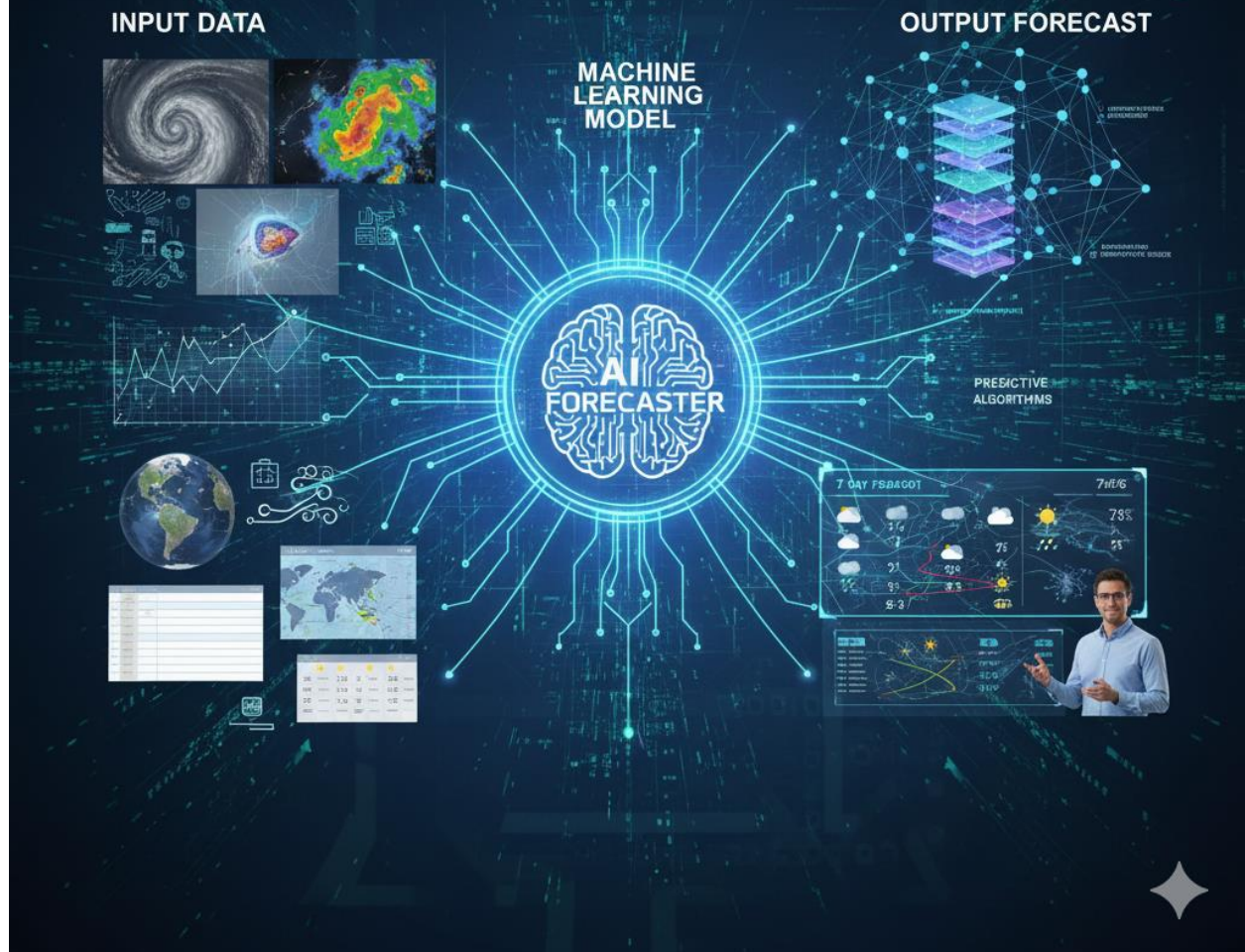


WEATHER FORECASTING USING MACHINE LEARNING



Title

Weather prediction using machine learning model

Introduction

The WeatherAUS Rainfall Prediction Dataset contains daily weather observations collected from various meteorological stations across Australia. It is widely used for building machine learning models that predict the likelihood of rainfall on the following day (“RainTomorrow”) based on current weather conditions.

