dataset for classification and clustering, and a monthly electricity generation dataset for time series analysis. With Orange's visual and interactive widgets like File, Select Columns, Preprocess, Classification Models, Clustering Tools, From Time Series, ARIMA, and Line Chart, we designed an end-to-end data analysis pipeline in a modular structure. This project underscores the need to know the data prior to modeling and shows that the convergence of various techniques can result in correct, understandable, and informative outputs.

Relevance / Importance of chosen topic

- Classification permits us to label data points according to features so that predictive work can be done. In our project, classifying temperature trends as "Warm" or "Cool" makes climatic action understandable in easy-to-grasp terms.
- Clustering allows the detection of implicit groupings within data without any pre-labeling. This unsupervised technique is necessary for finding natural segments in meteorological data, such as clustering days with comparable humidity, wind, and rainfall profiles.
- Time Series Analysis is important for learning about how data changes over time. Predicting electricity generation, say, has practical uses for resource planning, energy policy, and environmental sustainability

Description of the Mini Project and the tool used

This mini project is an in-depth study of principal machine learning methods employing real-world data sets and the Orange Data Mining software. The primary objective is to utilize classification, clustering, and time series forecasting to obtain insights in a visual and intuitive manner. The project takes a standard data science process from data selection and preprocessing to modeling and visualization.

Several Orange widgets have been utilized along the way:

- **File** and **Select Columns** for bringing in and sorting out data.
- **Preprocess** for cleaning up and normalizing.
- Classification models such as **Logistic Regression, kNN, and Random Forest,** tested by using **Test & Score** and **Confusion Matrix**.
- **Clustering methods** such as K-Means and Hierarchical Clustering.
- **From Time Series**, **ARIMA**, and **Line Chart** for time series forecasting and visualization.

Implementation / Procedure / Steps to execute the task

1. Classification

The classification component of the project was executed using **Orange**, a visual programming tool that simplifies the application of machine learning algorithms through an intuitive drag-and-drop interface. The classification task focused on predicting the **temperature class (TempClass)** based on various meteorological features such as atmospheric pressure (p), temperature (T), dew point (Tdew), relative humidity (rh), and vapor pressure readings (VPmax, VPact, VPdef), among others.

The procedure began by importing the dataset using the **File** widget, followed by **Select Columns** to define the target variable (TempClass) and distinguish it from the features. The target attribute TempClass was a categorical label with two values—likely Cool and Warm.

- **Normalization**: Standardized all features to zero mean and unit variance.
- **Imputation**: Missing values were filled using the average or most frequent strategy.
- **Discretization**: Continuous attributes were discretized using **Equal Frequency** discretization to facilitate learning in algorithms like Naive Bayes.

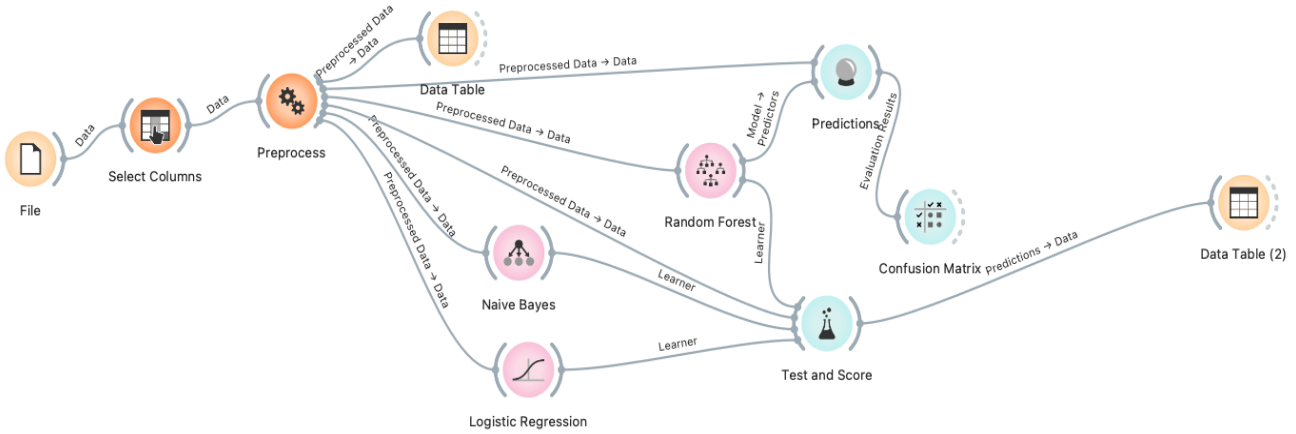
Post-preprocessing, the data was split and fed into three different classifiers:

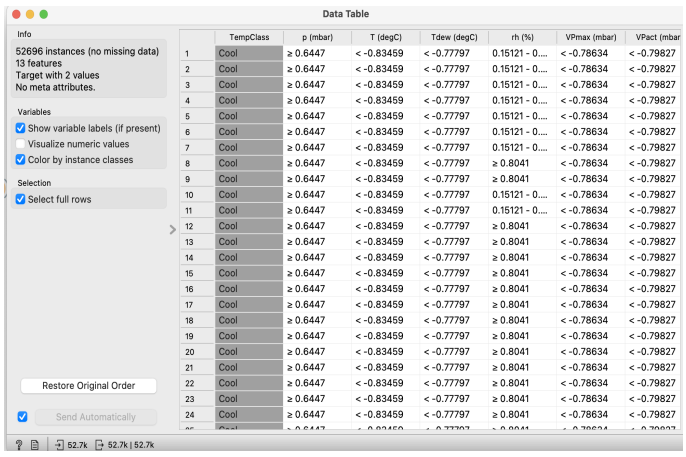
1. **Logistic Regression** – configured with Ridge (L2) regularization.
2. **Naive Bayes** – a probabilistic classifier based on Bayes' theorem.
3. **Random Forest** – an ensemble of 10 decision trees, where minimum subset size was set to 5 for pruning.

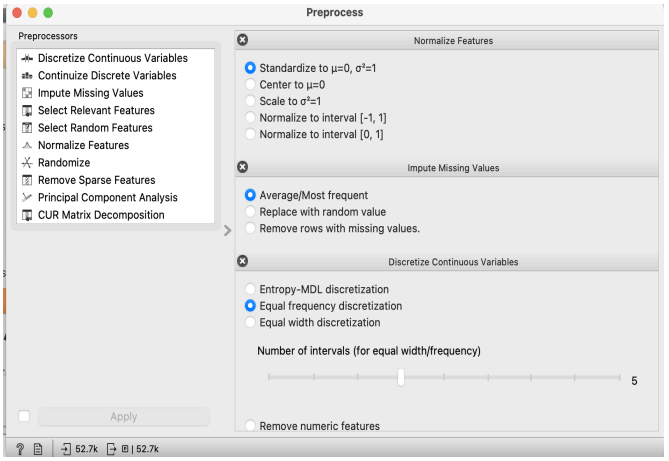
The models were evaluated using the **Test and Score** widget with **5-fold cross-validation** and **10 repeats**, stratified to maintain class balance. Evaluation metrics included:

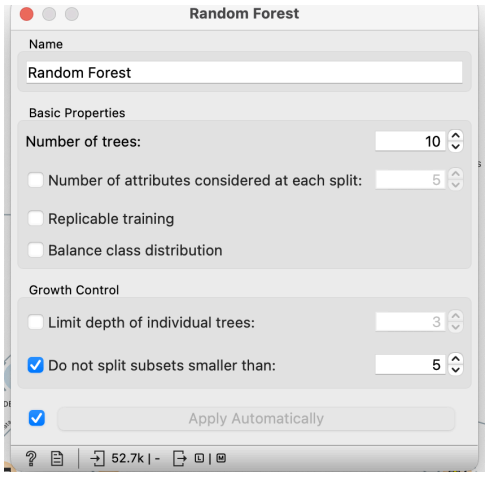
- **AUC** (Area Under Curve)
- **CA** (Classification Accuracy)
- **F1-score**
- **Precision**
- **Recall**

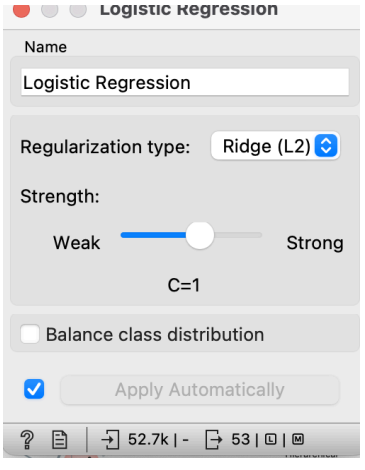
- **MCC** (Matthews Correlation Coefficient)

