Laporan Tugas Machine Learning

LINEAR REGRESSION

disusun untuk memenuhi Tugas
Machine Learning A

Oleh:

M. Nabil Maulana	(2208107010011)
Irfan Rizadi	(2208107010062)
Maulana Fikri	(2208107010042)
Indriani Miza Alfiyanti	(2208107010026)
Raihan Firyal	(2208107010084)

DEPARTEMEN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

BANDA ACEH

2025

1.1 Latar Belakang

Perubahan iklim dan kondisi cuaca yang tidak menentu semakin menuntut adanya sistem prediksi cuaca yang akurat dan andal. Informasi mengenai suhu udara sangat penting untuk berbagai sektor, seperti pertanian, penerbangan, kesehatan, dan kegiatan sehari-hari masyarakat. Oleh karena itu, peramalan suhu berdasarkan parameter cuaca menjadi salah satu aspek penting dalam sistem prediksi cuaca modern.

Dalam tugas ini, dilakukan pemodelan untuk memprediksi suhu udara berdasarkan berbagai parameter cuaca seperti kelembaban, tekanan udara, dan kecepatan angin. Model yang digunakan dalam prediksi ini adalah Linear Regression dan Polynomial Regression, dua metode regresi yang sering digunakan dalam masalah regresi kontinu.

Dataset yang digunakan berisi data cuaca historis dengan beberapa fitur yang berperan sebagai variabel independen, serta suhu sebagai variabel target. Tujuan utama dari proyek ini adalah untuk membangun model prediktif yang mampu memperkirakan suhu secara akurat, serta melakukan evaluasi terhadap kinerja masing-masing model regresi yang diterapkan.

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Untuk memahami hubungan antara parameter cuaca (seperti kelembaban, tekanan udara, dan kecepatan angin) terhadap suhu udara.
2. Untuk membangun dan membandingkan model Linear Regression dan Polynomial Regression dalam memprediksi suhu.
3. Untuk mengevaluasi kinerja model menggunakan metrik seperti Mean Squared Error (MSE), Mean Absolute Error (MAE), dan R^2 Score.
4. Untuk menentukan apakah model yang dibangun cukup efektif digunakan dalam memprediksi suhu udara berdasarkan data cuaca historis.

1.3 Pemahaman Dataset

1.3.1 Analisis Informasi Statistik

Dataset yang digunakan dalam penelitian ini berisi sebanyak **96.453 data cuaca historis** dengan beberapa fitur utama yang berkaitan dengan kondisi atmosfer pada waktu tertentu. Analisis statistik deskriptif dilakukan untuk memahami karakteristik dasar dari masing-masing variabel numerik.

(Isi Gambar)

Berikut adalah ringkasan dari statistik deskriptif variabel-variabel utama:

- **Temperature (°C)**
  - Rata-rata: 11.93 °C
  - Minimum: -21.82 °C
  - Maksimum: 39.91 °C
  - Standar deviasi: 9.55

Ini adalah target utama yang akan diprediksi oleh model. Nilai suhu bervariasi cukup lebar, menandakan adanya distribusi suhu dari musim dingin hingga musim panas.

- **Apparent Temperature (°C)**
  Suhu yang dirasakan, sedikit lebih tinggi atau lebih rendah tergantung kelembaban dan kecepatan angin.

- **Humidity**
  - Rata-rata: 0.73 (skala 0 sampai 1)
  - Hampir semua nilai berada pada rentang 0.6 hingga 1, menandakan kelembaban cukup tinggi secara umum.

- **Wind Speed (km/h)**
  - Rata-rata: 10.81 km/h
  - Maksimum: 63.85 km/h
  - Terdapat nilai 0, menunjukkan kondisi tanpa angin pada beberapa titik data.

- **Wind Bearing (degrees)**
  Arah angin dalam derajat (0–359), tersebar merata dari seluruh arah mata angin.

- **Visibility (km)**
  Rata-rata visibilitas sekitar 10.35 km, dengan nilai maksimum hingga 16 km, dan minimum 0 km (mungkin menandakan kabut atau badai).