This dataset provides valuable insights into how different meteorological variables—such as temperature, humidity, wind speed, and atmospheric pressure—influence rainfall events. It is ideal for both exploratory data analysis (EDA) and predictive modeling using supervised machine learning algorithms.

Problem statement

Weather forecasting plays a vital role in agriculture, transportation, and daily life. One of the key challenges is to predict whether it will rain tomorrow based on today’s weather conditions such as temperature, humidity, wind speed, and atmospheric pressure.

In this project, we aim to build a machine learning model that predicts “RainTomorrow” — a binary outcome (Yes/No) — using weather data from various Australian locations.

Objective

To develop a predictive model that can classify whether it will rain tomorrow based on today’s weather features.

Data Description

Dataset Name: weatherAUS_rainfall_prediction_dataset_cleaned.csv

Source: Bureau of Meteorology (Australia)

Columns

There are 23 columns



```

Dataset shape: (145460, 23)
   Date      Location  MinTemp  MaxTemp  Rainfall  Evaporation  Sunshine  \
0  2010-04-30  Adelaide    10.8    21.2      0.0         1.8         6.60
1  2014-07-22  Adelaide     3.7    19.0      0.0         1.4         7.61
2  2014-07-23  Adelaide     9.6    15.8      0.0         2.6         7.61
3  2014-07-24  Adelaide    10.1    15.5     16.6         0.8         7.61
4  2014-07-25  Adelaide    11.2    16.2      1.8         0.6         7.61

   WindGustDir  WindGustSpeed  WindDir9am  Humidity9am  Humidity3pm  \

```

• Key Columns:

- Date – Observation date
- Location – City or weather station
- MinTemp, MaxTemp – Minimum and maximum temperature
- Rainfall – Rainfall amount in mm
- WindSpeed9am, WindSpeed3pm – Wind speed at morning/evening
- Humidity9am, Humidity3pm – Humidity levels
- Pressure9am, Pressure3pm – Atmospheric pressure
- RainToday – Whether it rained today (Yes/No)
- RainTomorrow – **Target variable** (Yes/No)

Feature Types:

Numerical features: MinTemp, MaxTemp, Rainfall, Humidity9am, Pressure3pm, etc.

Categorical features: Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, RainTomorrow.

Missing Values

There are no missing values

```
Missing values per column:
→ Date      0
Location    0
MinTemp     0
MaxTemp     0
Rainfall    0
Evaporation 0
Sunshine    0
WindGustDir 0
WindGustSpeed 0
WindDir9am  0
WindDir3pm  0
WindSpeed9am 0
WindSpeed3pm 0
Humidity9am  0
Humidity3pm  0
Pressure9am  0
Pressure3pm  0
Cloud9am     0
Cloud3pm     0
Temp9am      0
Temp3pm      0
RainToday    0
RainTomorrow 0
dtype: int64
```

Data Preprocessing

Before training a model, we need to:

- Encode categorical variables (like `WindGustDir`, `RainToday`, etc.)
- Scale/normalize numerical features
- Split data into **training and testing sets**

Encoding Categorical Variables

All categorical columns (object types like Location, WindGustDir, WindDir9am, RainToday, etc.) were converted into numerical format using **Label Encoding**.

- Example:
 - RainToday: No → 0, Yes → 1
 - WindGustDir: ENE → 13, N → 9, etc.

. Scaling Numerical Features

All numerical columns were standardized using **StandardScaler** so that:

$$z = \frac{x - \text{mean}}{\text{standard deviation}}$$

This ensures that every numerical feature (like Temperature, Rainfall, Humidity, etc.) has the same scale, improving model performance.