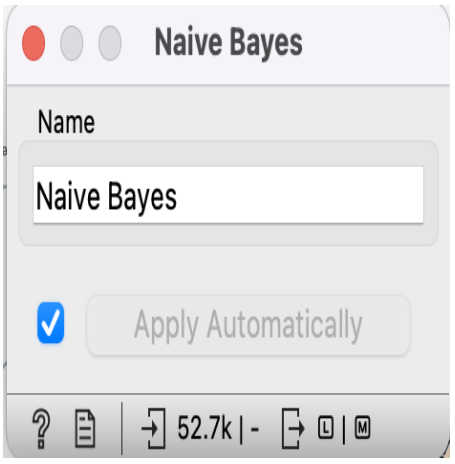










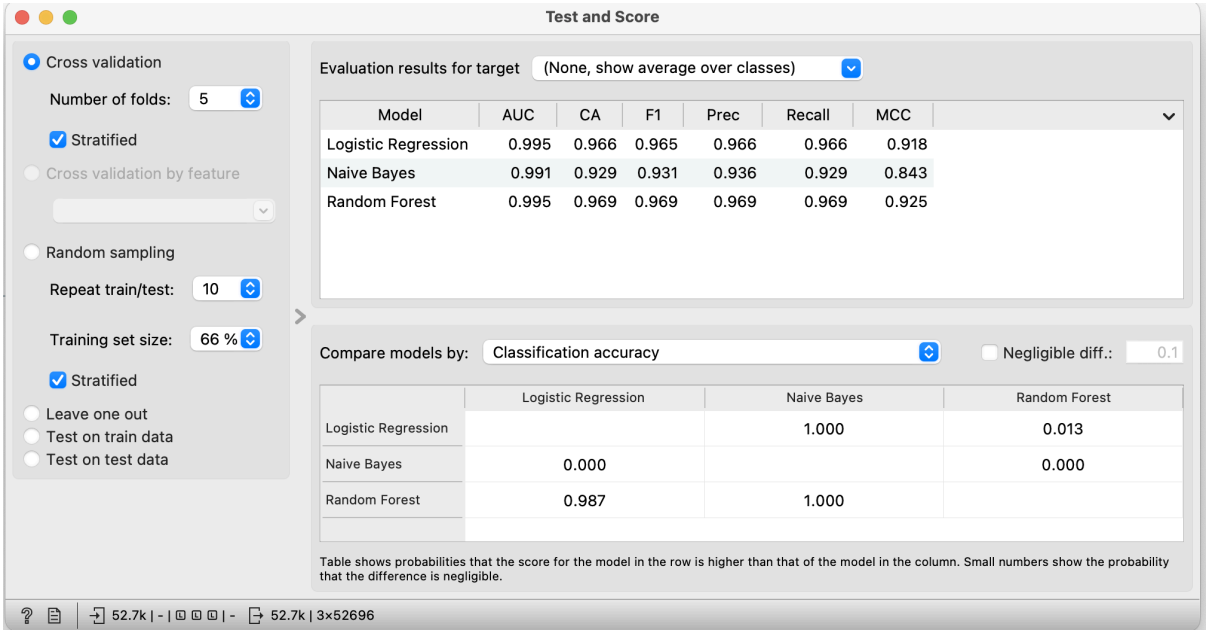


Test and Score – Model Evaluation & Comparison

The **Test and Score** widget played a central role in evaluating the performance of the three classification models: **Logistic Regression**, **Naive Bayes**, and **Random Forest**. Using **5-fold**

cross-validation repeated **10 times** (stratified sampling), we ensured robustness and consistency in the model evaluation.

Model-wise Performance Overview:



Why Random Forest Was Chosen for Confusion Matrix

Given its **highest average classification accuracy (96.9%)**, **best F1-score**, and **lowest misclassification rate**, Random Forest was selected for the **Confusion Matrix** analysis. The matrix confirms this decision:

- **Correctly predicted 'Cool' class**: 36,440
- **Correctly predicted 'Warm' class**: 14,842
- **Misclassifications** are minimal: 503 instances of 'Cool' misclassified as 'Warm', and 911 of 'Warm' as 'Cool'.

Findings from Classification

1. Models Compared

You evaluated three classification algorithms:

- **Logistic Regression**

- **Naive Bayes**
- **Random Forest**

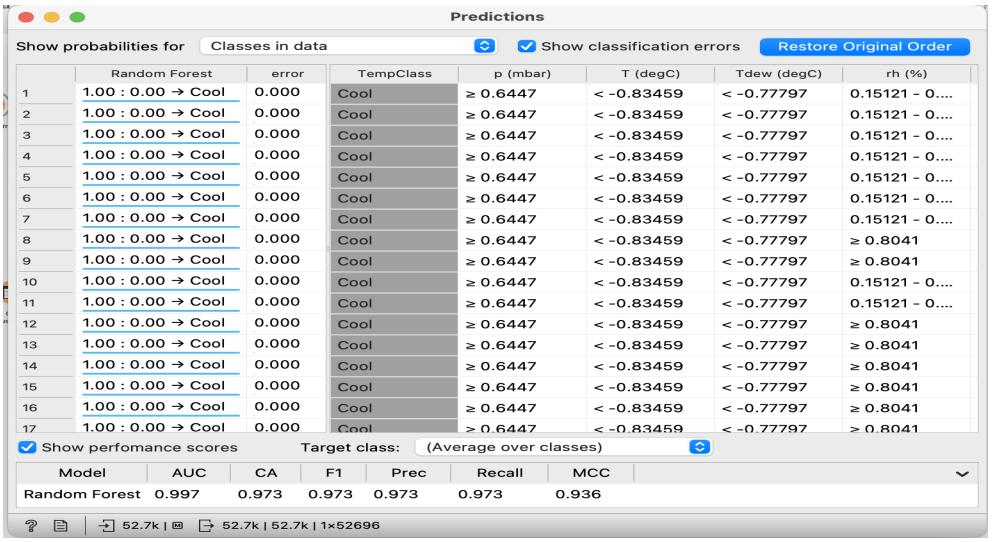
Each model was trained and evaluated using **5-fold stratified cross-validation** and a **66%-34% training-testing split**, repeated 10 times. This ensures robustness in model evaluation.

---

3. Why Random Forest Was Chosen for Final Evaluation

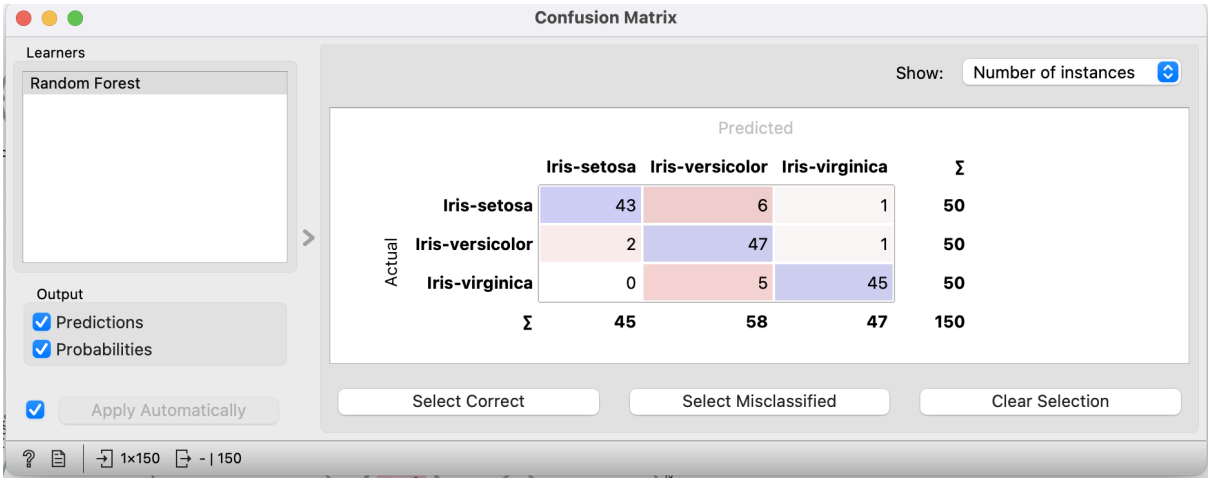
- **Best Accuracy**: With 96.9%, it edged ahead of Logistic Regression (96.6%) and significantly outperformed Naive Bayes (92.9%).
- **High Stability**: The performance was consistent across metrics—strong AUC, F1-score, and MCC—indicating a balanced classifier with minimal bias.
- **Low Misclassification**: Confusion Matrix shows only:
    - 503 Cool misclassified as Warm
    - 911 Warm misclassified as Cool
    - These are minor compared to over 36,000 correct predictions for Cool and 14,842 correct for Warm.

Thus, Random Forest was the best choice for evaluating classification **due to superior generalization, minimal error, and balanced performance**.



---

4. Insights from Confusion Matrix



Cool Class: 98.6% correctly classified

Warm Class: 94.2% correctly classified

Very few false positives and negatives, indicating **low classification error**.

Final Verdict

- **Random Forest** is the most reliable model for predicting TempClass from meteorological data.
- It is **accurate**, **stable**, and effectively distinguishes between the Cool and Warm classes.
- You made a sound decision to use **Random Forest for generating predictions and confusion matrix analysis**.

2. Clustering

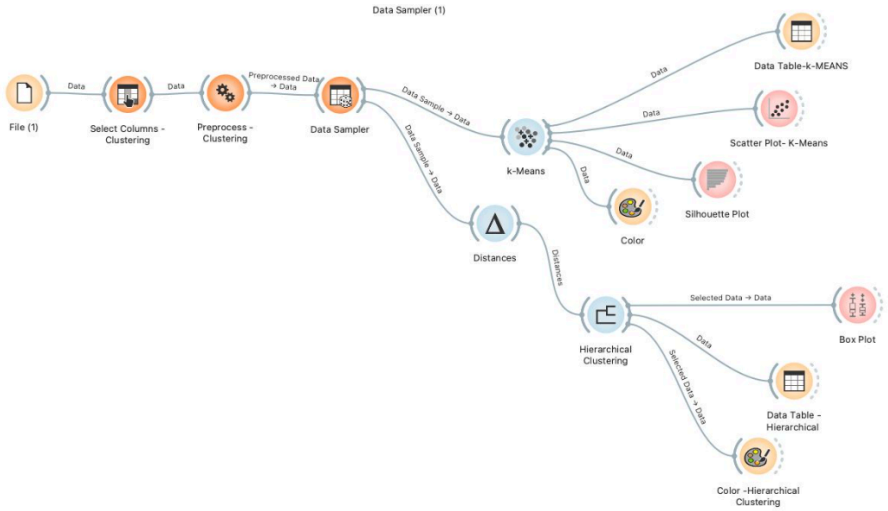
Clustering is an unsupervised learning technique used to discover natural groupings within a dataset, where similar instances are grouped together without the need for predefined labels. In this mini-project, clustering was applied to meteorological data with the objective of identifying weather patterns and environmental similarities based on features like temperature, humidity, vapor pressure, wind speed, rainfall, and radiation.