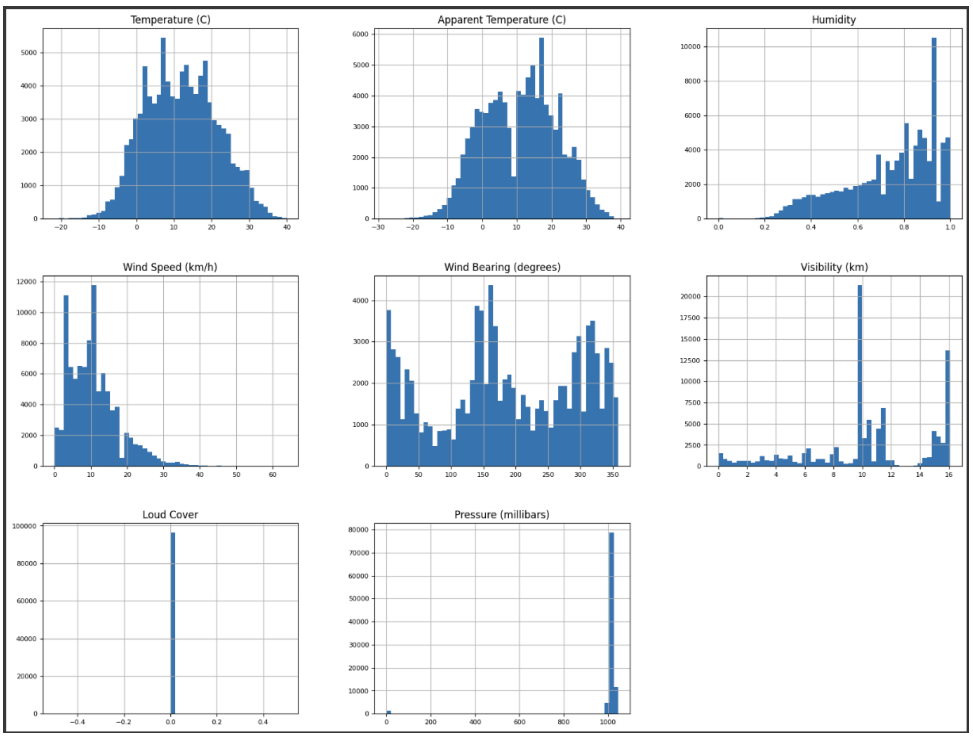
- **Loud Cover**
  Semua nilainya 0, menunjukkan fitur ini kemungkinan tidak memiliki variasi dan mungkin bisa dihapus saat pemodelan.

- **Pressure (millibars)**
  - Rata-rata: 1003.24 millibars
  - Variasi tekanan atmosfer dalam rentang normal, dengan standar deviasi sebesar 116.97 menunjukkan adanya beberapa nilai ekstrem.

Statistik deskriptif ini memberikan gambaran awal mengenai distribusi dan penyebaran data. Informasi ini sangat penting dalam tahapan prapemrosesan dan seleksi fitur sebelum membangun model prediksi.

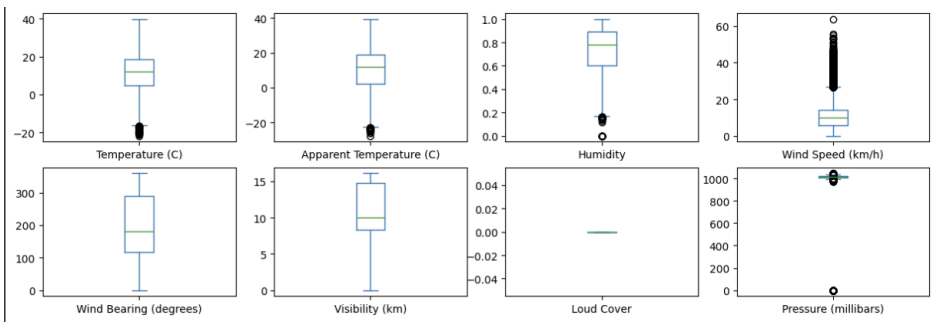
1.3.2 Visualisasi Distribusi Data

a. Distribusi Kolom Numerik (Histogram)



