Seminar Nasional Sains Data 2025 (SENADA 2025) E-ISSN 2808-5841
UPN "Veteran" Jawa Timur P-ISSN 2808-7283

Perbandingan Klasifikasi Cuaca Kota Denpasar dengan Regresi Logistik Multinomial dan Analisis Diskriminan

Muhamad Alfa Reza Gobel[1], Alfathrindra Agastyo Yulianto[2], Hizkia Marvel Abinaya[3] ,Ike Fitriyaningsih, M.Si[4]

[1, 2, 3] *S1 Sains Data, Universitas Negeri Surabaya*

[1] muhamadalfa.23207@mhs.unesa.ac.id

[2] alfathrindra.23094@mhs.unesa.ac.id

[3] hizkia.23061@mhs.unesa.ac.id

[4] *S1 Kecerdasan Artifisial, Universitas Negeri Surabaya*

[4] Ikefitriyaningsih@unesa.ac.id

Abstract: *Weather forecasting plays a vital role in supporting public activities, especially in tourist areas such as Denpasar City, Bali. This study aims to classify weather conditions in 2019 into four main categories: Clouds, Rain, Thunderstorm, and Clear, using Linear Discriminant Analysis (LDA) and Multinomial Logistic Regression (MLR). The dataset, obtained from Kaggle, consists of 264,925 hourly weather observations, covering 34 atmospheric variables such as temperature, humidity, pressure, wind speed, and precipitation. The analysis process included data exploration and cleaning, followed by the application of both methods to build accurate and interpretable classification models. Results show that MLR with PCA achieved the highest accuracy (0.7713), followed by MLR without PCA (0.7664), and LDA (0.6725). MLR outperformed LDA in weather classification, while PCA contributed to model simplification. The most influential variable in LDA was temperature (temp, 3.21), while in both MLR models, it was clouds_all (857.28 without PCA and 252.40 with PCA). These findings highlight the importance of multivariate approaches in developing data-driven weather classification systems.*
Keywords: *Weather Classification, Denpasar City, Discriminant Analysis, Multinomial Regression***.**

Abstrak: Prakiraan cuaca berperan penting dalam menunjang aktivitas masyarakat, terutama di wilayah wisata seperti Kota Denpasar, Bali. Penelitian ini bertujuan mengklasifikasikan kondisi cuaca tahun 2019 ke dalam empat kategori utama: Clouds, Rain, Thunderstorm, dan Clear, menggunakan Analisis Diskriminan Linear (LDA) dan Regresi Logistik Multinomial (MLR). Dataset diperoleh dari Kaggle, terdiri dari 264.925 observasi cuaca yang dicatat setiap satu jam, dengan 34 variabel atmosfer seperti suhu, kelembapan, tekanan, kecepatan angin, dan curah hujan. Proses analisis dimulai dari eksplorasi dan pembersihan data, dilanjutkan dengan penerapan kedua metode untuk membangun model klasifikasi yang akurat dan interpretatif. Hasil menunjukkan MLR dengan PCA memberikan akurasi tertinggi (0,7713), disusul MLR tanpa PCA (0,7664), dan LDA (0,6725). MLR lebih unggul dalam klasifikasi cuaca, sedangkan PCA membantu menyederhanakan model. Variabel paling berpengaruh dalam LDA adalah suhu (temp, 3,21), sedangkan dalam MLR tanpa PCA atau dengan PCA adalah clouds_all (857,28 dan 252,40)**.** Temuan ini menegaskan pentingnya pendekatan multivariat dalam pengembangan sistem klasfikasi cuaca berbasis data.
Kata kunci: Klasifikasi Cuaca, Kota Denpasar, Analisis Diskriminan, Regresi Multinomial.