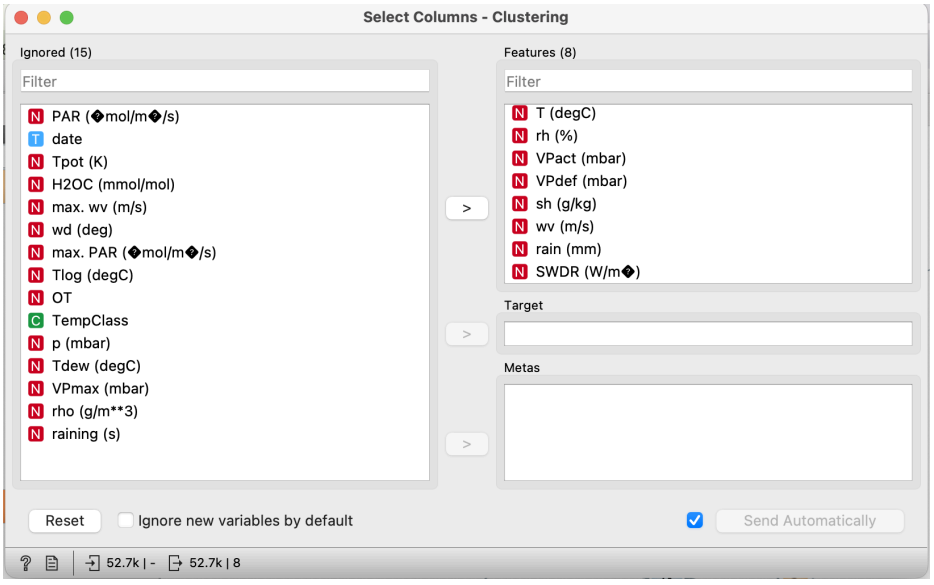
1. **Hierarchical Clustering**
   A tree-based method that builds nested clusters by successively merging or splitting them, visualized using dendrograms and distance matrices.

2. **K-Means Clustering**
   A centroid-based clustering method that partitions the data into *k* pre-defined groups, optimizing intra-cluster similarity and minimizing inter-cluster variation.

Each algorithm was followed by supporting visualization techniques, such as:

- **Scatter plots** for spatial distribution of clusters,
- **Color maps** to highlight cluster membership across feature ranges,
- **Silhouette plots** for cluster quality assessment (K-Means),
- **Distance matrices** for hierarchical clustering relationships,
- **Clustered Data Tables** displaying discretized variables for deeper interpretation.





Data Preparation

Relevant continuous features were selected (e.g., T (degC), rh (%), VPact, VPdef, sh, wv, rain, and SWDR) using Orange's column selection widget. The target variable TempClass and unrelated fields like date were excluded to retain purely unsupervised clustering behavior.

To ensure reliable clustering:

- Features were **standardized** to have mean 0 and variance 1.
- **Missing values** were imputed using the most frequent value.
- **Equal frequency discretization** was applied to improve interpretability of feature intervals.
- The dataset was **downsampled to 3000 instances** using the Data Sampler widget for faster execution and visual clarity.

Hierarchical Clustering: Implementation & Findings

Hierarchical Clustering was also used on the meteorological dataset to identify natural groupings without specifying the number of clusters ahead of time. The dendrogram visually depicted how data points clustered based on common atmospheric features. Clusters C1, C2, and C3 were easily distinguished in the tree, indicating different patterns in the data. These clusters were selected because they were most prominent and understandable in the dendrogram layout. This hierarchical method supplemented other clustering techniques by providing a tree-based view of data similarity. Overall, it yielded useful information on the natural structure of meteorological variables.

Key Cluster Insights

Feature	Cluster C1 – Cold, Dry, Stable	Cluster C2 – Moderate, Transitional	Cluster C3 – Warm, Humid, Rain-Rich
Temperature (°C)	Very Low (< -0.71098)	Low to Moderate (−0.83459 to −0.07943)	High (≥ 0.71858)
Relative Humidity (rh %)	Very Low (< -0.71098)	Slightly Improved (−0.71098 to 0.15121)	High to Very High (≥ 0.8041)
Vapour Pressure (VPact)	Low (< -0.79827)	Mid-range (−0.79827 to −0.19432)	High (≥ 0.59108)

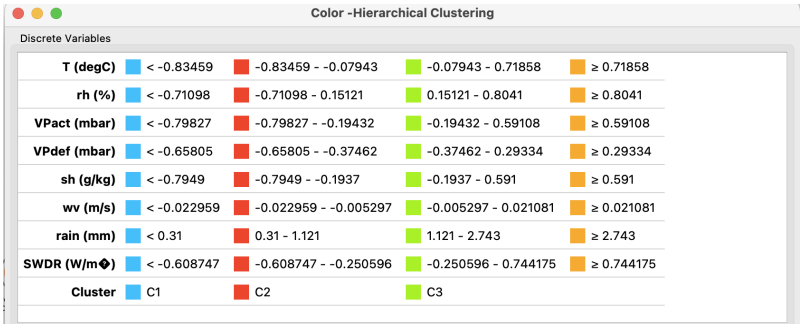
Vapour Pressure Deficit	Low (< -0.65805)	Mid-range (−0.65805 to −0.37462)	High (≥ 0.29334)
Specific Humidity (g/kg)	Low (< -0.7949)	Moderate (−0.7949 to −0.1937)	High (≥ 0.591)
Wind Velocity (m/s)	Mostly Calm (< -0.022959)	Moderate (−0.022959 to −0.005297)	High (≥ 0.021081)
Rainfall (mm)	Minimal (< 0.31)	Low to Moderate (0.31 to 1.121)	Heavy (≥ 2.743)
Shortwave Radiation (W/m^2)	Weak (< -0.608747)	Moderate (−0.608747 to −0.250596)	Strong (≥ 0.744175)
Interpretation	Cold, dry, low-energy, stable (winter-like)	Transitional, spring/early monsoon-like	Tropical/monsoonal, warm, storm-prone

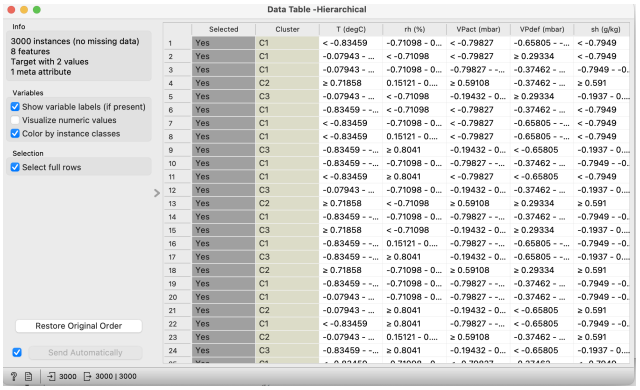
Summary of Cluster Utility

- **C1** helps detect **cold and dry spells**.
- **C2** captures **moderate or transitioning periods**.
- **C3** focuses on **warm, wet extremes**—ideal for identifying heatwaves or monsoon effects.

Observations:

- The clustering was robust, with minimal overlap between groups and clear separation of climate types.
- Unlike DBSCAN, which captured nuanced local densities, hierarchical clustering provided a **global perspective of similarity**, ideal for **multi-level data exploration**.
- The **dendrogram structure** aids in visualising not just final groupings but also how closely related various instances are at different levels of the tree.
- **C1**: Cold & dry → suggests winter or desert-like climate zones.
- **C2**: Moderate weather → acts as a transitional buffer.
- **C3**: Hot, humid, storm-prone → typical of tropical monsoon climates.
- Hierarchical clustering provided **multi-scale understanding** of weather similarities and helped differentiate between **low, medium, and high-intensity environmental segments**.





K-Means Clustering: Implementation & Findings

Key Setup and Configuration:

- **Preprocessing:** Normalization was enabled to ensure all features contribute equally during distance calculation.
- **Initialization:** KMeans++ was used to improve centroid selection for faster convergence and better clustering quality.
- **Cluster Range Evaluated:** From 2 to 3 clusters.
- **Best Silhouette Score:**
    - **3 clusters** achieved a higher silhouette score of **0.160** compared to **0.137** for 2 clusters, indicating better cluster separation.