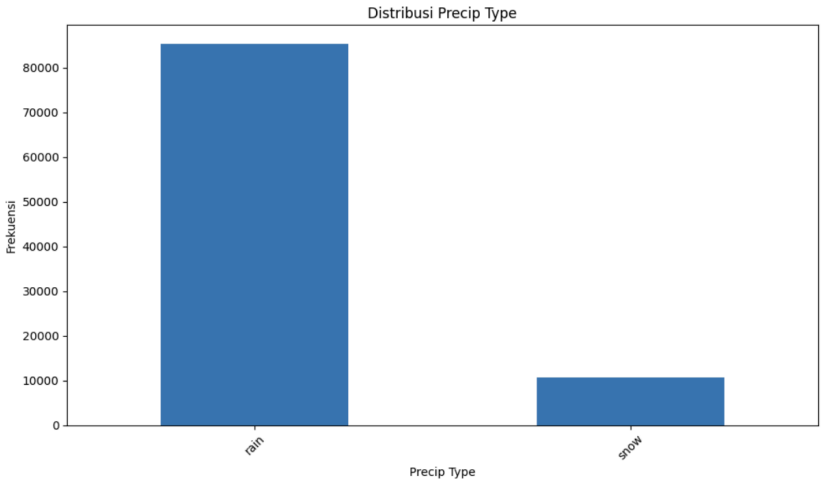
Distribusi kolom numerik melalui histogram menunjukkan pola sebaran data untuk tiap fitur numerik dalam dataset. Temperatur dan Apparent Temperature terlihat mengikuti distribusi normal yang cukup simetris, menunjukkan bahwa sebagian besar nilai berkisar di tengah-tengah rentang data. Sementara itu, Humidity atau kelembaban cenderung tinggi, dengan mayoritas nilai mendekati angka 1, menandakan lingkungan yang umumnya lembab. Kecepatan angin (Wind Speed) memiliki distribusi yang condong ke kanan (right-skewed), di mana sebagian besar nilai berada di kisaran rendah, namun terdapat beberapa nilai ekstrem yang lebih tinggi. Wind Bearing, atau arah angin, tersebar merata dari 0 hingga 360 derajat, mencerminkan bahwa arah angin bervariasi secara acak. Visibility atau jarak pandang menunjukkan banyak konsentrasi pada nilai tertentu yang tinggi, menandakan bahwa sebagian besar kondisi cuaca memiliki jarak pandang yang baik. Untuk Cloud Cover, datanya tampak konstan pada satu nilai, mengindikasikan bahwa fitur ini mungkin tidak memberikan variasi informasi. Terakhir, tekanan udara (Pressure) sangat terkonsentrasi di sekitar satu nilai tertentu, menunjukkan kestabilan tekanan atmosfer dalam data.

b. Boxplot Kolom Numerik



Visualisasi boxplot membantu mengidentifikasi outlier dan distribusi umum dari data numerik. Dari boxplot tersebut, terlihat bahwa beberapa fitur seperti Humidity, Wind Speed, dan Pressure memiliki sejumlah outlier, ditunjukkan oleh titik-titik di luar whisker. Temperatur dan Apparent Temperature juga menunjukkan beberapa outlier pada sisi bawah, tetapi secara umum memiliki rentang nilai yang stabil dan simetris. Wind Bearing memiliki jangkauan nilai yang luas namun tidak menunjukkan outlier yang signifikan. Sementara itu, fitur Cloud Cover kembali menunjukkan bahwa datanya tidak bervariasi, dengan satu nilai mendominasi keseluruhan data. Visibility cenderung stabil tanpa banyak penyimpangan atau nilai ekstrem.

c. Distribusi Kolom Kategorikal (Bar Plot)



Untuk kolom kategorikal, distribusi pada *Precip Type* menunjukkan bahwa mayoritas data berkaitan dengan hujan (rain), sementara jumlah observasi untuk salju (snow) jauh lebih sedikit. Hal ini menunjukkan bahwa kondisi hujan jauh lebih umum terjadi dibandingkan salju dalam data cuaca yang dianalisis, dan bisa menjadi indikasi penting dalam pemodelan atau interpretasi data terkait pola cuaca.