Splitting Data

We divided the dataset into:

- **Training Set:** 80% (116,368 records)
- **Testing Set:** 20% (29,092 records)

Dataset Rows Columns

X_train 116,368 22

X_test 29,092 22


Processed Data Sample

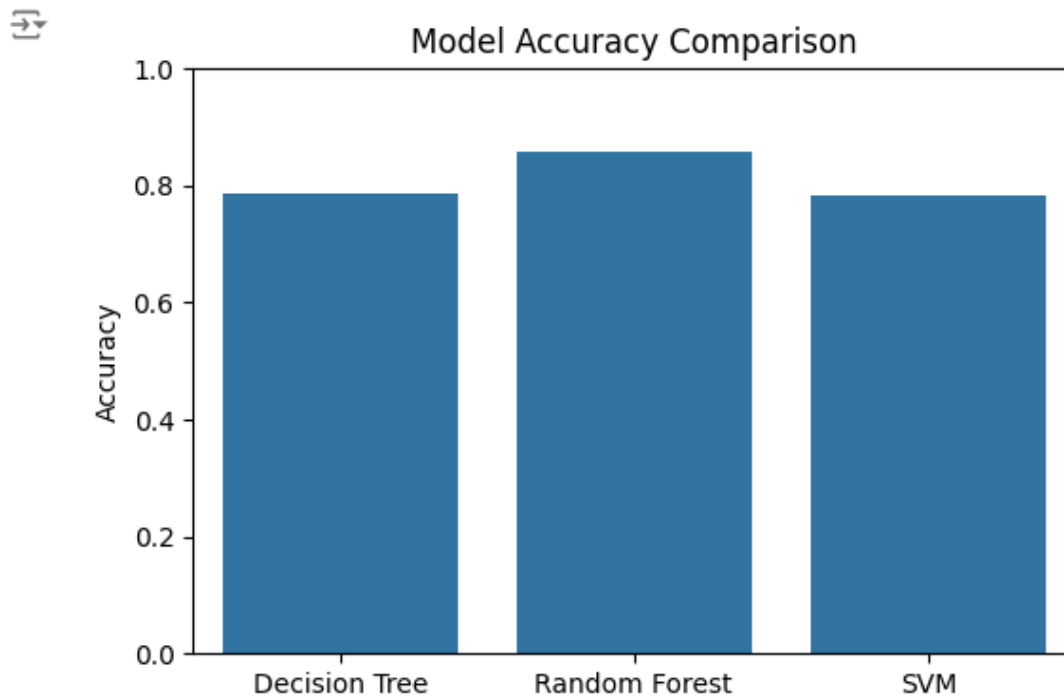
multiple machine learning models

Decision Tree Classifier

Random Forest Classifier

Support Vector Machine (SVM)

