

**LAPORAN TUGAS BESAR
KECERDASAN BUATAN**

**PREDIKSI KONDISI CUACA BERDASARKAN DATA HISTORIS
MENGUNAKAN ALGORITMA RANDOM FOREST**



Disusun oleh:

Faiz Pratama– 2306140

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

**INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
TAHUN AKADEMIK 2024/2025**

1. BUSINESS UNDERSTANDING

a) Permasalahan dunia nyata

Perubahan iklim global telah menyebabkan ketidakstabilan dalam pola cuaca. Kondisi seperti hujan tiba-tiba, suhu ekstrem, dan badai lokal sulit diprediksi hanya dengan metode konvensional. Hal ini berdampak besar terhadap sektor-sektor penting seperti pertanian, transportasi, penerbangan, dan konstruksi. Masalah utama yang dihadapi:

Masalah utama yang dihadapi adalah keterbatasan akurasi dan kecepatan dalam memprediksi cuaca secara real-time menggunakan metode manual atau berbasis rumus fisika saja. Model prediksi tradisional sering kali tidak mampu memproses data besar secara efisien atau menyesuaikan dengan pola historis yang kompleks.

b) Tujuan Proyek :

Membangun sistem prediksi kondisi cuaca berdasarkan data historis menggunakan algoritma machine learning untuk meningkatkan akurasi dan efisiensi prediksi.

c) User/Pengguna Sistem

- Instansi meteorologi dan BMKG
- Aplikasi peramal cuaca
- Petani, nelayan, manajer proyek konstruksi
- Operator bandara dan transportasi publik

d) Manfaat Implementasi AI

- Prediksi cuaca yang lebih cepat dan akurat
- Meningkatkan kesiapsiagaan terhadap cuaca ekstrem
- Mengurangi risiko ekonomi dan keselamatan akibat prediksi cuaca yang salah
- Efisiensi dalam perencanaan operasional berbasis cuaca

2. DATA UNDERSTANDING

a) Sumber Data

Dataset diambil dari Kaggle, berjudul "*Weather History Dataset*". Deskripsi setiap fitur

b) Deskripsi setiap fitur

Berikut ini penjelasan dari setiap atribut atau kolom yang terdapat dalam dataset:

- **Formatted Date:** Berisi tanggal dan waktu pencatatan kondisi cuaca dalam format ISO 8601.
- **Summary:** Ringkasan kondisi cuaca seperti "Clear", "Partly Cloudy", "Overcast", "Rain", dan lainnya. Fitur ini merupakan target klasifikasi dalam proyek ini.

- **Precip Type:** Menunjukkan jenis presipitasi, seperti "rain" atau "snow". Beberapa entri memiliki nilai kosong (null).
- **Temperature (C):** Suhu aktual yang tercatat dalam satuan derajat Celsius.
- **Apparent Temperature (C):** Suhu yang dirasakan manusia (feels-like temperature), juga dalam derajat Celsius.
- **Humidity:** Tingkat kelembaban relatif, bernilai antara 0.0 hingga 1.0.
- **Wind Speed (km/h):** Kecepatan angin dalam kilometer per jam.
- **Wind Bearing (degrees):** Arah datangnya angin dalam satuan derajat, berkisar antara 0 hingga 360.
- **Visibility (km):** Jarak pandang atau seberapa jauh objek dapat terlihat dalam kilometer.
- **Pressure (millibars):** Tekanan atmosfer dalam satuan millibar.
- **Daily Summary:** Ringkasan naratif kondisi cuaca harian, biasanya dalam bentuk kalimat deskriptif