I. PENDAHULUAN

Kota Denpasar, sebagai ibu kota Provinsi Bali, memiliki peran strategis dalam sektor ekonomi, pariwisata, dan sosial. Fluktuasi cuaca di wilayah ini berdampak langsung terhadap aktivitas masyarakat dan pariwisata sebagai sektor ekonomi utama. Oleh karena itu, prakiraan cuaca yang akurat dan tepat waktu menjadi kebutuhan krusial dalam mendukung perencanaan aktivitas dan pengambilan keputusan. Prakiraan cuaca merupakan salah satu aspek penting dalam meteorologi yang memerlukan pemahaman mendalam terhadap berbagai variabel atmosfer yang saling berinteraksi secara dinamis[1]. Untuk memahami dan mengklasifikasikan kondisi cuaca secara akurat, diperlukan pendekatan statistik yang mampu menangani berbagai variabel sekaligus secara simultan, dalam hal ini, analisis multivariat menjadi pilihan yang relevan dan efektif.

Analisis multivariat adalah salah satu teknik dalam statistika yang digunakan untuk menganalisis secara simultan variabel lebih dari satu[2]. Dalam konteks meteorologi, metode ini sangat berguna untuk mengekstraksi pola tersembunyi dari sekumpulan variabel atmosfer yang saling berkorelasi. Pendekatan ini tidak hanya membantu dalam memahami hubungan antar variabel, tetapi juga dalam

membangun model klasifikasi yang mampu memetakan kondisi atmosfer ke dalam kategori cuaca tertentu seperti Clouds, Rain, Thunderstorm, dan Clear.

Sejalan dengan pendekatan tersebut, proyek ini mengaplikasikan dua metode utama dalam analisis multivariat yaitu Analisis Diskriminan dan Regresi Multinomial. Analisis diskriminan merupakan suatu metode yang digunakan untuk mengklasifikasikan suatu individu (objek) ke dalam suatu kelompok[3]. Sedangkan, Regresi multinomial merupakan salah satu metode analisis yang digunakan dalam mencari hubungan antara variabel respon yang terdiri lebih dari dua kategori dengan suatu variabel prediktor yang bersifat kategori maupun kontinu[4]. Regresi Logistik Multinomial (MLR) juga dikenal sebagai Regresi Soft-max adalah metode klasifikasi yang menggeneralisasi regresi logistik multikelas yaitu, dengan lebih dari dua kemungkinan hasil yang berbeda[5].

Pada penelitian sebelumnya yang berjudul *Weather Prediction Using Multi Linear Regression Algorithm* yang ditulis oleh N. Anusha et al. (2019), metode Multi Linear Regression diterapkan untuk memprediksi curah hujan di wilayah Uttar Pradesh, India. Penelitian tersebut berfokus pada peramalan nilai kuantitatif curah hujan, namun belum mengklasifikasikan kondisi cuaca ke dalam kategori seperti Clouds, Rain, Thunderstorm, dan Clear yang lebih aplikatif untuk pengambilan keputusan. Berbeda dengan penelitian tersebut, penelitian ini menggunakan pendekatan klasifikasi untuk mengelompokkan kondisi cuaca di Kota Denpasar berdasarkan variabel atmosfer.

Analisis ini difokuskan untuk mengidentifikasi hubungan antar variabel atmosfer, meliputi suhu, kelembapan, tekanan udara, kecepatan angin, dan curah hujan, dalam upaya klasifikasi kondisi cuaca di Kota Denpasar pada tahun 2019 secara komprehensif. Permasalahan utama yang dikaji adalah efektivitas penerapan metode multivariat, dalam mengembangkan model klasifikasi cuaca dengan kategori Clouds, Rain, Thunderstorm, dan Clear. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi variabel-variabel yang memberikan kontribusi signifikan terhadap pembentukan kategori cuaca serta mengevaluasi tingkat akurasi dan kemudahan interpretasi model yang dihasilkan. Tujuan penelitian ini adalah menerapkan analisis diskriminan dan regresi multinomial untuk membangun model klasifikasi kondisi atmosfer di Kota Denpasar yang akurat dan dapat diinterpretasikan dengan mudah. Melalui pendekatan tersebut, penelitian berupaya menilai peran variabel atmosfer dalam meningkatkan performa model klasifikasi serta menyediakan alat yang dapat digunakan oleh praktisi dan peneliti di bidang meteorologi dan pengelolaan risiko cuaca.