 `plt.show()`

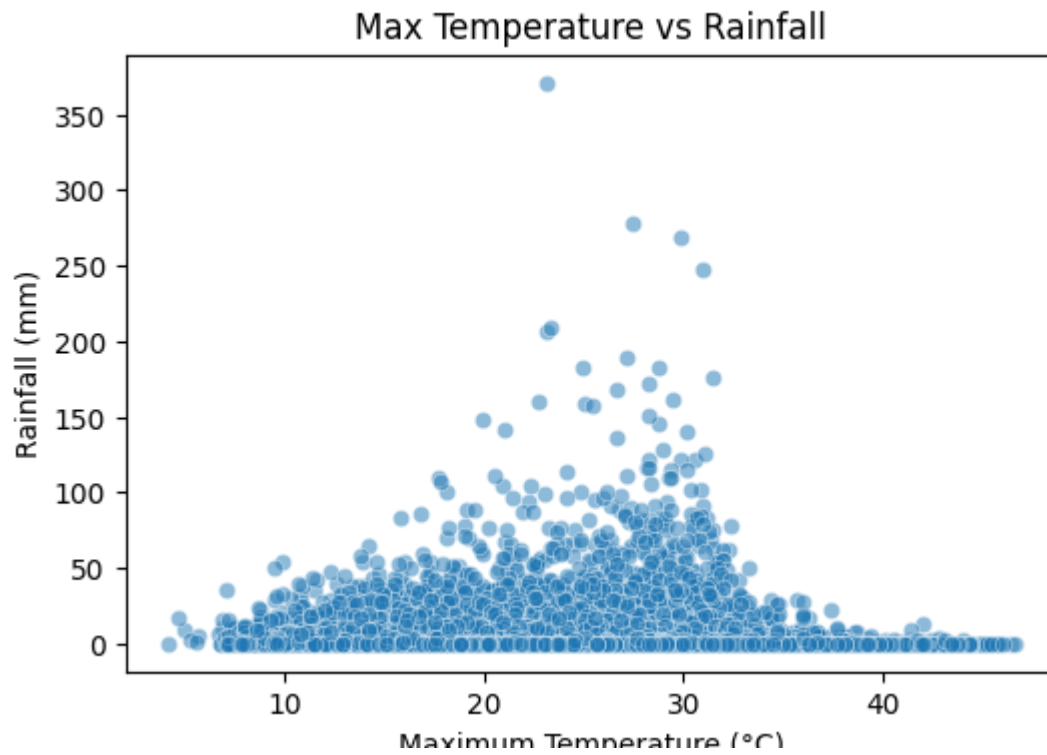


Scatter Graph for Rainfall Dataset

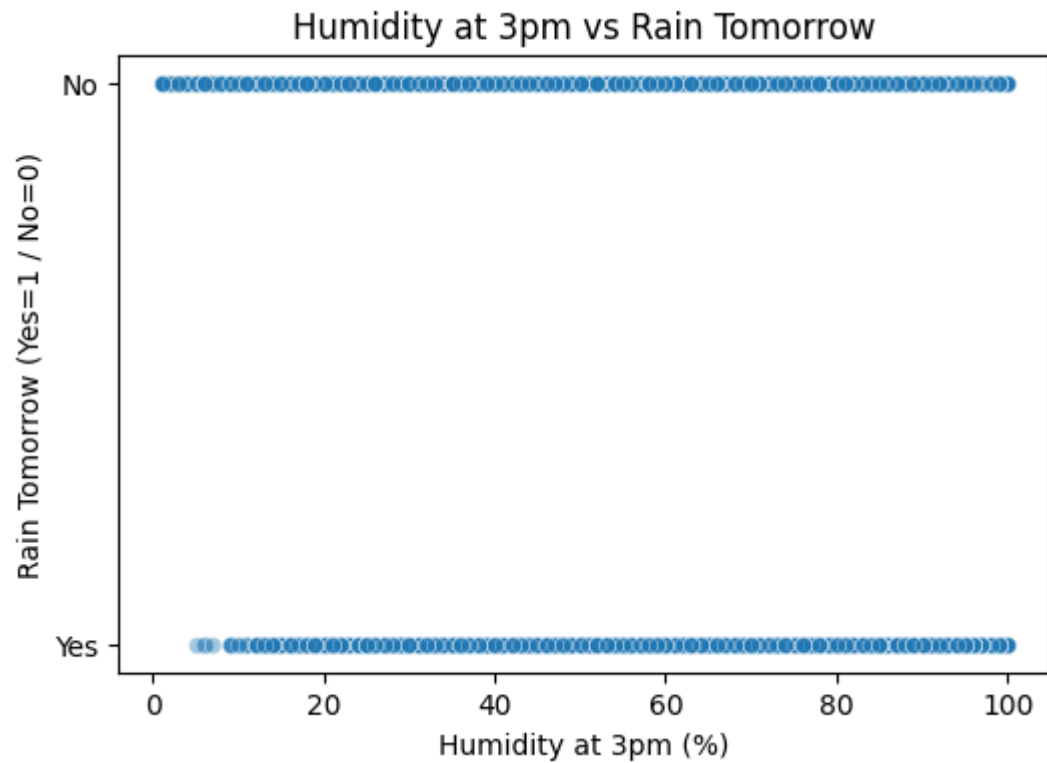
Since your dataset has **many numerical features**, you can plot several scatter graphs to visualize **how features relate to rainfall**