1.4 Eksplorasi Data dan Pra-Pemrosesan

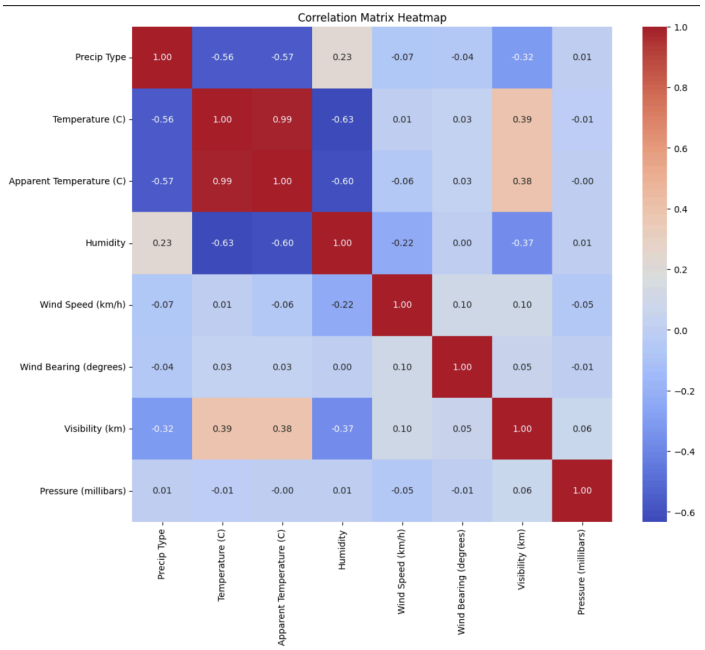
1.4.1 Eksplorasi Data

Langkah awal eksplorasi data dilakukan dengan memeriksa apakah terdapat missing value (nilai kosong) di dalam dataset. Dengan menggunakan fungsi df.isnull().sum(), dapat diketahui jumlah nilai yang hilang di setiap kolom. Langkah ini penting untuk memastikan kualitas data yang akan digunakan dalam analisis maupun pelatihan model. Data yang memiliki terlalu banyak nilai kosong bisa menyebabkan bias atau kesalahan pada hasil analisis.

Setelah mengetahui bahwa terdapat nilai kosong, langkah berikutnya adalah menghapus baris yang mengandung missing value menggunakan df.dropna(inplace=True). Meskipun pendekatan ini cukup ekstrem karena dapat mengurangi jumlah data, cara ini memastikan bahwa hanya data yang lengkap yang digunakan dalam proses analisis. Penghapusan ini dipilih karena bisa jadi imputasi tidak tepat atau data kosong tidak dapat direkonstruksi secara akurat.

Selanjutnya, dilakukan pengecekan terhadap data duplikat dengan fungsi df.duplicated().sum(). Duplikasi data dapat mengganggu hasil analisis dan membuat model terlalu "belajar" dari data yang sama. Jika ditemukan duplikat, biasanya akan dipertimbangkan untuk dihapus agar data yang dianalisis benar-benar merepresentasikan kondisi sebenarnya.

Tahap berikutnya adalah encoding pada kolom kategorikal, khususnya kolom 'Precip Type', yang diubah dari bentuk teks menjadi numerik (rain menjadi 0, dan snow menjadi 1). Ini diperlukan agar kolom tersebut dapat diikutsertakan dalam analisis korelasi maupun dalam proses machine learning yang membutuhkan input numerik.



Hasil heatmap menunjukkan bahwa Temperature (C) dan Apparent Temperature (C) memiliki korelasi sangat kuat positif (0.99), yang mengindikasikan bahwa keduanya hampir identik. Sebaliknya, Humidity berkorelasi negatif dengan Temperature dan Apparent Temperature, artinya saat suhu meningkat, kelembapan cenderung menurun. Korelasi lainnya seperti Precip Type dengan Temperature dan Apparent Temperature juga menunjukkan hubungan negatif, menunjukkan bahwa salju cenderung terjadi pada suhu yang lebih rendah. Korelasi yang ditampilkan pada heatmap ini menjadi dasar penting dalam memilih fitur yang paling relevan untuk model prediktif.