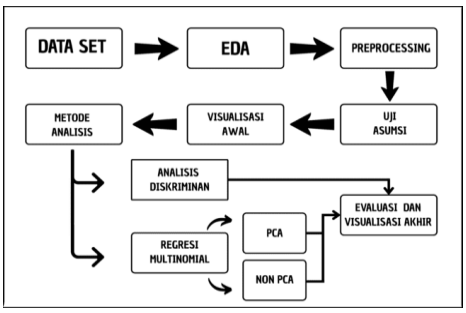
II. METODE PENELITIAN

1. Dataset

Dataset yang digunakan dalam proyek ini berasal dari Kaggle dengan judul "Denpasar Bali Historical Weather Data", yang berisi data cuaca historis Kota Denpasar dari tahun 1990 hingga 2020 dengan total 264.925 baris data. Target variabel klasifikasi yaitu kolom weather_main memiliki beberapa kategori unik, antara lain Clouds, Rain, Thunderstorm, Clear, Haze, Smoke, dan Mist. Namun, untuk kebutuhan proyek ini, fokus klasifikasi diarahkan pada empat kategori utama yang relevan dengan kondisi cuaca Denpasar tahun 2019, yaitu Clouds, Rain, Thunderstorm, dan Clear.

2. Langkah Penelitian

2.1 Diagram Alir



Gambar 1. Diagram Alir

2.2 Exploratory Data Analysis (EDA)

Proses analisis diawali dengan pemuatan dan pemeriksaan awal terhadap dataset cuaca guna memastikan integritas struktur data. Dataset dibaca dan ditampilkan sebagian untuk memverifikasi kesesuaian format, tipe data, serta konsistensi variabel. Selanjutnya, dilakukan pengukuran dimensi dataset melalui perhitungan jumlah baris dan kolom, yang bertujuan untuk memperoleh gambaran awal mengenai skala dan kompleksitas data. Tahapan berikutnya adalah identifikasi nilai hilang (missing values) dengan menghitung jumlah entri NA pada masing-masing variabel, sebagai dasar dalam pengambilan keputusan terhadap strategi penanganan yang tepat, baik melalui imputasi maupun penghapusan. Selain itu, dilakukan pula analisis distribusi pada variabel target weather_main dengan menghitung frekuensi kemunculan masing-masing kategori cuaca. Langkah ini penting untuk mengevaluasi proporsi kelas dan mengidentifikasi potensi ketidakseimbangan data yang dapat memengaruhi performa model klasifikasi.

2.3 Pre-Processing

Dalam proses pre-processing data, dilakukan seleksi terhadap kolom-kolom numerik yang relevan seperti suhu (temp, temp_min, temp_max), tekanan (pressure), kelembaban (humidity), kecepatan dan arah angin (wind_speed, wind_deg), serta tingkat awan (clouds_all), sambil menghapus kolom dengan banyak nilai hilang. Seleksi fitur dilakukan dengan memilih variabel yang berkaitan langsung dengan kolom target serta mempertimbangkan korelasi antar fitur. Data kemudian difilter hanya untuk empat kategori cuaca utama (Clouds, Rain, Thunderstorm, dan Clear) guna menyederhanakan klasifikasi menjadi multi-kelas terbatas. Variabel target weather_main diubah ke format numerik (0–3) melalui label encoding, dan dataset dibagi menjadi data latih (70%) dan data uji (30%) menggunakan createDataPartition dari library caret dengan menjaga proporsi kelas. Distribusi data target setelah filtering menunjukkan adanya ketidakseimbangan kelas, yaitu Clear (Kelas 0) sebanyak 22 data, Clouds (Kelas 1) sebanyak 8075 data, Rain (Kelas 2) sebanyak 571 data, dan Thunderstorm (Kelas 3) sebanyak 143 data. Untuk mengatasi ketidakseimbangan data latih, dilakukan balancing melalui oversampling, undersampling, dan metode SMOTE. Selain itu, korelasi antar fitur numerik dianalisis menggunakan heatmap (corrplot) untuk mengidentifikasi hubungan linear yang dapat mendukung pemilihan fitur lebih lanjut.