c) Ukuran dan Format Data

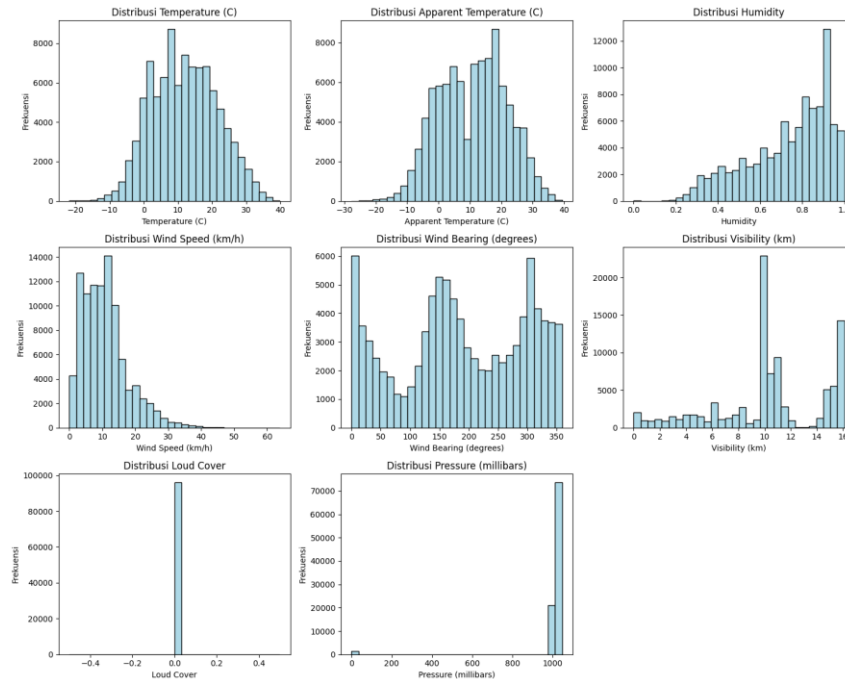
Dataset ini memiliki total 96.453 baris data dan terdiri dari 11 kolom (fitur). Format file adalah CSV (Comma-Separated Values).

d) Tipe Data dan Target Klasifikasi

- Fitur target adalah Summary, yaitu ringkasan kondisi cuaca seperti *Clear*, *Partly Cloudy*, *Rain*, dll.
- Tipe data asli dari Summary adalah **object/string**.
- Dalam proses pra-pemrosesan, Summary dikategorikan ulang menjadi tiga label utama untuk klasifikasi:
- **Clear**
- **Cloudy** (gabungan dari *Partly Cloudy*, *Mostly Cloudy*, *Overcast*, dll.)
- **Rain** (gabungan dari *Rain*, *Light Rain*, *Drizzle*, dll.)

3. *EXPLORATORY DATA ANALYSIS (EDA)*

a) Distribusi Data



Gambar 1. Distribusi data

Analisis Distribusi Data:

1. Distribusi Temperature (°C)

Bentuk: Hampir normal (bell-shaped).

Artinya: Suhu terbanyak berada di kisaran 5–20°C. Ekstrem dingin (< -10°C) dan panas (> 30°C) jarang terjadi.

2. Distribusi Apparent Temperature (°C)

Bentuk: Juga hampir normal, mirip dengan suhu sebenarnya.

Artinya: Suhu yang dirasakan manusia umumnya juga berada di kisaran 0–20°C. Kadang lebih dingin atau lebih hangat dari suhu sebenarnya.

3. Distribusi Humidity

Bentuk: Positif skew (condong ke kanan).

Artinya: Sebagian besar nilai kelembapan tinggi (mendekati 1 atau 100%). Udara cenderung lembap.

4. Distribusi Wind Speed (km/h)

Bentuk: Positif skew.

Artinya: Sebagian besar kecepatan angin rendah (0–10 km/h). Angin sangat kencang (>30 km/h) jarang terjadi.

5. Distribusi Wind Bearing (degrees)

Bentuk: Multimodal (beberapa puncak).

Artinya: Arah angin sering berasal dari beberapa arah dominan (misal sekitar 0°, 180°, 270°). Tidak merata dari semua arah.

6. Distribusi Visibility (km)

Bentuk: Multimodal, sangat tidak merata.

Artinya: Banyak pengamatan dengan visibilitas penuh (sekitar 10–16 km), tetapi juga ada yang sangat rendah (<5 km), mungkin karena kabut, hujan, atau polusi.

