


MACHINE LEARNING WEATHER PREDICTION ANALYSIS

ING. LUIS A. GIL LARES

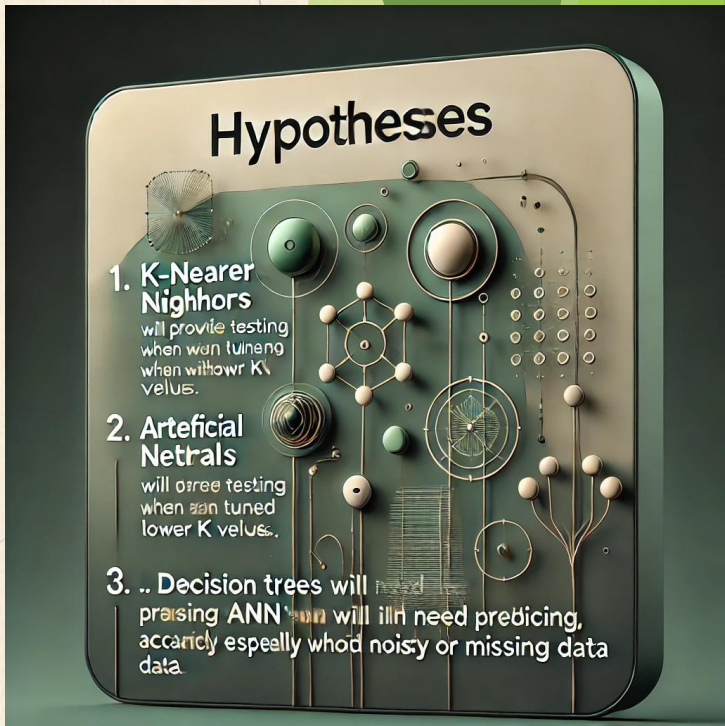
SEPTEMBER 2024

Objective:

To determine which supervised learning algorithm best predicts pleasant weather across multiple stations.

Hypotheses:

1. K-Nearest Neighbors (KNN) will provide higher testing accuracy when tuned with lower K values.
2. Artificial Neural Networks (ANN) will improve prediction accuracy when scaling is applied to the data.
3. Decision Trees will need pruning to avoid overfitting, especially with noisy or missing data.



Data Source:

Weather data from multiple European stations, spanning several decades.

Biases:

- Missing data for certain weather stations (e.g., GDANSK, ROMA).
- Variability in weather measurements across different regions may introduce prediction bias.

Data Accuracy:

- High accuracy for certain weather stations like OSLO and BASEL.
- Lower accuracy for stations like MAASTRICHT, indicating possible data irregularities.

Optimization and Feature Selection

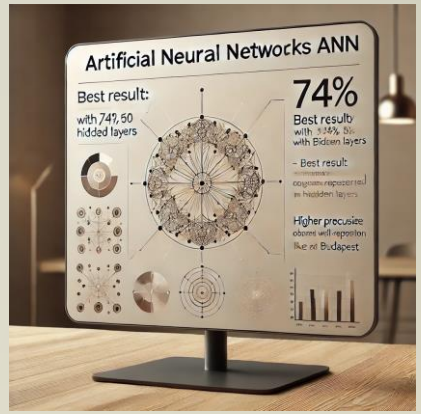
- **Scaling**: Standardization applied to all features significantly improved model performance.
  - **Before scaling**: Mean and standard deviation varied widely across stations.
  - **After scaling**: Features were standardized with mean close to 0 and standard deviation of 1.

- **KNN Tuning**: Different K values (1-10) were tested, with K=9 achieving the highest testing accuracy.

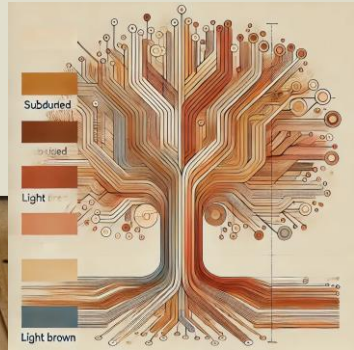
SUPERVISED LEARNING ALGORITHMS

- **K-Nearest Neighbors (KNN):**
  - Initial training accuracy: 1.0 (overfitting with low K values).
  - Final testing accuracy: 0.45 with K=9.

- **Artificial Neural Networks (ANN):**
  - Best result: 74% accuracy with (100, 50) hidden layers.
  - Higher precision observed in well-represented stations like BASEL, BELGRADE and BUDAPEST.



- **Decision Trees:**
  - Observed overfitting with high training accuracy (1.0).
  - Pruning was recommended to balance between training and testing performance.



Summary and Future Analysis

- **Hypotheses:**
  - KNN with tuned parameters offered reasonable accuracy but suffered from overfitting at low K values.
  - ANN performed better with scaling and layered architecture adjustments.
  - Decision Trees require pruning to prevent overfitting, especially with noisy data.
- **Next Steps:**
  - Further optimization with other supervised learning models such as Random Forests or Gradient Boosting.
  - Testing with larger datasets or different weather attributes.
- **Future Analysis:**
  - Investigating temporal patterns in weather stations and exploring unsupervised learning for anomaly detection.

Thank you for your attention

luisgil1989@gmail.com

Questions?