1.4.2 Preprocessing

Pada tahap preprocessing, dilakukan serangkaian langkah untuk mempersiapkan data sebelum digunakan dalam proses pemodelan machine learning. Tahapan pertama adalah menghapus fitur yang memiliki multikolinearitas tinggi, yaitu fitur `Apparent Temperature (C)`. Hal ini dilakukan karena fitur tersebut memiliki korelasi sangat kuat (mendekati 1) dengan `Temperature (C)` berdasarkan hasil heatmap sebelumnya. Fitur dengan korelasi yang sangat tinggi dapat menyebabkan redundansi informasi dan membingungkan model, sehingga salah satunya perlu dihapus untuk menjaga efisiensi dan akurasi analisis.

Langkah selanjutnya adalah menangani outlier atau pencilan pada data numerik. Untuk ini digunakan pendekatan berbasis interquartile range (IQR), di mana nilai-nilai yang berada di bawah batas bawah (Q1 - 1.5 * IQR) dianggap sebagai outlier dan kemudian diganti dengan nilai Q3 (kuartil atas). Proses ini dilakukan secara bertahap pada kolom-kolom numerik seperti suhu, kelembapan, kecepatan angin, arah angin, jarak pandang, dan tekanan udara. Pendekatan ini membantu menjaga stabilitas data tanpa menghilangkan terlalu banyak informasi, dengan mengganti outlier ekstrem menjadi nilai yang lebih masuk akal.

Tahap akhir preprocessing adalah melakukan normalisasi data menggunakan teknik RobustScaler. Metode ini sangat efektif karena lebih tahan terhadap outlier dibandingkan scaler lain seperti MinMaxScaler atau StandardScaler. RobustScaler bekerja dengan mengurangi median dan membagi data dengan rentang antar kuartil (IQR), sehingga distribusi data menjadi lebih seimbang dan tidak terpengaruh oleh nilai ekstrem. Setelah proses scaling, data diubah kembali menjadi DataFrame baru dengan kolom-kolom asli agar tetap mudah dianalisis dan digunakan dalam model selanjutnya.

1.5 Implementasi Model

1.5.1 Data Splitting dan Inisialisasi Model Linear Regression

Tahap akhir preprocessing adalah melakukan normalisasi data menggunakan teknik RobustScaler. Metode ini sangat efektif karena lebih tahan terhadap outlier dibandingkan scaler lain seperti MinMaxScaler atau StandardScaler. RobustScaler bekerja dengan mengurangi median dan membagi data dengan rentang antar kuartil (IQR), sehingga distribusi data menjadi lebih seimbang dan tidak terpengaruh oleh nilai ekstrem. Setelah proses scaling, data diubah kembali menjadi DataFrame baru dengan kolom-kolom asli agar tetap mudah Pada tahap implementasi model, langkah pertama yang dilakukan adalah proses pemisahan data (data splitting) dan inisialisasi model regresi linear (Linear Regression). Pemisahan data ini bertujuan untuk membagi dataset yang telah diproses menjadi dua bagian, yaitu data pelatihan (training set) dan data pengujian (testing set). Fitur atau variabel independen (`X`) diambil dari seluruh kolom numerik hasil scaling, kecuali kolom `Temperature (C)` yang digunakan sebagai variabel target atau dependen (`y`). Pemisahan dilakukan dengan proporsi 80% data untuk pelatihan dan 20% untuk pengujian, serta menggunakan parameter `random_state=42` untuk memastikan hasil pemisahan konsisten setiap kali kode dijalankan.