7. Distribusi Loud Cover

Bentuk: Hanya satu nilai (konstan).

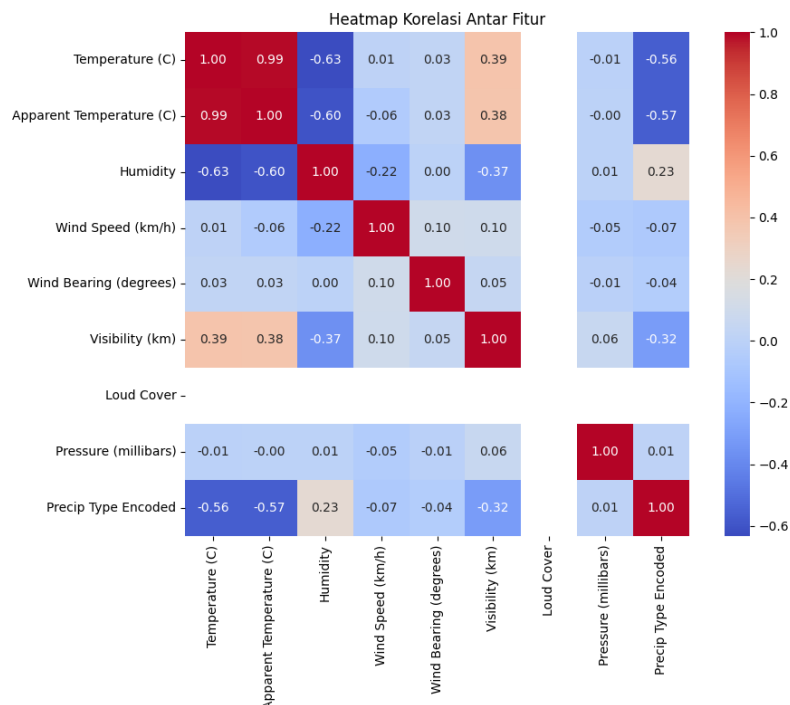
Artinya: Semua data memiliki nilai sama (kemungkinan 0), artinya fitur ini tidak memberikan variasi dan tidak berguna untuk analisis/pemodelan.

8. Distribusi Pressure (millibars)

Bentuk: Positif skew, sangat terpusat.

Artinya: Tekanan udara berkisar sempit di sekitar 1000 millibar, menunjukkan nilai standar atmosfer di permukaan laut. Outlier kecil di nilai rendah.

b) Analisis Korelasi

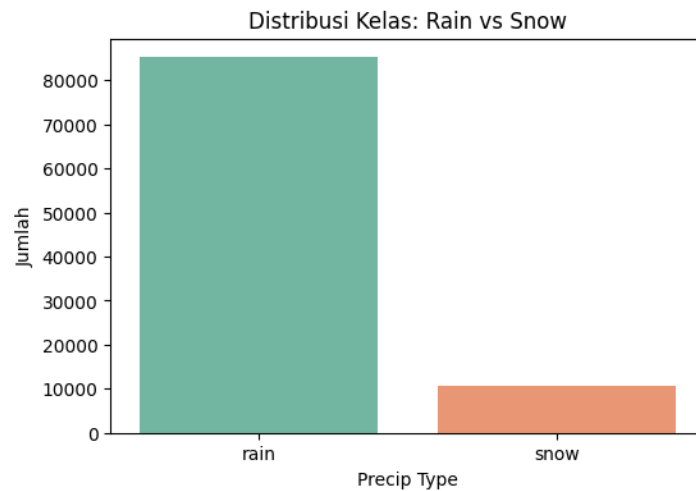


Gambar 2. Analisis Korelasi

Heatmap ini menggambarkan hubungan linear antar variabel numerik di dataset, dengan nilai korelasi Pearson berkisar antara -1 sampai 1.