Silhouette Analysis:

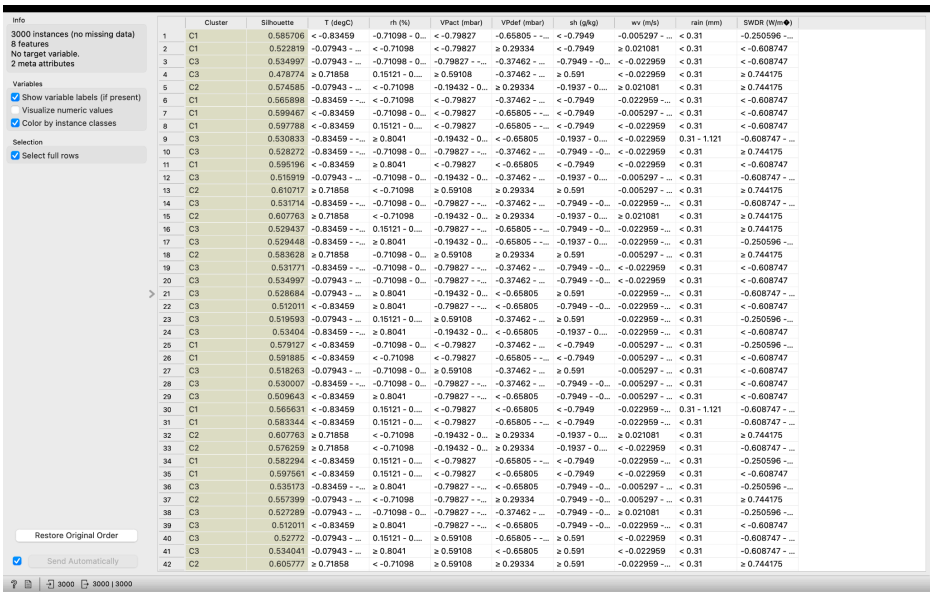
- The silhouette plot grouped by Manhattan distance showed:
    - **Cluster C1 (Blue)** had a moderately high silhouette score, indicating well-separated points.
    - **Cluster C2 (Red)** had the **highest density** of instances with **fair intra-cluster compactness**.

- **Cluster C3 (Green)** had slightly lower silhouette values, suggesting some borderline points or overlap with others.

Scatter Plot Interpretation:

- **Scatter Plot (T vs rh)**:
  - Clear visual separation was seen among clusters on the temperature (T) and relative humidity (rh) axes.
  - **C1:** Occupies lower temperature and lower humidity regions (dry and cold).
  - **C2:** Middle of the range – moderate temperature and humidity (balanced conditions).
  - **C3:** Higher temperature and humidity range – indicating warm and moist conditions.

Cluster Characteristics (from Data Table and Colour Map):



Info

3000 instances (no missing data)
8 features
No target variable.
2 meta attributes

Variables
- Show variable labels (if present)
- Visualize numeric values
- Color by instance classes

Selection
- Select full rows

Restore Original Order
- Send Automatically

3000 | 3000

	Cluster	Silhouette	T (degC)	rh (%)	VPact (mbar)	VPdef (mbar)	sh (g/kg)	wv (m/s)	rain (mm)	SWDR (W/m²)
1	C1	0.585706	< -0.83459	-0.71098 - 0...	< -0.79827	-0.65805 - -...	< -0.7949	-0.005297 - ...	< 0.31	-0.250596 -...
2	C1	0.522819	-0.07943 - ...	< -0.71098	< -0.79827	≥ 0.29334	< -0.7949	≥ 0.021081	< 0.31	< -0.608747
3	C3	0.534997	-0.07943 - ...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	< -0.022959	< 0.31	< -0.608747
4	C3	0.478774	≥ 0.71858	0.15121 - 0....	≥ 0.59108	-0.37462 - ...	≥ 0.591	< -0.022959	< 0.31	≥ 0.744175
5	C2	0.574585	-0.07943 - ...	< -0.71098	-0.19432 - 0...	≥ 0.29334	-0.1937 - 0....	≥ 0.021081	< 0.31	≥ 0.744175
6	C1	0.565898	-0.83459 - -...	< -0.71098	< -0.79827	-0.37462 - ...	< -0.7949	-0.022959 -...	< 0.31	< -0.608747
7	C1	0.599467	< -0.83459	-0.71098 - 0...	< -0.79827	-0.65805 - -...	< -0.7949	-0.005297 - ...	< 0.31	< -0.608747
8	C1	0.597788	< -0.83459	0.15121 - 0....	< -0.79827	-0.65805 - -...	< -0.7949	< -0.022959	< 0.31	< -0.608747
9	C3	0.530833	-0.83459 - -...	≥ 0.8041	-0.19432 - 0...	< -0.65805	-0.1937 - 0....	< -0.022959	0.31 - 1.121	-0.608747 - ...
10	C3	0.528272	-0.83459 - -...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	< -0.022959	< 0.31	≥ 0.744175
11	C1	0.595196	< -0.83459	≥ 0.8041	< -0.79827	< -0.65805	< -0.7949	< -0.022959	< 0.31	< -0.608747
12	C3	0.515919	-0.07943 - ...	-0.71098 - 0...	-0.19432 - 0...	-0.37462 - ...	-0.1937 - 0....	-0.005297 - ...	< 0.31	-0.608747 - ...
13	C2	0.610717	≥ 0.71858	< -0.71098	≥ 0.59108	≥ 0.29334	≥ 0.591	-0.005297 - ...	< 0.31	≥ 0.744175
14	C3	0.531714	-0.83459 - -...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	-0.022959 -...	< 0.31	-0.608747 - ...
15	C2	0.607763	≥ 0.71858	< -0.71098	-0.19432 - 0...	≥ 0.29334	-0.1937 - 0....	≥ 0.021081	< 0.31	≥ 0.744175
16	C3	0.529437	-0.83459 - -...	0.15121 - 0....	-0.79827 - -...	-0.65805 - -...	-0.7949 - -0...	-0.022959 -...	< 0.31	≥ 0.744175
17	C3	0.529448	-0.83459 - -...	≥ 0.8041	-0.19432 - 0...	-0.65805 - -...	-0.1937 - 0....	-0.022959 -...	< 0.31	-0.250596 -...
18	C2	0.583628	≥ 0.71858	-0.71098 - 0...	≥ 0.59108	≥ 0.29334	≥ 0.591	-0.005297 - ...	< 0.31	≥ 0.744175
19	C3	0.531771	-0.83459 - -...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	< -0.022959	< 0.31	< -0.608747
20	C3	0.534997	-0.07943 - ...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	< -0.022959	< 0.31	< -0.608747
21	C3	0.528684	-0.07943 - ...	≥ 0.8041	-0.19432 - 0...	< -0.65805	≥ 0.591	-0.022959 -...	< 0.31	-0.608747 - ...
22	C3	0.512011	< -0.83459	≥ 0.8041	-0.79827 - -...	< -0.65805	-0.7949 - -0...	-0.022959 -...	< 0.31	< -0.608747
23	C3	0.519593	-0.07943 - ...	0.15121 - 0....	≥ 0.59108	-0.37462 - ...	≥ 0.591	-0.022959 -...	< 0.31	-0.250596 -...
24	C3	0.53404	-0.83459 - -...	≥ 0.8041	-0.19432 - 0...	< -0.65805	-0.1937 - 0....	-0.022959 -...	< 0.31	< -0.608747
25	C1	0.579127	< -0.83459	-0.71098 - 0...	< -0.79827	-0.37462 - ...	< -0.7949	-0.005297 - ...	< 0.31	-0.250596 -...
26	C1	0.591885	< -0.83459	< -0.71098	< -0.79827	-0.65805 - -...	< -0.7949	-0.005297 - ...	< 0.31	< -0.608747
27	C3	0.518263	-0.07943 - ...	-0.71098 - 0...	≥ 0.59108	-0.37462 - ...	≥ 0.591	-0.005297 - ...	< 0.31	< -0.608747
28	C3	0.530007	-0.83459 - -...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	-0.005297 - ...	< 0.31	< -0.608747
29	C3	0.509643	< -0.83459	≥ 0.8041	-0.79827 - -...	< -0.65805	-0.7949 - -0...	-0.005297 - ...	< 0.31	< -0.608747
30	C1	0.565631	< -0.83459	0.15121 - 0....	< -0.79827	< -0.65805	< -0.7949	-0.022959 -...	0.31 - 1.121	-0.608747 - ...
31	C1	0.583344	< -0.83459	0.15121 - 0....	< -0.79827	-0.65805 - -...	< -0.7949	-0.022959 -...	< 0.31	-0.608747 - ...
32	C2	0.607763	≥ 0.71858	< -0.71098	-0.19432 - 0...	≥ 0.29334	-0.1937 - 0....	≥ 0.021081	< 0.31	≥ 0.744175
33	C2	0.576259	≥ 0.71858	< -0.71098	-0.19432 - 0...	≥ 0.29334	-0.1937 - 0....	< -0.022959	< 0.31	-0.608747 - ...
34	C1	0.582294	< -0.83459	0.15121 - 0....	< -0.79827	-0.65805 - -...	< -0.7949	-0.022959 -...	< 0.31	-0.250596 -...
35	C1	0.597561	< -0.83459	0.15121 - 0....	< -0.79827	< -0.65805	< -0.7949	< -0.022959	< 0.31	< -0.608747
36	C3	0.535173	-0.83459 - -...	≥ 0.8041	-0.79827 - -...	< -0.65805	-0.7949 - -0...	-0.005297 - ...	< 0.31	-0.250596 -...
37	C2	0.557399	-0.07943 - ...	< -0.71098	-0.79827 - -...	≥ 0.29334	-0.7949 - -0...	-0.005297 - ...	< 0.31	≥ 0.744175
38	C3	0.527289	-0.07943 - ...	-0.71098 - 0...	-0.79827 - -...	-0.37462 - ...	-0.7949 - -0...	≥ 0.021081	< 0.31	-0.250596 -...
39	C3	0.512011	< -0.83459	≥ 0.8041	-0.79827 - -...	< -0.65805	-0.7949 - -0...	-0.022959 -...	< 0.31	< -0.608747
40	C3	0.52772	-0.07943 - ...	0.15121 - 0....	≥ 0.59108	-0.65805 - -...	≥ 0.591	< -0.022959	< 0.31	-0.608747 - ...
41	C3	0.534041	-0.07943 - ...	≥ 0.8041	≥ 0.59108	< -0.65805	≥ 0.591	< -0.022959	< 0.31	-0.608747 - ...
42	C2	0.605777	≥ 0.71858	< -0.71098	≥ 0.59108	≥ 0.29334	≥ 0.591	-0.022959 -...	< 0.31	≥ 0.744175