2.4 Normalisasi Min Max

Min-max normalization adalah teknik yang sangat umum digunakan untuk menyesuaikan fitur pada dataset ke rentang [0,1]. Teknik ini membantu mempercepat konvergensi pada algoritma yang menggunakan optimasi gradien (seperti neural network) dan mencegah fitur dengan rentang besar mendominasi fitur lain, khususnya pada algoritma berbasis jarak seperti KNN[6].

2.5 Uji Asumsi Analisis Diskriminan (Menggunakan data yang belum dinormalisasi)

2.5.1 Uji Normalitas per Fitur dalam Tiap Kelas (Shapiro-Wilk Test)

Menurut Quraisy (2020), uji Shapiro-Wilk merupakan salah satu metode yang umum dan efektif digunakan untuk menguji normalitas data dengan sampel kecil hingga sedang [7]. Karena uji Shapiro-Wilk hanya berlaku untuk pengujian univariat, maka pengujian dilakukan secara terpisah pada masing-masing kolom fitur (misalnya temperatur, kelembapan, dan lain-lain) di setiap kelas. Jika nilai p dari uji Shapiro-Wilk > 0.05, maka distribusi data fitur tersebut dianggap normal dalam kelas yang bersangkutan.

2.5.2 Uji Homogenitas Matriks Kovarian (Box's M Test)

Box's M test merupakan asumsi penting dalam analisis multivariat varians (MANOVA) dan juga dalam analisis diskriminan[8]. Uji ini bertujuan untuk mengevaluasi apakah matriks kovarians antar kelompok cuaca bersifat homogen, yang merupakan syarat penting dalam Linear Discriminant Analysis (LDA). Jika nilai p (p-value) dari uji ini lebih besar dari 0,05, maka dapat disimpulkan bahwa tidak terdapat perbedaan signifikan pada matriks kovarians antar kelompok, sehingga asumsi homogenitas terpenuhi.

2.5.3 Uji Normalitas Multivariat (Mardia)

Uji Mardia mengevaluasi normalitas multivariat dengan menghitung dua statistik utama yaitu skewness (kemencengan) dan kurtosis (keruncingan). Jika nilai p untuk kedua statistik ini lebih besar dari 0,05, maka asumsi normalitas multivariat dianggap terpenuhi. Namun, jika salah satu atau kedua nilai p kurang dari 0,05, maka asumsi tersebut tidak terpenuhi, yang dapat memengaruhi validitas hasil analisis diskriminan. Dalam penelitian oleh Anis et al. (2021), dibandingkan beberapa metode uji normalitas multivariat, termasuk Uji Mardia. Hasilnya menunjukkan bahwa Uji Mardia memiliki kekuatan uji yang lebih tinggi dan tingkat kesalahan tipe II yang lebih rendah dalam berbagai kondisi data, menjadikannya pilihan yang andal untuk menguji normalitas multivariat dalam konteks LDA [9].

2.6 Uji Asumsi untuk Multinomial Logistic Regression (data yang sudah dinormalisasi)

2.6.1 Uji Multikolinearitas (Variance Inflation Factor)

Dalam studi oleh Sura et al. (2023), VIF digunakan sebagai bagian dari evaluasi model untuk mengidentifikasi multikolinearitas dan memastikan validitas hasil regresi [10]. Nilai VIF yang tinggi menunjukkan adanya multikolinearitas yang dapat mempengaruhi keandalan model. Sebagai pedoman umum, nilai VIF di atas 10 dianggap menunjukkan multikolinearitas yang signifikan dan memerlukan perhatian lebih lanjut.

