Data Analytics Project

A Data-Driven Approach to Weather Pattern Clustering & Rainfall Forecasting

Using Australian Meteorological Data

Ella Lee

Table of Content

Project Overview

- **Project Type:** Data Science & Machine Learning
- **Tools Used:** Python, pandas, seaborn, scikit-learn, KMeans, Random Forest, Neural Network
- **Target Users:** Agricultural managers, logistics planners, meteorological agencies, urban operations departments
- **Objective:** Utilize weather and agricultural data to support demand forecasting and optimize operational efficiency

01 Project Overview

01 Background

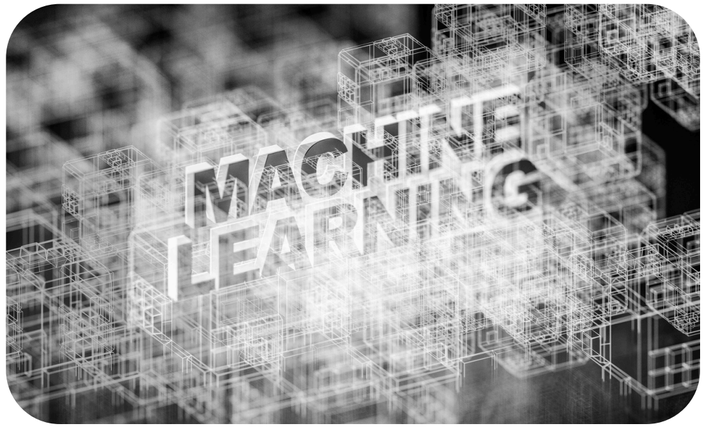
Extreme weather events such as **sudden rain or drought** are becoming more frequent due to climate change, posing serious risks to agriculture, logistics, and infrastructure.

02 Business Impact

Unpredictable rain causes crop loss, delivery delays, and energy imbalance. Current forecasts lack **local pattern detection**

03 Project Objective

Analyze weather patterns, group similar conditions, and build a model to predict rain for better **decision-making**



02 Data Description

Australian Government
Bureau of Meteorology

Dataset Overview

- Australian weather dataset with daily records across multiple regions
- **24 variables** spanning temperature, wind, humidity, pressure, and rainfall
- Structured for clustering and supervised prediction using mixed data types

---

Feature type

- **Ratio:** Temperature, Rainfall, Pressure, WindSpeed, Evaporation
- **Interval:** Humidity, Sunshine
- **Ordinal:** Cloud cover (oktas)
- **Nominal:** Date, Location, Wind Direction, RainToday, RainTomorrow

---

Focus Variables

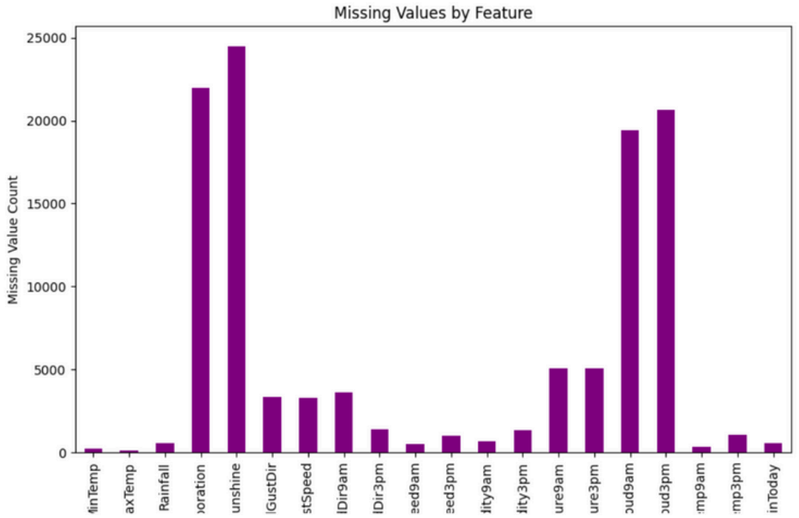
- **Target:** RainTomorrow – binary label for next-day rainfall prediction
- **Key Predictors:**

– Temperature (MaxTemp, Temp3pm) → thermal trends
– Humidity (Humidity3pm) → strongest correlation
– Rainfall → historical patterns
– Wind (WindGustSpeed, Direction) → storm indicators

03 Preprocessing

01 Handling Missing Values

02 Imputation Methods



Data Type	Imputation Method	Example Features	Reason
Numerical Features	Median	MaxTemp, Rainfall	Robust to outliers
Categorical Features	Mode	WindDir9am,RainToday	Preserve category consistency

```
No missing values in numerical features after imputation.
No missing values in categorical features after imputation.
```

- Numerical features (e.g., MaxTemp, Rainfall):
  → Imputed using **median**, which is robust against skewed distributions and outliers.
- Categorical features (e.g., WindDir9am, RainToday):
  → Filled using **mode**, preserving label consistency across groups.