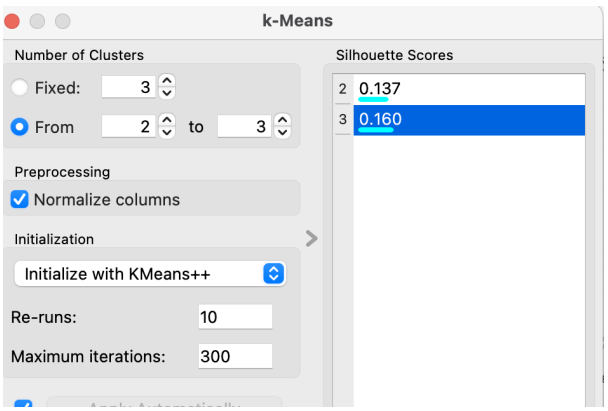
Feature	Cluster C1 (Blue) *Cold and Dry*	Cluster C2 (Red) *Moderate/Transitional*	Cluster C3 (Green) *Warm and Moist*
Temperature (°C)	Very Low (< -0.83459)	Mid-range (-0.83459 to -0.07943)	High (≥ 0.71858)
Relative Humidity (rh %)	Very Low (< -0.71098)	Low to Moderate (-0.71098 to 0.8041)	Very High (≥ 0.8041)
Rainfall (mm)	Minimal (< 0.31)	Moderate (0.31 to 1.121)	Heavy (≥ 1.121), some > 2.743
VPact (Vapour Pressure)	Low (< -0.79827)	Moderately High	High (≥ 0.59108)
VPdef (Vapour Pressure Deficit)	Low (< -0.65805)	Moderately High	High (≥ 0.29334)
Specific Humidity (sh g/kg)	Low (< -0.7949)	Moderate	High (≥ 0.591)
Wind Velocity (wv m/s)	Very Low (< -0.022959)	Moderate (-0.022959 to 0.021081)	Moderate to High (> 0.021081)
Shortwave Radiation (SWDR)	Low (< -0.608747)	Mid-level (-0.250596 to 0.744175)	Moderate to High (≥ 0.744175)

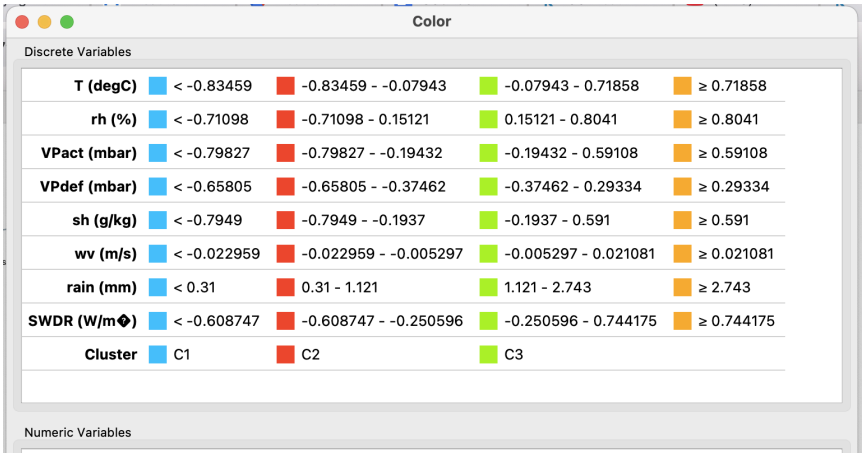
Silhouette Score Range	~0.52 – 0.61 (well-formed)	~0.41 – 0.61 (moderate compactness)	~0.10 – 0.53 (less compact, possible overlap)
Interpretation	Cold, dry, stable (winter/arid)	Seasonal shift (spring/autumn/monsoon onset)	Tropical, humid, rainy (monsoon/post-monsoon)

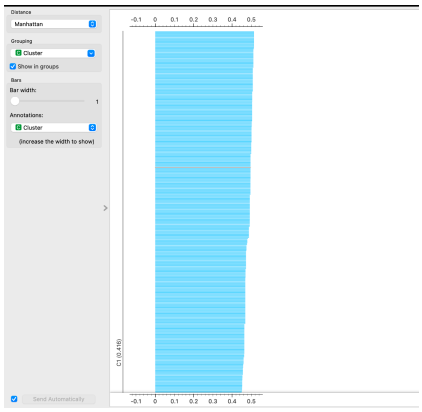
Key Insights:

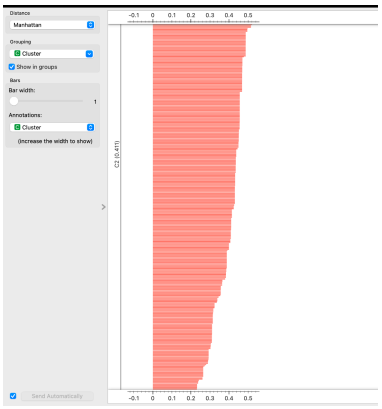
- **Interpretability:** K-Means yielded clearly interpretable clusters with reasonably good intra-cluster cohesion, evident from consistent Silhouette scores across instances.
- **Efficiency:** K-Means is computationally efficient, and its performance is evident with faster convergence and good visual separation.
- **Limitation:** Some boundary instances in C3 had lower silhouette values, suggesting overlapping zones which may be better handled by or Hierarchical methods

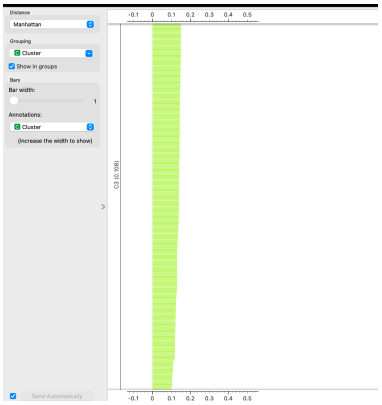
---

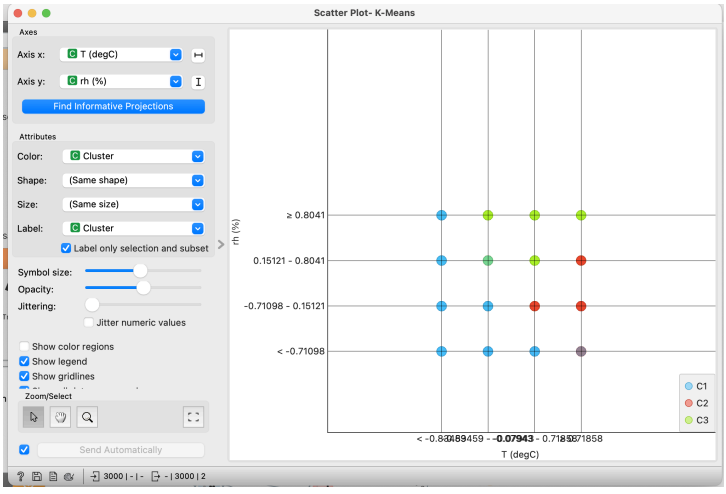

k-Means
Number of Clusters
Fixed: 3
From 2 to 3
Preprocessing
Normalize columns
Initialization
Initialize with KMeans++
Re-runs: 10
Maximum iterations: 300
Silhouette Scores
2 0.137
3 0.160

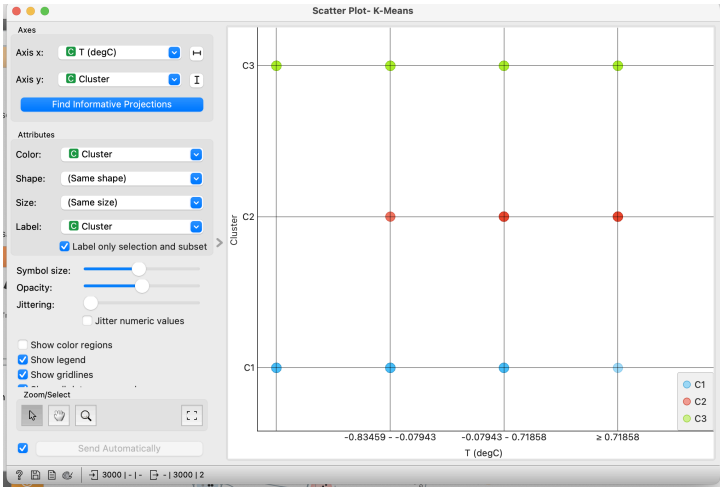

Color
Discrete Variables
T (degC) < -0.83459 -0.83459 - -0.07943 -0.07943 - 0.71858 ≥ 0.71858
rh (%) < -0.71098 -0.71098 - 0.15121 0.15121 - 0.8041 ≥ 0.8041
VPact (mbar) < -0.79827 -0.79827 - -0.19432 -0.19432 - 0.59108 ≥ 0.59108
VPdef (mbar) < -0.65805 -0.65805 - -0.37462 -0.37462 - 0.29334 ≥ 0.29334
sh (g/kg) < -0.7949 -0.7949 - -0.1937 -0.1937 - 0.591 ≥ 0.591
wv (m/s) < -0.022959 -0.022959 - -0.005297 -0.005297 - 0.021081 ≥ 0.021081
rain (mm) < 0.31 0.31 - 1.121 1.121 - 2.743 ≥ 2.743
SWDR (W/m²) < -0.608747 -0.608747 - -0.250596 -0.250596 - 0.744175 ≥ 0.744175
Cluster C1 C2 C3
Numeric Variables