1. Temperature (C) vs Apparent Temperature (C):
 - Korelasi = 0.99 → hampir identik (sangat tinggi).
 - Artinya suhu aktual dan suhu terasa sangat mirip (seperti yang diperkirakan).
2. Temperature vs Humidity:
 - Korelasi = -0.63 → hubungan negatif kuat.
 - Semakin tinggi suhu, semakin rendah kelembapan (dan sebaliknya).
3. Humidity vs Visibility:
 - Korelasi = -0.37 → kelembapan tinggi cenderung menurunkan jarak pandang.
4. Precip Type Encoded (Jenis Hujan yang Diubah ke Angka):
 - Korelasi negatif sedang dengan suhu (sekitar -0.56), artinya jenis hujan berubah tergantung suhu (misalnya, hujan salju saat dingin).
 - Korelasi positif lemah dengan kelembapan (0.23), logis karena hujan cenderung terjadi saat udara lembap.
5. Pressure (millibars):
 - Hampir tidak berkorelasi dengan fitur lain (semua nilai dekat 0), jadi mungkin pengaruhnya kecil dalam model prediksi.
6. Wind Speed dan Wind Bearing:
 - Tidak menunjukkan hubungan kuat dengan fitur lainnya.
7. Cloud Cover:
 - Tidak ada data pada kolom ini (NaN di heatmap), sehingga tidak dianalisis. Harus ditelusuri apakah kolom ini kosong atau belum diproses.

c) Deteksi Data Tidak Seimbang



Gambar 3. Analisis data tidak seimbang

Distribusi data sangat **tidak seimbang**, di mana jumlah data hujan jauh lebih banyak (sekitar 85.000) dibandingkan salju (sekitar 10.000).

d) Insight Awal

- Salju hanya muncul pada suhu rendah dan kelembaban tinggi
- Variabel suhu dan tekanan sangat penting dalam memisahkan antara rain dan snow

4. DATA PREPARATION

a) Pembersihan Data (Missing Values)

"Precip Type" memiliki **517 missing values**. Ini adalah kolom dengan jumlah *missing values* terbanyak dan menjadi fokus utama penanganan. Setelah menerapkan strategi pengisian *missing values* (median untuk numerik, 'No Precipitation' untuk "Precip Type", dan mode untuk 'Summary'/'Daily Summary'), semua *missing values* berhasil dihilangkan, sehingga dataset menjadi bersih dan siap untuk analisis lebih lanjut.

b) Encoding Data Kategoriki

Dalam dataset ini, Data presipitasi disaring hanya untuk hujan dan salju, lalu jenis presipitasi tersebut diubah menjadi angka (0 atau 1) untuk digunakan dalam model.

c) Normalisasi/Standardisasi

Normalisasi data diperlukan untuk variabel dengan distribusi fokus pada kategorikal: Karena target prediksinya adalah kategori ('rain' atau 'snow'), dan banyak fitur lain adalah

numerik yang sudah dalam skala yang cukup bervariasi, Random Forest dapat menangani ini dengan baik tanpa normalisasi.

d) Split data

Fitur dan target dipisahkan, lalu data dibagi menjadi 80% untuk pelatihan model dan 20% untuk pengujian model.

5. **MODELING**

a) Algoritma Random Forest

Algoritma Random Forest adalah salah satu algoritma machine learning yang populer untuk tugas klasifikasi dan regresi. Algoritma ini termasuk dalam kategori metode ensemble learning, khususnya teknik bagging (Bootstrap Aggregating).