03 Preprocessing

“These transformations ensured comparability, reduced feature complexity, and improved model convergence”

04 Normalization

Min-Max scaling & Z-score standardization for stable and comparable features.

03 Binning

① Equal-width binning

② Equal-depth binning

05 Discretization

Grouped wind speeds into discrete bands
(0–15–30–45 km/h)

→ captures wind intensity levels and simplifies input variability

06 Binarization

Converted wind direction to binary feature

→ 1 if direction is North/NNE/NE/ENE, 0 otherwise

→ Reduces dimensionality while capturing key wind behavior

Scaling ensures fair contribution of features and improves convergence in model training

Multiple binning strategies enhance data interpretability

Wind speed grouping optimizes analysis and model inputs.

Binarization simplifies directionality and focuses on high-impact directional patterns

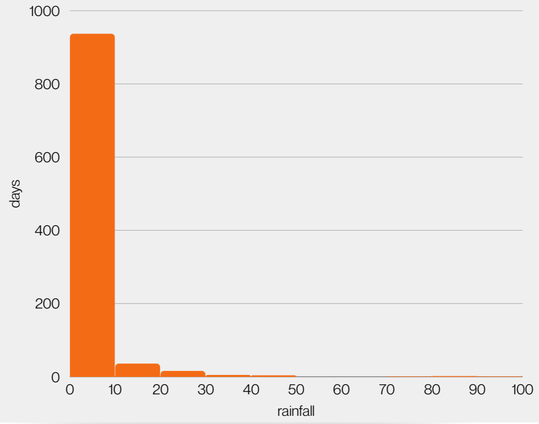
04 EDA Summary

WindGustSpeed :

40km/h

- Average 40km/h
- Max 104km/h

Rainfall :

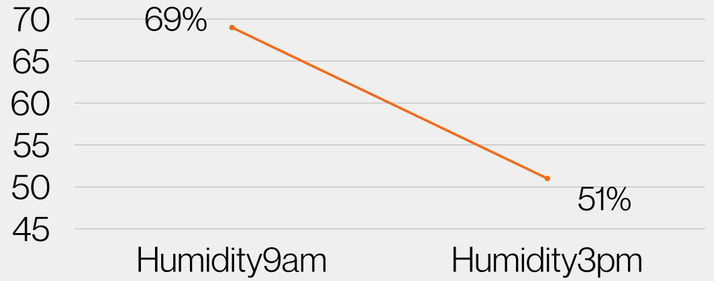


0~2 mm

70% of days < 2mm rainfall
→ Dry Days Dominant

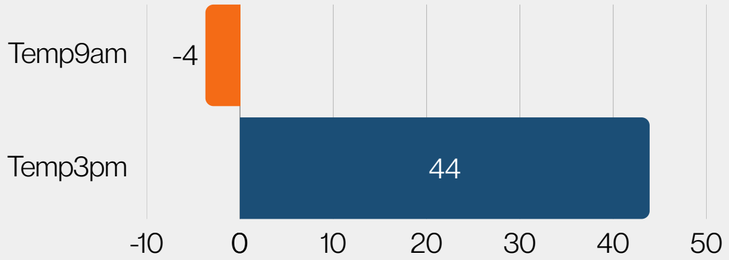
Humidity Change :

9am : 69%
→ 3pm : 51%



Temperature Range (°C) :

Min : -3.7°C
Max : 43.9°C



EDA-Driven Understanding

Visual exploration of key weather variables revealed distinct behavioral patterns across temperature, humidity, and rainfall.
These findings serve as the foundation for deeper correlation analysis and clustering, ultimately enhancing predictive modeling for rainfall.