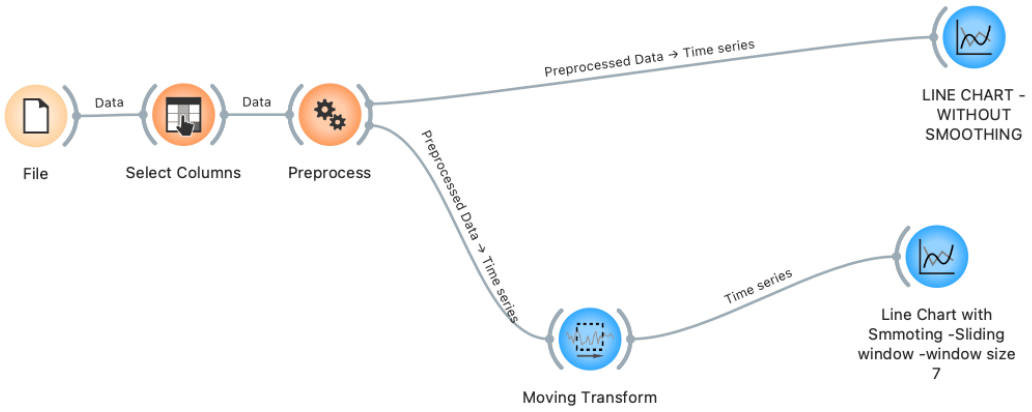





Scatter Plot- K-Means
Axis x: T (degC)
Axis y: rh (%)
Find Informative Projections
Color: Cluster
Shape: (Same shape)
Size: (Same size)
Label: Cluster
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Send Automatically
≥ 0.8041
0.15121 - 0.8041
-0.71098 - 0.15121
< -0.71098
C1
C2
C3
T (degC)


Scatter Plot- K-Means
Axis x: T (degC)
Axis y: Cluster
Find Informative Projections
Color: Cluster
Shape: (Same shape)
Size: (Same size)
Label: Cluster
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Send Automatically
C3
C2
C1
Cluster
-0.83459 - -0.07943
-0.07943 - 0.71858
≥ 0.71858
T (degC)

3. Time Series Analysis – Smoothing using Sliding Window



Workflow Summary:

- **Selected Features**: Temperature (T °C), Rainfall (mm), Wind Velocity (wv), Relative Humidity, VPact, VPdef, sh, and SWDR.
- **Preprocessing**:
  - **Normalization**: Standardized to mean = 0, variance = 1.
  - **Imputation**: Handled missing values using the "Most Frequent" method.
- **Time Series Settings**:
  - **Sliding Window**: Applied with a window width of 7 days to smooth fluctuations.
  - **Charts Used**: Line Chart without Smoothing and Line Chart with Smoothing.

Observations (Without Smoothing):

Temperature (T °C)

- Shows a clear **seasonal trend** with a gradual rise from January to July (peaking in summer), followed by a decline post-August.
- Significant day-to-day **variability**—likely due to fluctuating weather conditions.

Rainfall (mm)

- Highly **sporadic and spiky**, indicating **intermittent rain events**.
- Highest peaks are seen around **July**, matching **monsoon season** behavior in many regions.
- Remains **mostly low** throughout other months.

Observations (Smoothed with Sliding Window):

Temperature (T °C)

- The 7-day moving average smooths out noise, revealing a **more consistent seasonal curve**.
- The rising and falling trend is **better visualised**—useful for forecasting temperature trends.

Wind Velocity (wv)

- The smoothed graph shows clearer **wind speed oscillations**.
- Slight rise around mid-year (possibly stormy months or cyclonic winds).
- Despite smoothing, wind speeds remain relatively **low on average**.

Insights:

- **Smoothed temperature data** is excellent for identifying **seasonal patterns**—useful for climate modelling or agricultural planning.
- **Rainfall trends**, though erratic, can be correlated with **specific months (like July-August)** to identify monsoon periods.
- **Wind speed changes** are subtle, but smoothing helps highlight longer-term behaviour compared to noisy daily values.

3.1 . ARIMA Time Series Forecasting – Combined

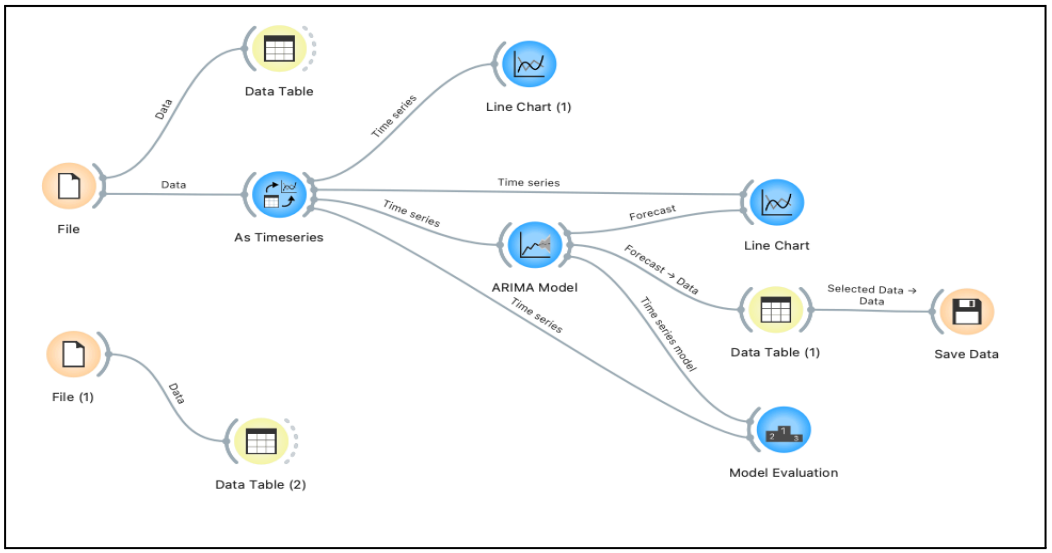


Fig: For Employment Rate and Daily Minimum Temperature

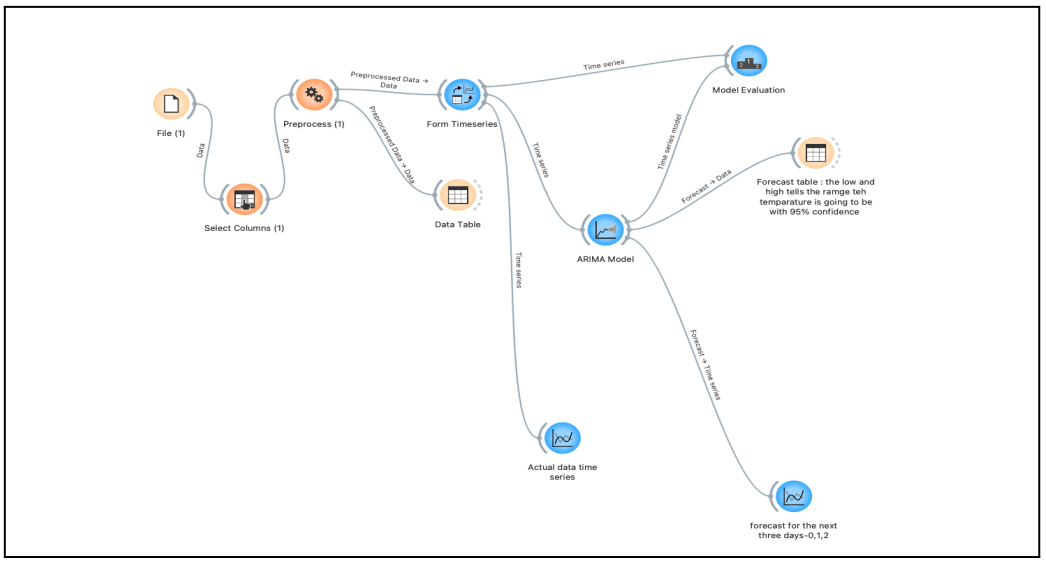


Fig: Forecast For Weather Data

Step-by-Step Procedure (Common to All 3 Datasets)

1. **Load Dataset**
   Use the File widget in Orange to upload the dataset.

2. **Visualize Raw Data**
   Connect to Data Table and Line Chart to inspect the dataset visually.

3. **Convert to Time Series Format**
   Use As Timeseries or Form Timeseries. Select the date column.

4. **Apply ARIMA Model**
    - Configure ARIMA with suitable parameters (p, d, q).
    - Set forecast steps ahead (e.g., 3 or 12 days).
    - Enable confidence intervals to see forecast range.

5. **Model Evaluation**
    - Add Model Evaluation widget.
    - Use cross-validation (e.g., 8 or 20 folds).
    - Observe RMSE, MAE, R^2, etc.

6. **Visualize Forecasts**
    - Connect Line Chart and Data Table to ARIMA output.
    - View forecast values and their confidence ranges.