Setelah data berhasil dipisah, tahap selanjutnya adalah inisialisasi model regresi linear. Model ini merupakan salah satu algoritma machine learning paling sederhana namun efektif, yang bekerja dengan mencari hubungan linear antara variabel input dan output. Model Linear Regression yang telah diinisialisasi kemudian dilatih (fit) menggunakan data pelatihan (`X_train` dan `y_train`). Proses pelatihan ini memungkinkan model untuk mempelajari pola dalam data sehingga dapat digunakan untuk melakukan prediksi terhadap suhu berdasarkan fitur-fitur lainnya.dianalisis dan digunakan dalam model selanjutnya.

1.5.2 Inisialisasi Model Polynomial Regression dengan Hyperparameters

Pada tahap ini, dilakukan inisialisasi model Polynomial Regression dengan kombinasi regularisasi Ridge Regression, yang kemudian dioptimalkan menggunakan GridSearchCV untuk menemukan kombinasi hyperparameter terbaik. Model ini dibuat dalam bentuk pipeline agar proses transformasi dan pelatihan dapat berjalan secara berurutan dan otomatis. Pertama,

data input akan ditransformasi ke dalam bentuk polinomial melalui PolynomialFeatures, kemudian dilanjutkan dengan proses regresi menggunakan Ridge Regression, yang membantu mengurangi overfitting melalui teknik regularisasi.

```
Best hyperparameters: {'polynomialfeatures__degree': 4, 'ridge__alpha': 0.1}
Mean Squared Error: 0.0015404365117202 39
R-squared: 0.9965991745476339
Mean Absolute Error: 0.018373209021861008
```

Untuk mengoptimalkan performa model, dilakukan pencarian hyperparameter terbaik dengan GridSearchCV, yaitu dengan mengevaluasi beberapa kombinasi derajat polinomial (degree: 2, 3, 4) dan kekuatan regularisasi (alpha: 0.1, 1, 10). Evaluasi dilakukan dengan cross-validation sebanyak 5 kali (cv=5) dan menggunakan metrik negative mean squared error sebagai acuan pemilihan model terbaik. Hasil dari pencarian ini menunjukkan bahwa kombinasi terbaik diperoleh pada degree = 4 dan alpha = 0.1.

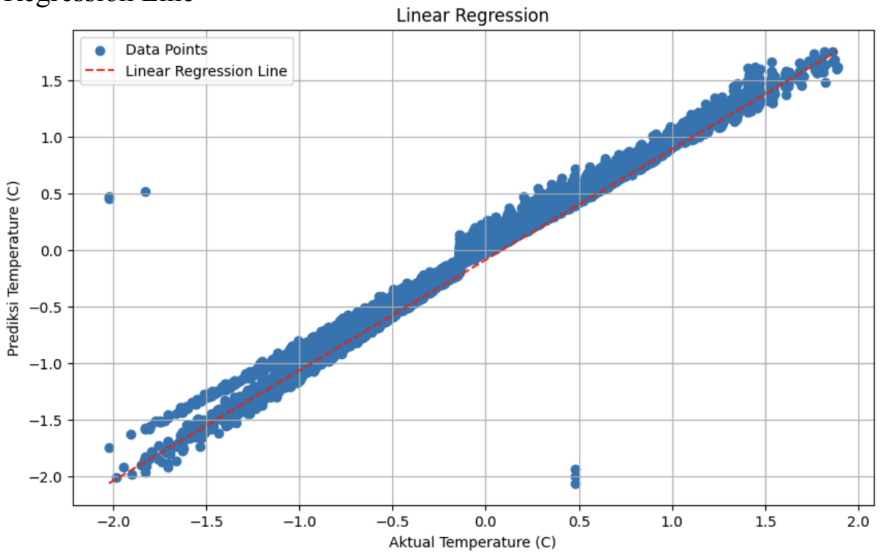
Setelah model terbaik ditemukan, dilakukan evaluasi terhadap data uji menggunakan beberapa metrik evaluasi, yakni Mean Squared Error (MSE), R-squared (R^2), dan Mean Absolute Error (MAE). Nilai MSE yang sangat kecil (0.00154), R^2 yang sangat mendekati 1 (0.9966), serta MAE yang rendah (0.0184), menunjukkan bahwa model Polynomial Regression dengan derajat 4 dan regularisasi yang lemah mampu memberikan prediksi suhu yang sangat akurat berdasarkan fitur-fitur cuaca lainnya.