2.6.2 Uji Chi-Square

Uji Chi-Square digunakan untuk menilai kesesuaian model (goodness-of-fit) dan signifikansi prediktor. Jika nilai p dari uji Chi-Square ini signifikan (biasanya $p < 0,05$), maka kita menyimpulkan bahwa model secara keseluruhan lebih baik daripada model tanpa prediktor, artinya setidaknya ada satu variabel independen yang secara signifikan mempengaruhi variabel dependen. Selain itu, uji Chi-Square juga dapat dilakukan pada tiap prediktor untuk melihat kontribusi individunya terhadap model. Namun, penting untuk memastikan bahwa frekuensi yang diharapkan dalam setiap kategori tidak terlalu kecil, agar uji Chi-Square tetap valid [11].

2.7 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) adalah metode statistik yang digunakan untuk memetakan data multivariat ke dalam ruang dua atau tiga dimensi sehingga hubungan atau kemiripan antar objek dapat divisualisasikan secara lebih jelas. MDS bekerja dengan mengubah informasi mengenai jarak atau kemiripan antar pasangan objek menjadi posisi titik dalam ruang visual, sehingga objek-objek yang mirip akan terletak berdekatan, dan yang tidak mirip akan berjauhan [12].

III. HASIL & PEMBAHASAN

1. Uji Asumsi

1.1 Analisis Diskriminan

1.1.1 Uji Normalitas Tiap Fitur per Kelas (Shapiro-Wilk Test)

```
Kelas: 0
  temp: W=0.9501, p=0.0000
  temp_min: W=0.9297, p=0.0000
  temp_max: W=0.9372, p=0.0000
  pressure: W=0.9269, p=0.0000
  humidity: W=0.9227, p=0.0000
  wind_speed: W=0.9064, p=0.0000
  wind_deg: W=0.8181, p=0.0000
  clouds_all: W=0.5711, p=0.0000
```

Gambar 2a.

```
Kelas: 1
  temp: W=0.9913, p=0.0000
  temp_min: W=0.9909, p=0.0000
  temp_max: W=0.9860, p=0.0000
  pressure: W=0.9777, p=0.0000
  humidity: W=0.9642, p=0.0000
  wind_speed: W=0.9659, p=0.0000
  wind_deg: W=0.8957, p=0.0000
  clouds_all: W=0.6243, p=0.0000
```

Gambar 2b.

Gambar 2 : (2a) Shapiro-Wilk Test Kelas 0 (2b) Shapiro-Wilk Test Kelas 1

Seluruh fitur pada kelas 0 menunjukkan nilai p < 0.05, yang menandakan bahwa tidak ada fitur yang berdistribusi normal. Fitur 'clouds_all' dan 'wind_deg' memiliki nilai W paling rendah (0.5711 dan 0.8181), menunjukkan penyimpangan terbesar dari distribusi normal. Dengan demikian, asumsi normalitas tidak terpenuhi pada kelas ini. Pada kelas 1, meskipun beberapa fitur seperti 'temp' dan 'temp_min' memiliki nilai W mendekati 1, seluruh p-value tetap < 0.05, yang berarti semua fitur juga tidak memenuhi asumsi normalitas. Fitur 'clouds_all' (W = 0.6243) dan 'wind_deg' (W = 0.8957) merupakan fitur dengan penyimpangan paling besar.

```
Kelas: 2
  temp: W=0.9908, p=0.0000
  temp_min: W=0.9812, p=0.0000
  temp_max: W=0.9809, p=0.0000
  pressure: W=0.9644, p=0.0000
  humidity: W=0.9269, p=0.0000
  wind_speed: W=0.8685, p=0.0000
  wind_deg: W=0.8491, p=0.0000
  clouds_all: W=0.8317, p=0.0000
```

Gambar 3a.

```
Kelas: 3
  temp: W=0.9769, p=0.0000
  temp_min: W=0.9738, p=0.0000
  temp_max: W=0.9716, p=0.0000
  pressure: W=0.9638, p=0.0000
  humidity: W=0.9331, p=0.0000
  wind_speed: W=0.8883, p=0.0000
  wind_deg: W=0.8891, p=0.0000
  clouds_all: W=0.7901, p=0.0000
```

Gambar 3b.