Rainfall

- IIndicates a dry climate pattern, causing class imbalance in RainTomorrow
- Needs to be addressed during model training (e.g., resampling, weighted loss)

Humidity

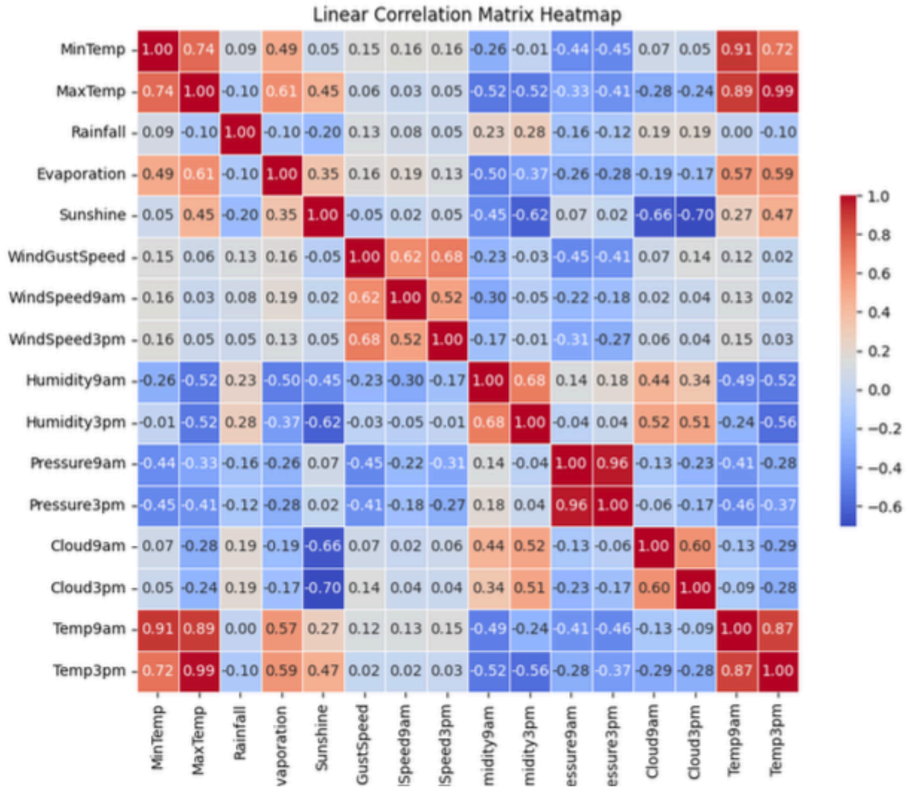
- Suggests inverse relationship with temperature
- Indicates high evaporation activity during daytime, which is key in climate-based segmentation

Temperature

- Temp3pm is nearly equivalent to MaxTemp
- Allows for feature simplification and helps prioritize variables in clustering and prediction

05 Correlation Analysis

Pearson correlation was used to explore how weather variables relate.
Key relationships guided variable grouping for clustering.



Temperature Variables :

- Strong positive correlation (0.74~0.99)

→ Used as one cluster feature: consistent daily temp trend

Humidity <--> Temperature:

- Negative correlation (-0.56)