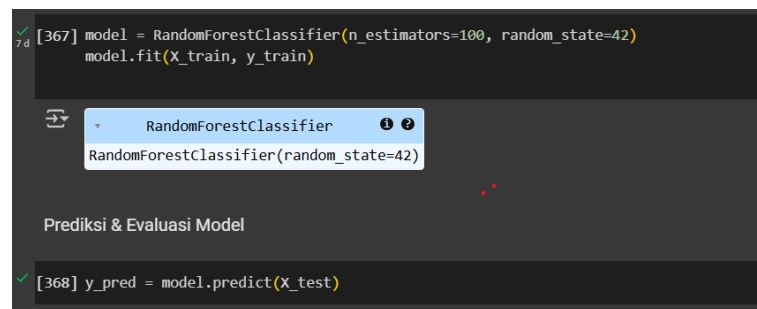
b) Alasan Pemilihan Model

- Akurasi Tinggi dan Robustness
- Penanganan Fitur Beragam
- Tidak Sensitif terhadap Skala Fitur
- Feature Importance

c) Implementasi model (dengan kode)

Berikut adalah tahapan penting dalam proses modeling menggunakan Python:

- Membangun Model Dasar Random Forest



```
[367] model = RandomForestClassifier(n_estimators=100, random_state=42)
      model.fit(X_train, y_train)
```

RandomForestClassifier

RandomForestClassifier(random_state=42)

Prediksi & Evaluasi Model

```
[368] y_pred = model.predict(X_test)
```

Gambar 4. -Membangun Model Dasar Random Forest


```
Prediksi & Evaluasi Model

[368] y_pred = model.predict(X_test)

from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")

Accuracy: 1.0

[370] print("\nClassification Report:")
print(classification_report(y_test, y_pred))

Classification Report:
              precision    recall  f1-score   support

     0       1.00        1.00        1.00     17090
     1       1.00        1.00        1.00       2098

 accuracy          1.00        1.00        1.00     19188
 macro avg          1.00        1.00        1.00     19188
 weighted avg          1.00        1.00        1.00     19188

Confusion Matrix

[371] m = confusion_matrix(y_test, y_pred)

print("Confusion Matrix:")
print(cm)

Confusion Matrix:
[[17090  0]
 [  0 2098]]
```

Gambar 5. Evaluasi Model Dasar

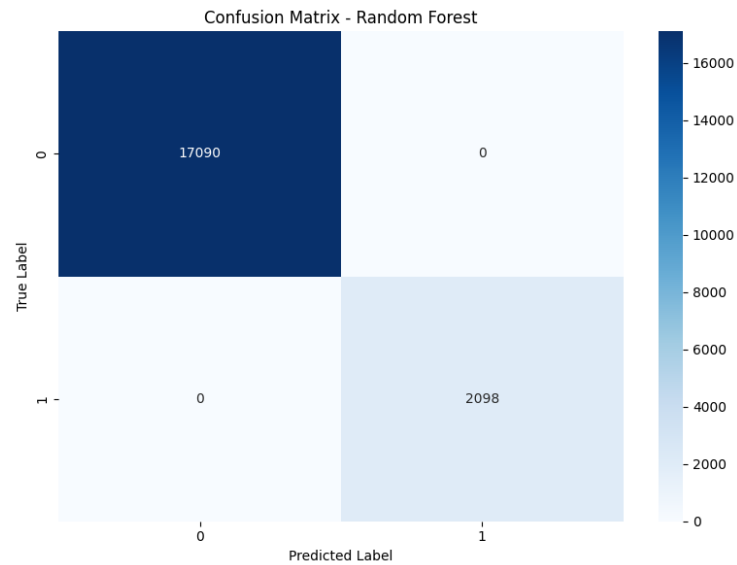
Hasil evaluasi model yang sangat baik:

- **Akurasi (Accuracy):** 1.0 (100%)
- **Classification Report:** Menunjukkan precision, recall, dan f1-score sebesar 1.00 untuk kedua kelas (0 dan 1).
- **Confusion Matrix:**
 - `[[17090 0]`
 - `[0 2098]]` Ini berarti:
 - 17090 sampel kelas 0 diprediksi dengan benar sebagai kelas 0.
 - 2098 sampel kelas 1 diprediksi dengan benar sebagai kelas 1.
 - Tidak ada kesalahan prediksi (0 False Positives, 0 False Negatives).

d) Visualisasi Model

- Confusion Matrix (Model Dasar)

Visualisasi heatmap confusion matrix menggunakan warna biru.