1.6 Evaluasi Model

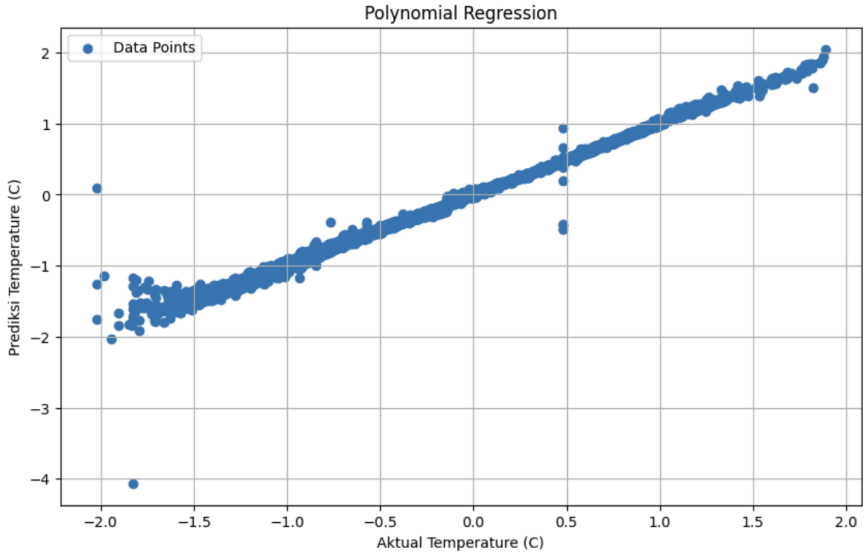
Pada tahap evaluasi model, dilakukan perbandingan kinerja antara model regresi linear dan regresi polinomial menggunakan tiga metrik evaluasi, yaitu Mean Squared Error (MSE), R-squared (R^2), dan Mean Absolute Error (MAE). Model regresi linear menghasilkan nilai MSE sebesar 0.0062, R^2 sebesar 0.9862, dan MAE sebesar 0.0521, yang menunjukkan bahwa model ini mampu menjelaskan sekitar 98.62% variasi data dan memiliki kesalahan prediksi yang relatif kecil. Sementara itu, model regresi polinomial menunjukkan performa yang lebih baik dengan nilai MSE yang lebih rendah sebesar 0.0015, R^2 sebesar 0.9966, dan MAE sebesar 0.0184. Hasil ini menunjukkan bahwa model polinomial memiliki kemampuan prediksi yang lebih akurat dibandingkan regresi linear, terutama dalam menangkap pola data yang bersifat non-linear.