→ Basis for identifying dry vs. humid patterns

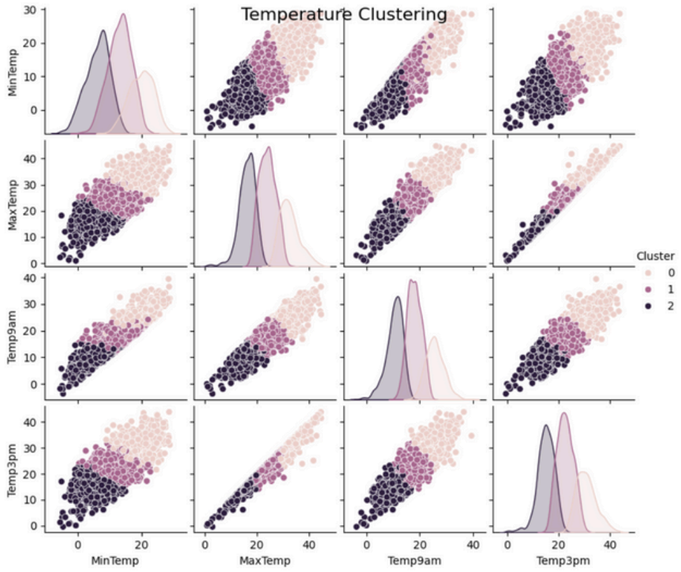
Wind Speed <--> Direction

- Near-zero correlation

→ Wind direction excluded from clustering

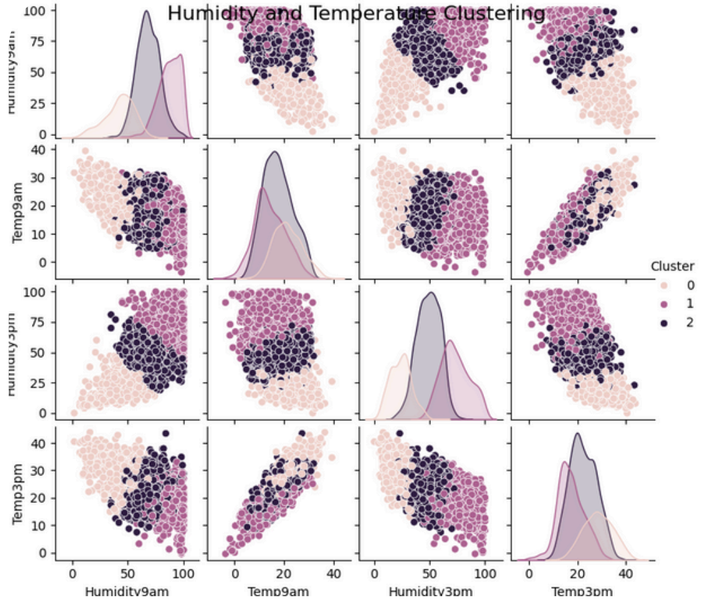
06 Clustering

K-means clustering on three variables revealed meaningful weather patterns, guided by prior correlation analysis



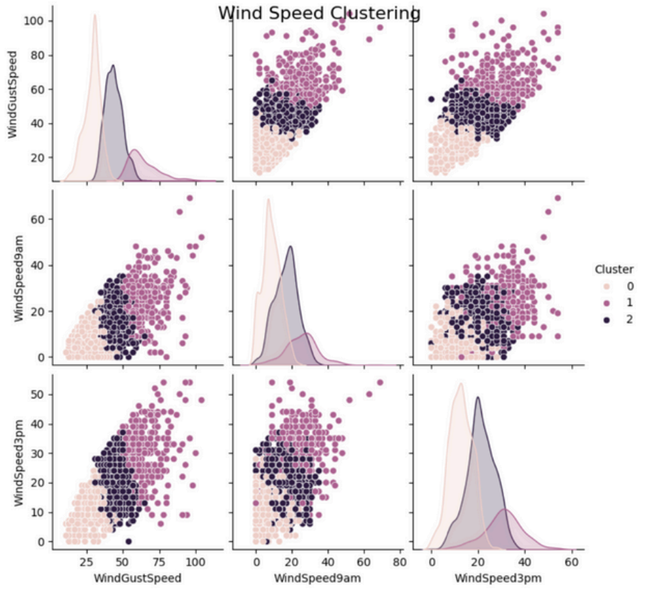
01 Temperature

- Clusters distinguish typical vs. extreme heat
- Supports seasonal planning and energy demand estimation



02 Humidity

- Identifies humid/dry periods via daily gap
- Useful for irrigation, comfort index, and crop risk



03 Wind Speed

- Detects gust-heavy days with operational risk.
- Informs logistics, aviation, and infrastructure safety

07 Modeling & Evaluation

Goal:

- Use weather variables to build predictive models
- Assess model performance for rainfall classification

Class Imbalance Handling

Since dry days dominate the dataset, class weighting was applied during model training to ensure balanced learning and improve rain prediction accuracy.

Steps:

1. Data Partitioning
2. Hyperparameter Tuning
3. Model Comparison & Final Evaluation

07 Modeling Overview

Data Partitioning

Split the dataset into **training, validation, and test sets** to prevent overfitting and ensure robust model evaluation.

Hyperparameter Tuning

Used **grid search** and **cross-validation** to optimize model parameters and improve generalization

Model Performance Comparison

Evaluated and compared multiple models using **consistent metrics** such as accuracy, precision, recall, and F1-score.

Final Model Selection & Interpretation

Selected the **best-performing model** based on evaluation results and interpreted its logic using confusion matrix and loss curves

07 Modeling

To develop a robust and generalizable predicive model, partitioning the dataset into training, validation, and testing sets is crucial. This process ensures the model is evaluated on unseen data and is not overfit to the training set.