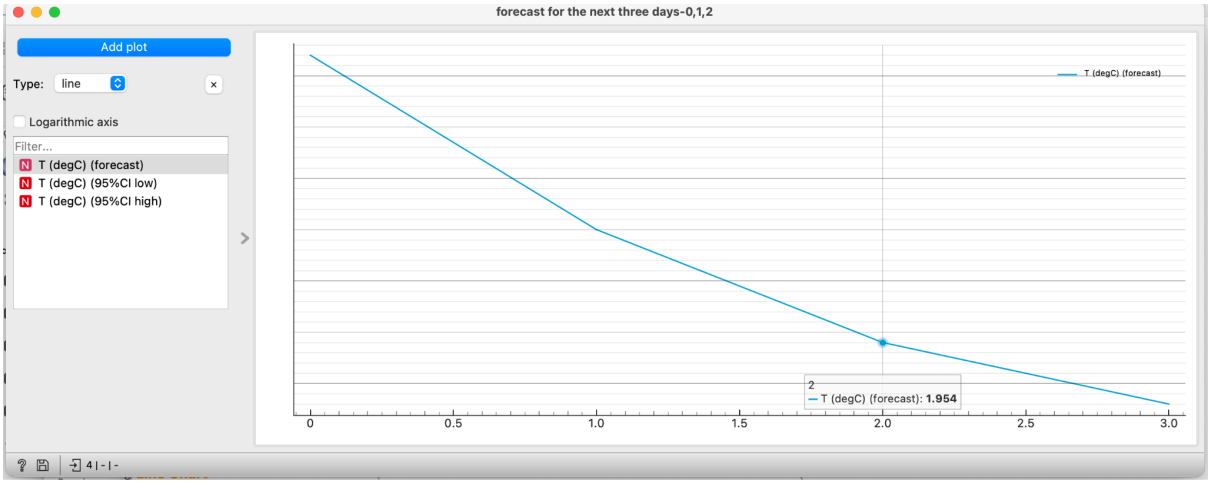
7. **Save Results (Optional)**
    - Use the Save Data widget to export the forecasted data.

Dataset 1: Weather Temperature Forecast

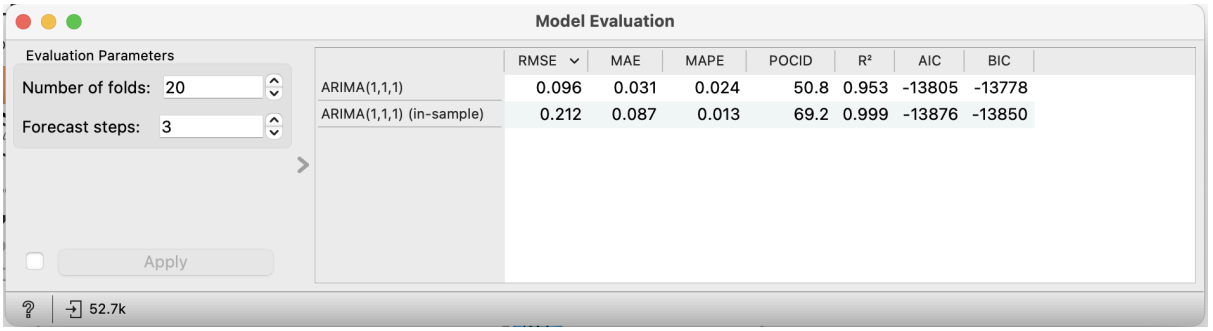
- **ARIMA(1,1,1)**, Forecast 4 days ahead
- Used daily temperature (T degC) data

Observations:

- The model provides tight confidence intervals, indicating high reliability.
- R^2 value of 0.953 shows excellent fit on test data.
- Temperature shows a gradual decline across the forecast horizon.



Forecast Line Chart



	RMSE	MAE	MAPE	POCID	R^2	AIC	BIC
ARIMA(1,1,1)	0.096	0.031	0.024	50.8	0.953	-13805	-13778
ARIMA(1,1,1) (in-sample)	0.212	0.087	0.013	69.2	0.999	-13876	-13850

Evaluation Metrics

	T (degC) (forecast)	T (degC) (95%CI low)	T (degC) (95%CI high)
1	1.98184	1.56613	2.39756
2	1.9647	1.20825	2.72116
3	1.95427	0.879414	3.02913
4	1.94793	0.581481	3.31437

ARIMA Forecast

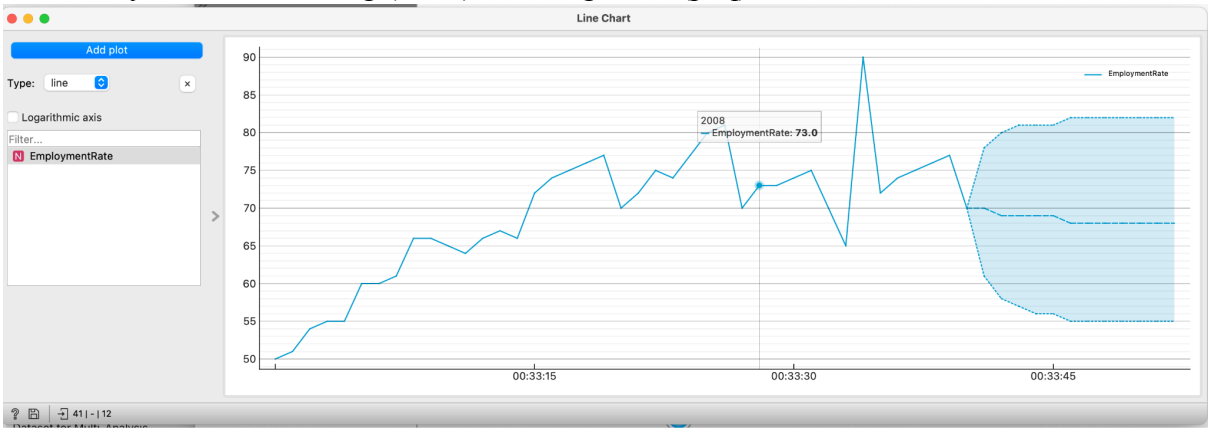
Dataset 2: Employment Rate Forecast

- **ARIMA(1,0,0)**, Forecast 12 steps ahead
- Used annual employment rate (percentage)

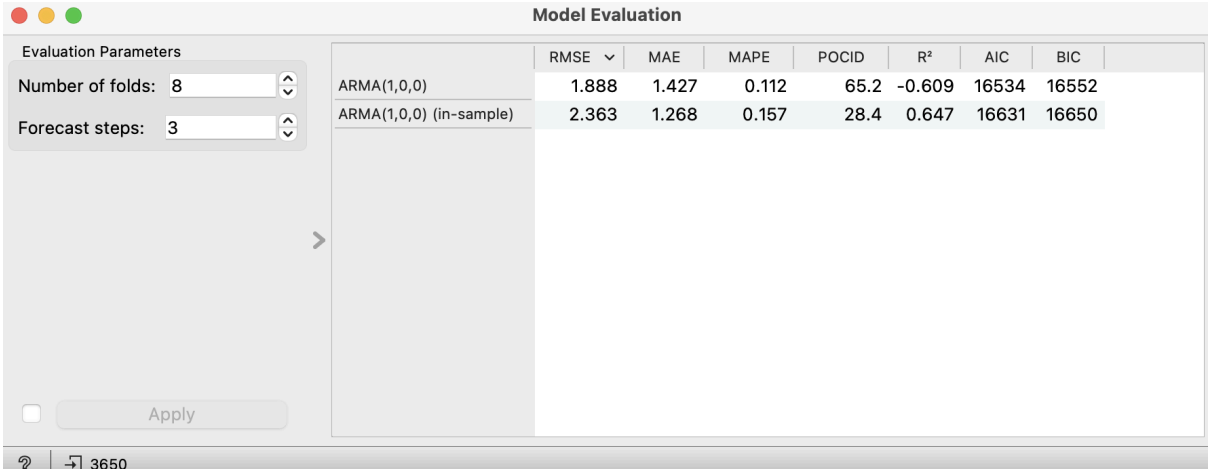
Observations:

- Negative R^2 suggests poor prediction power on unseen data.
- Confidence interval is wide – indicating high variance in prediction.

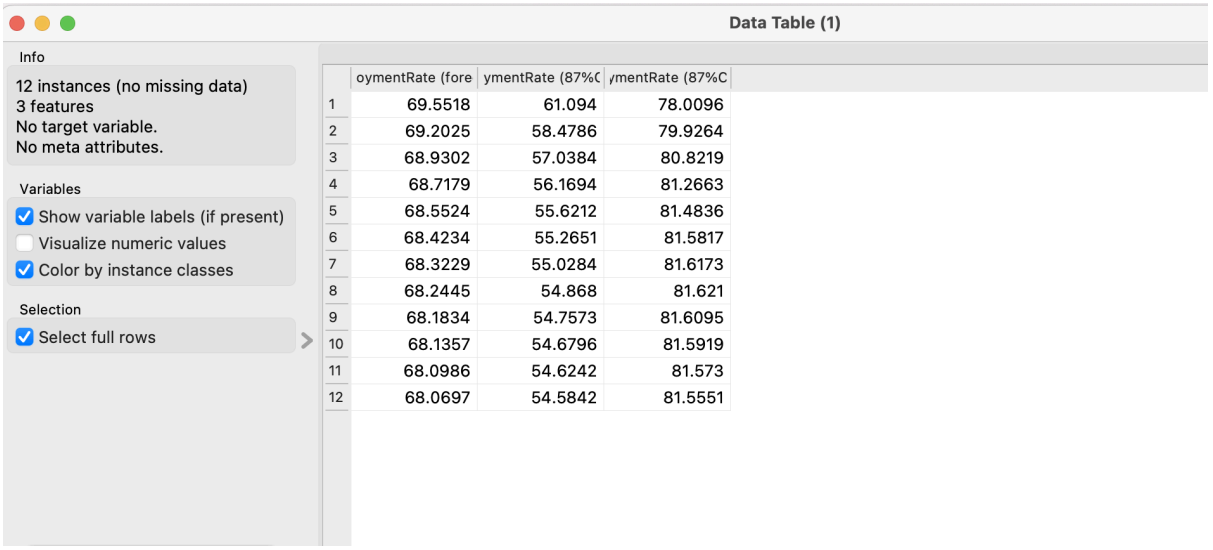
- Model may need differencing (d > 0) or tuning other (p, q) values.