1.7 Analisis Hasil

a. Visualisasi Regression Line

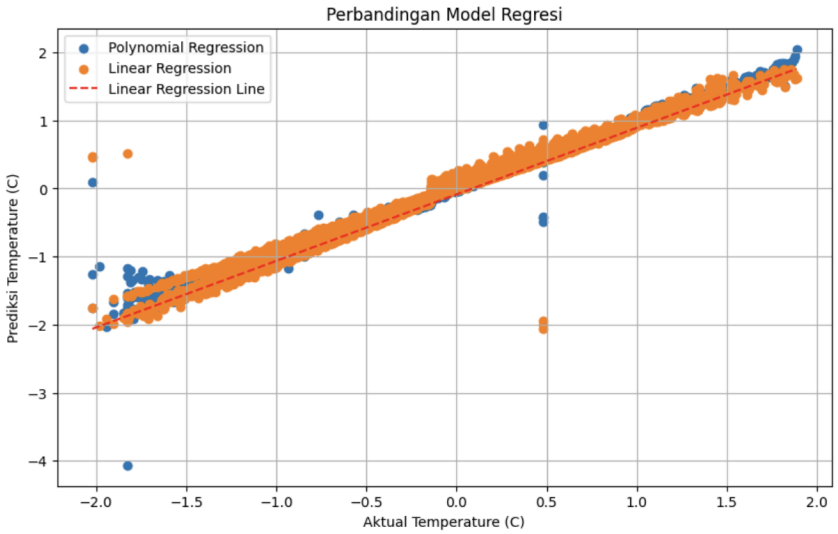


b. Visualisasi Polynomial Fit



Berdasarkan hasil visualisasi regresi linear dan regresi polinomial terhadap prediksi suhu, terlihat bahwa regresi linear menunjukkan hubungan yang sangat kuat antara suhu aktual dan suhu yang diprediksi, yang ditandai dengan distribusi titik data yang mengikuti garis regresi linear merah secara konsisten. Hal ini mengindikasikan bahwa model linear mampu menangkap pola hubungan suhu dengan cukup baik, meskipun terdapat beberapa outlier. Sementara itu, pada visualisasi regresi polinomial, distribusi titik data masih menunjukkan tren positif yang serupa, namun terdapat beberapa penyimpangan yang lebih besar, terutama pada nilai-nilai ekstrem, yang menunjukkan bahwa model polinomial cenderung lebih sensitif terhadap fluktuasi data dan kemungkinan mengalami overfitting. Dengan demikian, secara keseluruhan, model regresi linear memberikan hasil prediksi yang lebih stabil dan akurat dalam memodelkan hubungan antara suhu aktual dan suhu prediksi dibandingkan dengan model regresi polinomial pada data ini.

c. Perbandingan Model Regresi



```
Interpretasi Koefisien Regresi (Linear Regression):
Precip Type: 0.0051
Apparent Temperature (C): 1.0041
Humidity: -0.0387
Wind Speed (km/h): 0.0411
Wind Bearing (degrees): -0.0047
Visibility (km): 0.0012
Pressure (millibars): -0.0129
Intercept: 0.0545
```

Visualisasi pada grafik menunjukkan perbandingan performa antara model regresi linear dan regresi polinomial dalam memprediksi suhu berdasarkan data uji. Titik-titik oranye mewakili prediksi dari regresi linear, sedangkan titik biru adalah hasil dari regresi polinomial, dengan garis merah putus-putus menggambarkan garis regresi linear ideal. Hasil interpretasi koefisien regresi linear menunjukkan bahwa variabel “Apparent Temperature (C)” memiliki pengaruh paling dominan terhadap suhu aktual dengan koefisien sebesar 1.0041, sementara variabel lainnya seperti “Humidity”, “Wind Speed”, dan “Pressure” memiliki pengaruh yang relatif kecil. Intercept sebesar 0.0545 menandakan nilai suhu yang diprediksi saat semua variabel input bernilai nol. Visualisasi dan interpretasi ini membantu memahami hubungan antara fitur-fitur cuaca dengan suhu aktual serta membandingkan akurasi dua pendekatan regresi dalam memodelkan data.

Kesimpulan

Berdasarkan hasil evaluasi model, regresi linier menunjukkan performa yang cukup baik dengan nilai R^2 sebesar 98.62%, yang berarti model mampu menjelaskan sebagian besar variansi data aktual meskipun masih terdapat sedikit error. Namun, model Polynomial Regression memberikan hasil yang lebih unggul dengan R^2 sebesar 99.66% serta nilai MSE dan MAE yang lebih rendah, menunjukkan bahwa model ini lebih akurat dalam menangkap pola hubungan non-linear dalam data. Visualisasi mendukung hasil ini, di mana prediksi dari model polinomial (titik biru) lebih mendekati nilai aktual dibandingkan model linear (titik oranye), terutama pada data dengan nilai ekstrem. Interpretasi koefisien regresi linier juga menunjukkan bahwa variabel "Apparent Temperature (C)" memiliki pengaruh paling signifikan terhadap suhu aktual, disusul oleh kelembapan dan kecepatan angin. Secara keseluruhan, model Polynomial Regression direkomendasikan untuk digunakan jika tujuan utama adalah meminimalkan error dan meningkatkan akurasi prediksi suhu, terutama ketika hubungan antar variabel tidak bersifat linear.