Training Set

70%

Used to train the model and learn data patterns

Validation Set

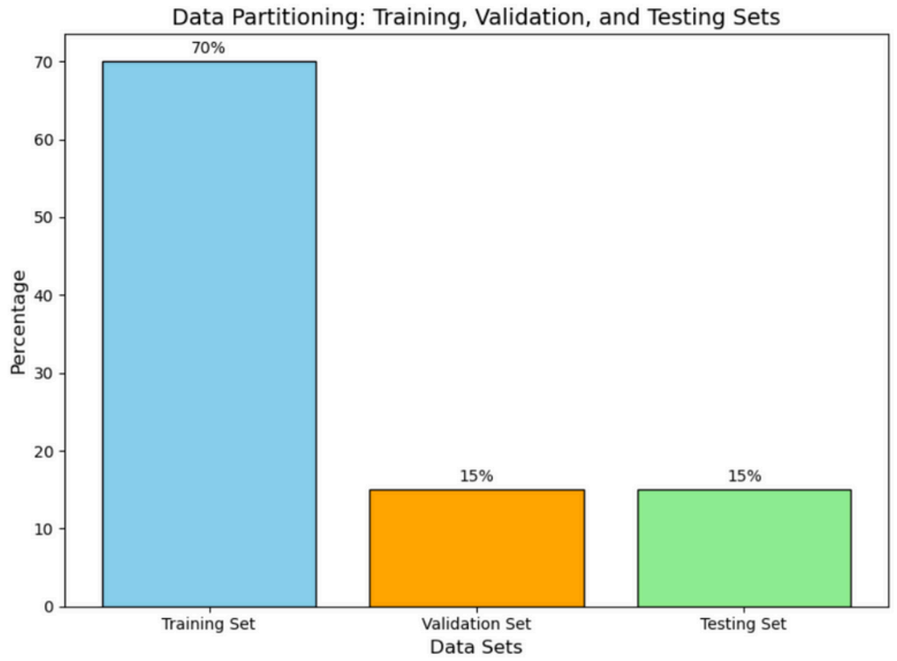
15%

Used to tune hyperparameters and prevent overfitting

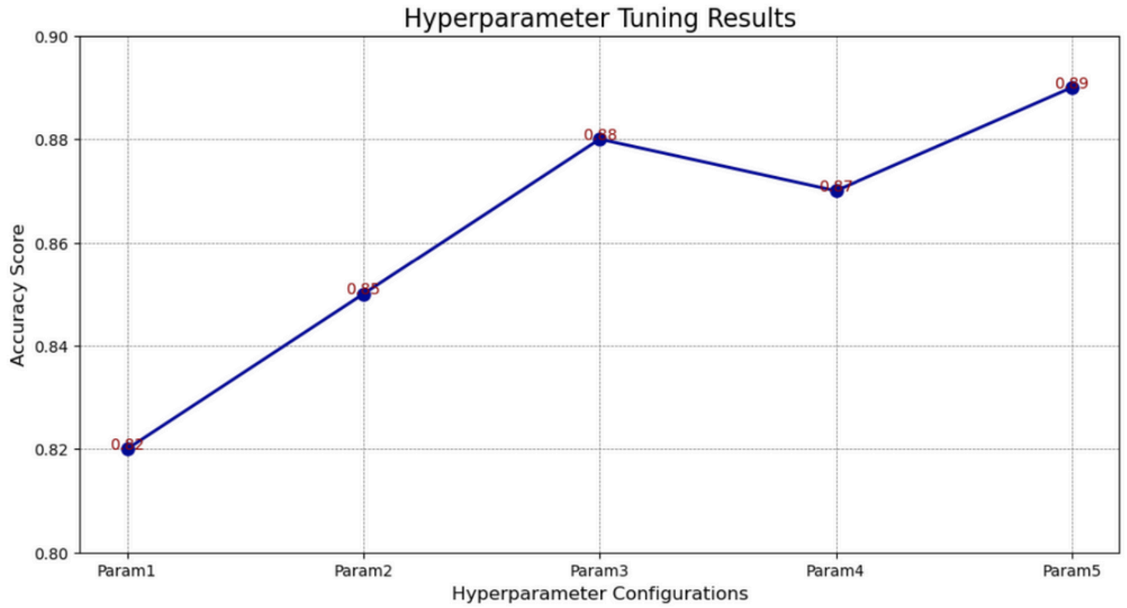
Testing Set

15%

Used for final performance check on unseen data



07 Modeling



Grid Search

A predefined range of hyperparameter values was explored to systematically identify the **optimal configuration**, ensuring model **consistency** without trial-and-error

Cross- Validation

k-Fold cross-validation was applied to validate model robustness across data splits and **prevent overfitting**