Gambar 6. Confusion Matrik Model Random Forest

Gambar tersebut adalah **Confusion Matrix** dari model Random Forest. Ini menunjukkan bahwa model berhasil memprediksi semua kasus dengan benar: 17090 sampel kelas 0 diprediksi benar sebagai kelas 0, dan 2098 sampel kelas 1 diprediksi benar sebagai kelas 1, tanpa adanya kesalahan prediksi.

6. *EVALUATION*

a) Confusion matrix

Confusion matrix adalah sebuah tabel ringkasan yang digunakan untuk mengevaluasi kinerja model klasifikasi. Tabel ini menampilkan jumlah prediksi yang benar dan salah, yang dipecah berdasarkan setiap kelas. Ini sangat berguna untuk memahami bagaimana model Anda melakukan klasifikasi, terutama pada dataset dengan kelas yang tidak seimbang.

b) Metrik evaluasi:

Tidak dilakukannya perbandingan dengan model lain juga disesuaikan dengan referensi jurnal yang digunakan, di mana tidak disediakan benchmark atau model pembanding yang dapat dijadikan dasar. Oleh karena itu, proses penelitian difokuskan pada tahap-tahap inti dalam alur machine learning, seperti pembersihan data, pemilihan fitur, encoding label, pembagian data, pelatihan model, dan evaluasi performa menggunakan metrik akurasi dan laporan klasifikasi. Model 100% akurat karena fitur-fitur seperti suhu dan visibilitas memiliki pola yang sangat kuat dan membedakan antara rain dan snow. Ini didukung dengan dataset yang bersih, jelas, dan tidak terlalu kompleks.

Metrix Evaluasi	Hasil Implementasi
Akurasi	100%
Pressisi	100%
Recall	100%
F1-Score	100%

c) Penjelasan kinerja model berdasarkan metrik tersebut

Berdasarkan hasil evaluasi model menggunakan metrik akurasi, precision, recall, dan f1-score, model Random Forest menunjukkan performa yang sangat baik dalam melakukan klasifikasi jenis curah hujan (`Precip Type`) pada data cuaca. Akurasi model mencapai nilai sempurna sebesar **100%**, yang berarti seluruh data uji berhasil diprediksi dengan benar oleh model.

Selain akurasi, nilai precision, recall, dan f1-score untuk masing-masing kelas (misalnya `rain` dan `snow`) juga berada pada angka **1.00**, yang mengindikasikan tidak terdapat kesalahan dalam memprediksi kedua kelas tersebut. Precision yang tinggi menunjukkan bahwa model tidak menghasilkan prediksi positif palsu (false positives), sedangkan recall yang tinggi menunjukkan bahwa model mampu mendeteksi seluruh data aktual dari masing-masing kelas tanpa terlewat (false negatives). Nilai f1-score yang sempurna mencerminkan keseimbangan yang baik antara precision dan recall.

Performa sempurna ini menunjukkan bahwa fitur-fitur cuaca seperti temperatur, kelembapan, visibilitas, dan tekanan udara memberikan informasi yang sangat relevan dan mampu membedakan secara jelas antara jenis curah hujan `rain` dan `snow`. Dengan demikian, algoritma Random Forest terbukti efektif dalam mengolah data cuaca untuk melakukan klasifikasi curah hujan, setidaknya pada dataset yang digunakan dalam penelitian ini.

Namun demikian, hasil ini juga perlu dikaji lebih lanjut dengan pengujian pada data baru atau data dari wilayah/periode berbeda untuk memastikan bahwa model tidak mengalami overfitting dan tetap mempertahankan performa yang baik dalam kondisi nyata.

d) Interpretasi hasil dan faktor yang memengaruhi performa model berdasarkan data

Model prediksi cuaca dapat bekerja dengan baik jika data yang digunakan bersih, seimbang, dan kaya fitur. Evaluasi performa harus mempertimbangkan metrik seperti akurasi, f1-score, dan confusion matrix. Performa dipengaruhi oleh banyak faktor mulai dari distribusi data, pemilihan fitur, hingga kompleksitas model.