Employment Line Chart with Forecast



Evaluation Table



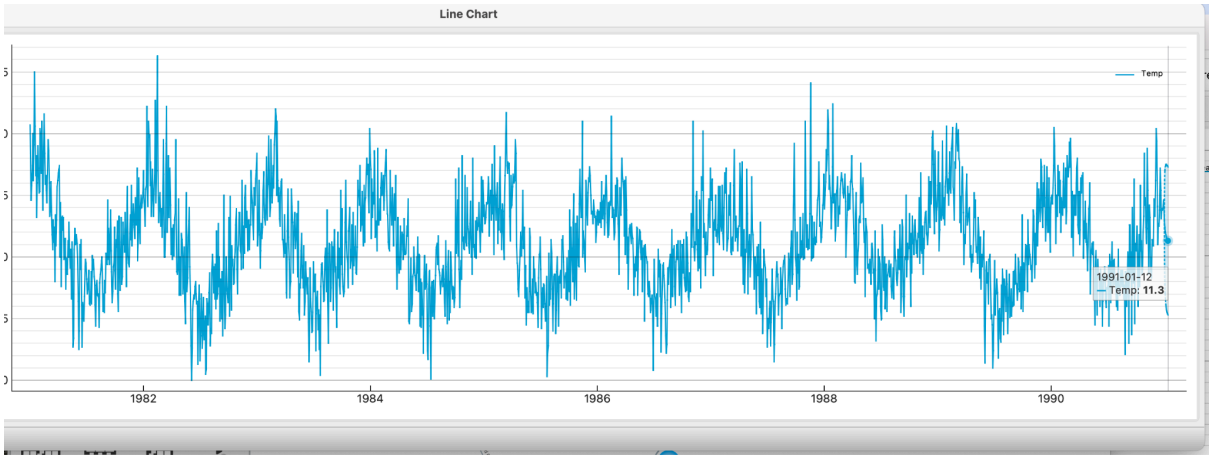
ARIMA Forecast

Dataset 3: Daily Minimum Temperature (Australia)

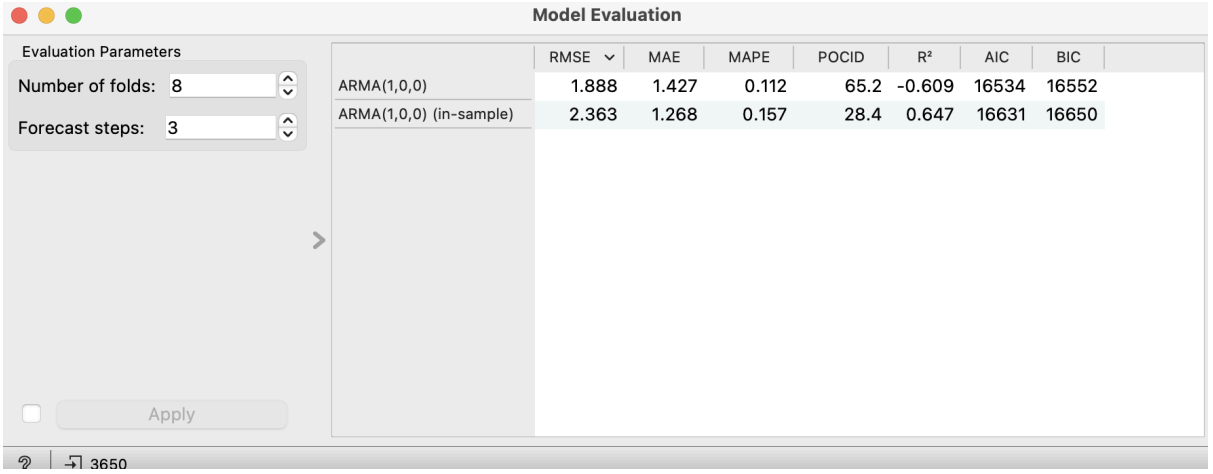
- **ARIMA(1,0,0)**, Forecast 3 steps
- *Used daily min temp dataset from CSV*

Observations:

- The model had moderate error scores but a negative R^2, indicating overfitting or trend misalignment.
- POCID of 65.2% shows decent directional accuracy.
- Adding differencing might stabilize the variance.

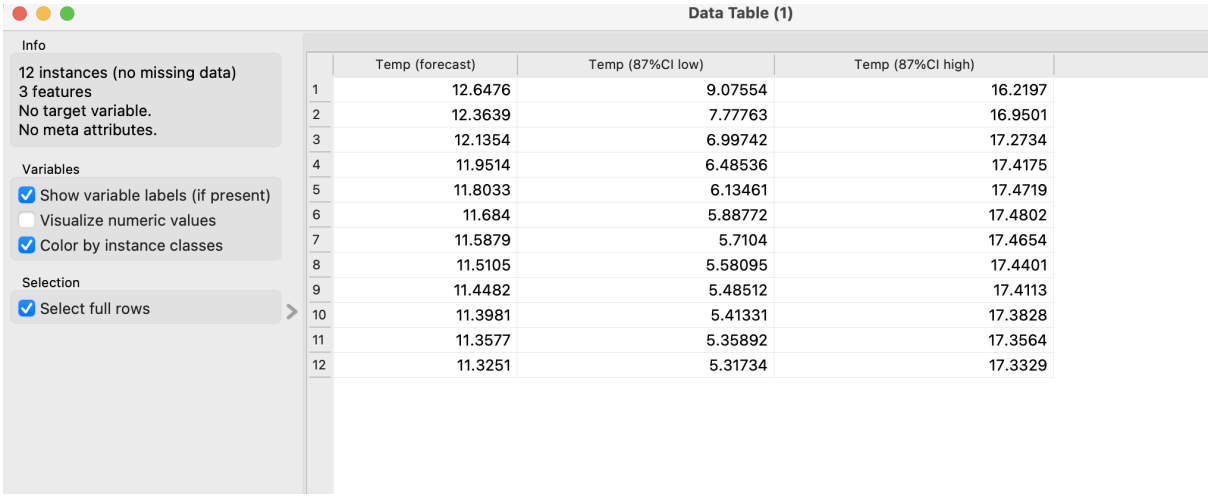


Min Temperature Line Chart



	RMSE	MAE	MAPE	POCID	R^2	AIC	BIC
ARMA(1,0,0)	1.888	1.427	0.112	65.2	-0.609	16534	16552
ARMA(1,0,0) (in-sample)	2.363	1.268	0.157	28.4	0.647	16631	16650

Evaluation Metrics



AIRM Forecast

Applications

- **Classification**

  The temperature classification task, where data was labeled as "Warm" or "Cool" using computed rules, has applications in environmental monitoring and seasonal analysis. Classifying temperature ranges can help identify climate trends, support agricultural planning, and assist energy consumption forecasting based on seasonal demands.

- **Clustering**

  Clustering enabled the identification of hidden patterns within meteorological data. Each algorithm—Hierarchical and K-Means—grouped instances with similar weather characteristics such as temperature, humidity, and radiation. These cluster-based groupings can be used for automated weather station profiling, regional climate zoning, and climate anomaly detection.

- **Time Series Forecasting**

  The ARIMA-based forecasting of temperature and employment rate provided insights into short-term trends. Temperature forecasting assists in climate adaptation strategies, public health alerts, and infrastructure planning. Employment trend forecasting can help policy makers, economists, and labor departments anticipate job market shifts and plan accordingly.

Limitations and Challenges

Classification

- Limited control over hyperparameter tuning

- No built-in support for advanced model customization
- Fewer model interpretability tools (e.g., SHAP, feature importance)

Clustering

- Manual K selection for K-Means — lacks auto-optimal detection
- Visualizations can lag with large datasets
- Limited options for density-based clustering (e.g., basic DBSCAN only)

Time Series Forecasting

- ARIMA model setup is manual with no auto-selection of (p,d,q)
- Lacks support for modern models like SARIMA, Prophet, or LSTM
- No seasonality detection or decomposition features

Conclusion

This project successfully applied Orange to perform classification, clustering, and time series forecasting across two real-world datasets: weather and employment. The classification component categorized temperature data into understandable labels that could assist in further prediction tasks. Each task was carried out using Orange's visual interface, highlighting the platform's ability to support rapid development and interpretation of machine learning workflows. The results obtained were realistic, interpretable, and offered practical insights into temporal behavior and data structure.

References

Cite all sources, including research papers, official documentation, and case studies. Use a standard referencing style (APA, IEEE, etc.).