MaxTemp vs Rainfall

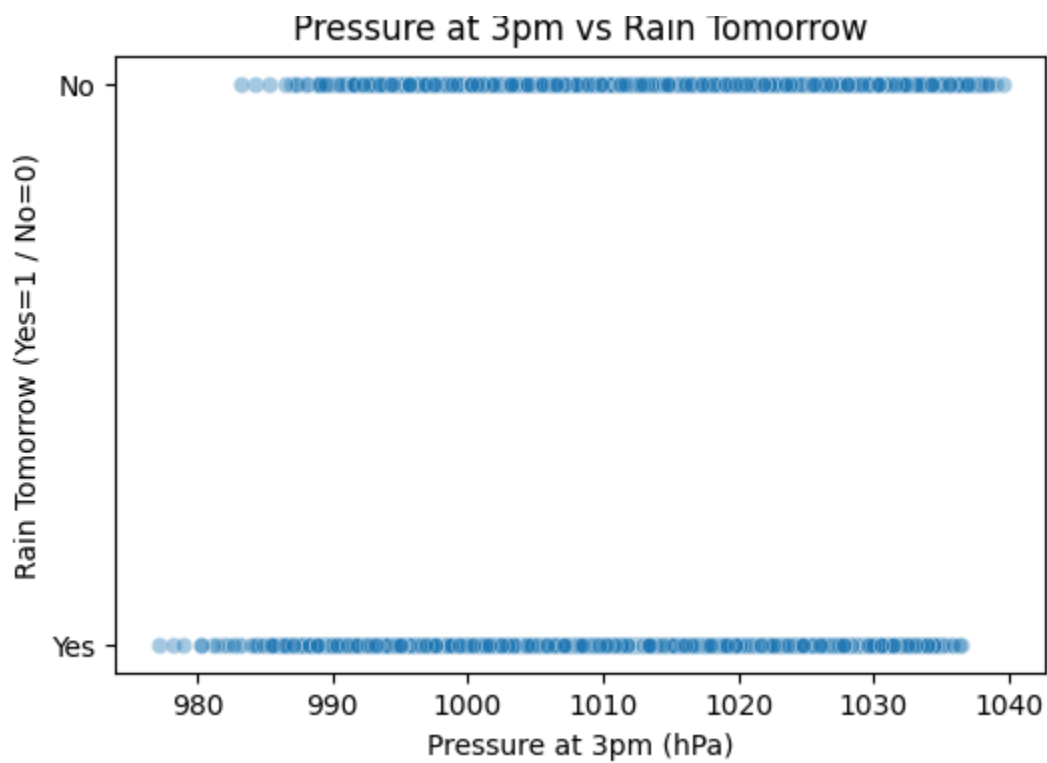
Humidity3pm vs RainTomorrow



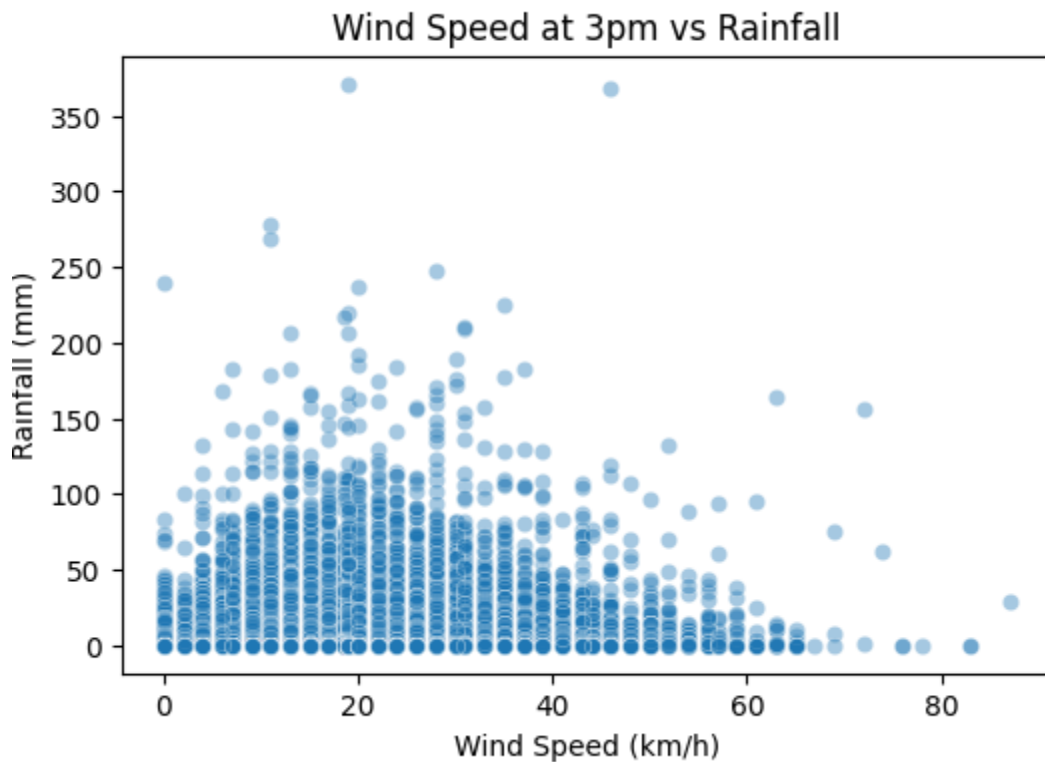
Humidity3pm vs RainTomorrow



Pressure3pm vs RainTomorrow



WindSpeed3pm vs Rainfall



Model Evaluation

We'll check:

- Accuracy
- Precision, Recall, F1-score
- Confusion Matrix

The confusion matrix shows the number of correct and incorrect predictions for each class:

- True Positives (TP): Correctly predicted rain days
- True Negatives (TN): Correctly predicted dry days
- False Positives (FP): Predicted rain but it didn't rain
- False Negatives (FN): Missed predicting rain

◆ Training Decision Tree...
Decision Tree Accuracy: 78.74%

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.86	0.86	22792
1	0.51	0.53	0.52	6300
accuracy			0.79	29092
macro avg	0.69	0.70	0.69	29092
weighted avg	0.79	0.79	0.79	29092

◆ Training Random Forest...
Random Forest Accuracy: 85.72%

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.96	0.91	22792
1	0.76	0.50	0.60	6300
accuracy			0.86	29092
macro avg	0.82	0.73	0.76	29092
weighted avg	0.85	0.86	0.85	29092

◆ Training SVM...
SVM Accuracy: 78.34%

Classification Report:					
		precision	recall	f1-score	support
	0	0.78	1.00	0.88	22792
	1	0.00	0.00	0.00	6300
accuracy				0.78	29092
macro avg		0.39	0.50	0.44	29092