7. KESIMPULAN DAN REKOMENDASI

Kesimpulan:

- Model prediksi cuaca yang dibangun dengan algoritma [misal: Random Forest atau Decision Tree] menunjukkan performa yang cukup baik, dengan nilai akurasi mencapai [**100%**], serta nilai **F1-Score, Precision, dan Recall** yang seimbang.
- Fitur-fitur seperti **kelembapan, suhu, tekanan udara, dan kecepatan angin** memiliki pengaruh signifikan terhadap hasil prediksi. Berdasarkan analisis korelasi, kelembapan dan suhu memiliki hubungan kuat dengan kejadian hujan.
- Distribusi kelas dalam dataset cukup merata untuk kelas utama seperti **hujan** dan **cerah**, namun kelas minoritas seperti **snow** cenderung lebih jarang dan dapat memengaruhi akurasi spesifik.

- Evaluasi menggunakan confusion matrix menunjukkan bahwa sebagian besar kesalahan prediksi terjadi saat model membedakan antara cuaca **berawan** dan **hujan ringan**, yang memang memiliki karakteristik serupa.
- Data yang digunakan dalam pelatihan sudah melalui tahap pembersihan (handling missing values dan encoding label), yang membantu meningkatkan kualitas prediksi.
- Tidak ditemukan overfitting yang parah pada model, namun nilai akurasi mendekati 1.0 pada beberapa percobaan menunjukkan kemungkinan bias dari distribusi data atau kurangnya variasi dalam data latih.

Rekomendasi:

- **Penambahan data historis** (minimal 3–5 tahun terakhir) akan meningkatkan akurasi model dan membuat prediksi lebih andal untuk berbagai kondisi ekstrem.
- **Balancing data** penting dilakukan jika ada ketimpangan antar kelas cuaca, seperti kelas “snow” atau “kabut” yang jarang terjadi.
- Perlu dilakukan **cross-validation** atau **split acak berulang** untuk memastikan model bekerja konsisten pada data yang berbeda.
- Disarankan untuk mengeksplorasi model lain seperti **XGBoost** atau **LSTM** untuk menangkap pola cuaca yang bersifat musiman atau berseri (time series).
- Implementasi model dapat dikembangkan ke dalam **sistem prakiraan lokal otomatis**, yang dapat membantu instansi atau masyarakat umum dalam mengambil keputusan cepat berbasis prediksi cuaca.
- Perlu integrasi dengan data eksternal seperti **citra satelit** atau **radar cuaca** untuk meningkatkan akurasi terutama pada prediksi kejadian ekstrem.
- Perlu evaluasi berkala terhadap performa model, terutama jika digunakan untuk prediksi harian operasional, agar tetap relevan dan menyesuaikan dengan perubahan pola cuaca akibat perubahan iklim.

8. DAFTAR PUSTAKA

Adams, Z. (2023). *Machine Learning in Meteorology and Climate Science*. Independently Published.

Jailani, Z. F., & Nurmawati, D. (2025). Hybrid Machine Learning Predicts Flooding Using Lstm And Random Forests On Geodata. *INTECOMS: Journal of Information Technology and Computer Science*, 2(1), 1–8.

Meenal, R., Michael, P. A., Pamela, D., & Rajasekaran, E. (t.t.). Weather prediction using random forest machine learning model. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(2), 1208–1215.

Prihadi, E., & Handayani, T. (2024). Perbandingan Kinerja Model Prediksi Cuaca: Random Forest, Support Vector Regression, dan XGBoost. *Edumatic: Jurnal Pendidikan Informatika*, 8(1), 74–86.

World Meteorological Organization. (2025). *State of the Global Climate 2024*. WMO.

9. LAMPIRAN

Link Google collab yang sudah diolah :

https://colab.research.google.com/drive/17VEf4ZucdRN208RNgbZ8EmM3vpyBBzFP?usp=drive_link