Evaluation Metrics

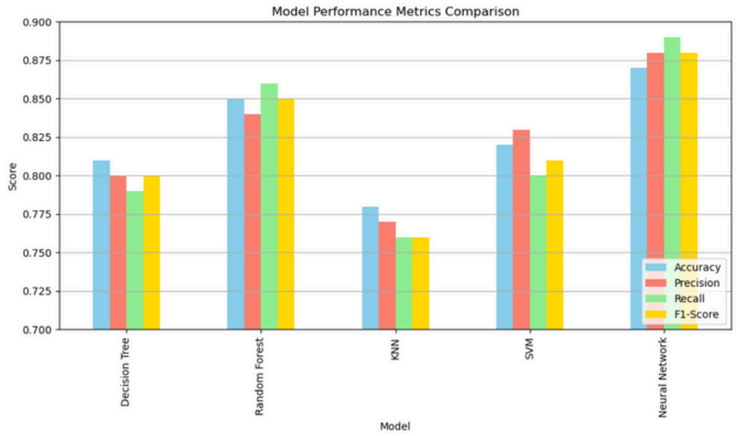
Accuracy, Precision, Recall, and **F1-Score** were used as consistent metrics across models for fair comparison.
These measures ensured both **correctness** and **balance** in performance evaluation.

07 Modeling

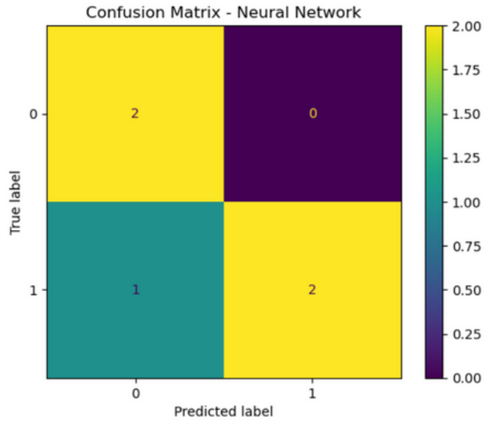
Model Comparison

5 classification models were trained and evaluated using key metrics

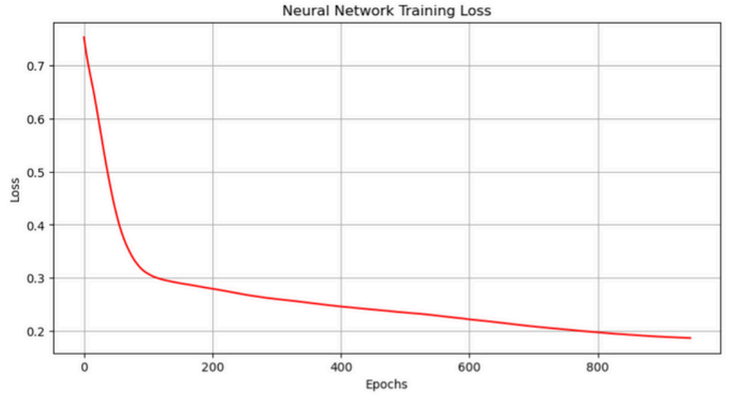
Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.81	0.80	0.79	0.80
Random Forest	0.85	0.84	0.86	0.85
KNN	0.78	0.77	0.76	0.76
SVM	0.82	0.83	0.80	0.81
Neural Network	0.87	0.88	0.89	0.88



Confusion Matrix & Training Loss



&



Both positive and negative classes were well separated, confirming strong predictive balance

The loss decreased smoothly and stabilized with no sign of overfitting

The neural network consistently outperformed other models across all metrics, showing the best balance of precision and recall. Future improvements could include SHAP or permutation importance to enhance interpretability.

08 Conclusion

This conclusion highlights the project's achievements, data-driven impact, and actionable next steps for weather-related decision making

Achievements & Strengths

- Built **a robust weather prediction model** with high accuracy.
- Achieved strong results across all metrics.
- Delivered **actionable insights** on temperature, humidity, and wind speed patterns.

Limitations & Challenges

- Weather data contained missing values and noisy outliers, requiring careful preprocessing (e.g., MICE imputation).
- Some variables showed multicollinearity, affecting interpretability.
- Lack of labeled data limited clustering validation and interpretability.

Business/Application Opportunities

Use Cases

- Agriculture: Forecast-informed irrigation
- Logistics: Route optimization to prevent rain-related delays
- Public safety: Early alerts for extreme weather

System Advantages

- Resource allocation before rainfall events
- Real-time deployment with streaming data