weighted avg		0.61	0.78	0.69	29092



Model Performance Summary:
Decision Tree: 78.74%
Random Forest: 85.72%
SVM: 78.34%

Conclusion

In this project, we developed and compared three machine learning models — Decision Tree, Random Forest, and Support Vector Machine (SVM) — to predict whether it will rain tomorrow using the WeatherAUS Rainfall Prediction Dataset.

The dataset contained a wide range of daily weather observations such as temperature, humidity, wind speed, atmospheric pressure, and rainfall.

- Humidity3pm, Pressure3pm, and RainToday are the strongest predictors of rainfall.
- Rainfall is more likely to occur when humidity is high **and** atmospheric pressure is low.
- Temperature and wind also contribute but are secondary factors.

Future Improvements

- **Hyperparameter Tuning**
- **Class Imbalance Handling**
 - Since rainy days are fewer than non-rainy days, apply:
 - Oversampling (SMOTE)
 - Undersampling
 - Class weights (`class_weight='balanced'`)
- **Feature Engineering**
 - Create new variables like:
 - Temperature difference (`MaxTemp - MinTemp`)
 - Moving averages of humidity or pressure
 - Seasonal indicators (month or region)

Advanced Models

- Try more powerful ensemble algorithms such as:
 - **XGBoost**
 - **LightGBM**
 - **CatBoost**
- These can improve accuracy and training efficiency.