Gambar 3 : (3a) Shapiro-Wilk Test Kelas 2 (3b) Shapiro-Wilk Test Kelas 3

Fitur-fitur pada kelas 2 juga tidak berdistribusi normal, dengan seluruh p-value < 0.05. Meskipun beberapa fitur seperti 'temp_min' dan 'temp_max' memiliki nilai W tinggi (di atas 0.98), fitur 'clouds_all', 'wind_speed', dan wind_deg menunjukkan deviasi yang cukup besar dari distribusi normal. Kelas 3 menunjukkan hasil serupa, dengan semua fitur tidak memenuhi asumsi normalitas. Fitur 'clouds_all' dan 'wind_deg' kembali menjadi fitur dengan penyimpangan tertinggi (W = 0.7901 dan 0.8891), sementara fitur suhu ('temp, temp_min', 'temp_max') juga menunjukkan deviasi meski nilainya lebih mendekati distribusi normal.

1.1.2 Uji Homogenitas Matriks Kovarian (Box's M Test)

```
Box's M-test for Homogeneity of Covariance Matrices

data:  train_balanced[, feature_cols]
Chi-Sq (approx.) = 20650, df = 108, p-value < 2.2e-16
```

Gambar 4. Box's M Test

Hasil uji menunjukkan nilai p < 2.2e-16, yang jauh di bawah ambang batas 0.05. Ini berarti terdapat perbedaan signifikan antar matriks kovarian tiap kelas, sehingga asumsi homogenitas tidak terpenuhi.

1.1.3 Uji Normalitas Multivariat (Mardia)

Tabel 1. Mardia

Test	Statistic	p value	Result

Mardia Skewness	42309.845377437	0	NO
data	68.1771549865525	0	NO

Hasil pengujian menunjukkan bahwa data tidak memenuhi asumsi tersebut, sebagaimana ditunjukkan oleh nilai skewness dan kurtosis Mardia. Pada uji skewness, diperoleh nilai statistik sebesar 42.309,85 dengan p-value sebesar 0, sedangkan pada uji kurtosis, nilai statistik tercatat sebesar 68,18 dengan p-value yang juga sama dengan 0. Kedua nilai p-value berada jauh di bawah tingkat signifikansi 0,05, sehingga hipotesis nol yang menyatakan bahwa data mengikuti distribusi normal multivariat ditolak. Dengan demikian, dapat disimpulkan bahwa data yang telah diolah dalam subset train_balanced tidak berdistribusi normal secara multivariat.

1.2 Multinomial Logistic Regression

1.2.1 Uji Multikolinearitas (VIF)

Tabel 2. VIF

pressure	Humidity	wind_speed	wind_deg	clouds_all	pca_temp
1.231212	2.452965	1.197297	1.132742	1.352676	2.167939

Hasil uji Variance Inflation Factor (VIF) menunjukkan bahwa seluruh variabel prediktor memiliki nilai VIF di bawah 5, dengan nilai tertinggi pada variabel humidity (2.45) dan pca_temp (2.17). Hal ini mengindikasikan tidak adanya masalah multikolinearitas yang signifikan di antara variabel-variabel dalam model. Oleh karena itu, asumsi VIF dalam regresi linier telah terpenuhi.

1.2.2 Uji Chi-Square

Tabel 3. Chi-Square

variable	p_value
pressure	1.228649e-254
humidity	0.000000e+00
wind_speed	0.000000e+00
wind_deg	0.000000e+00
clouds_all	0.000000e+00
pca_temp	3.199718e-153

Nilai p-value yang sangat kecil (< 0.05) menunjukkan bahwa terdapat hubungan yang signifikan secara statistik antara masing-masing variabel numerik tersebut dengan variabel target weather_main. Artinya, seluruh variabel tersebut berkontribusi secara signifikan dalam membedakan kondisi cuaca, dan layak dipertahankan sebagai fitur dalam model yang berarti uji asumsi telah terpenuhi.

III.1 PreProcessing

III.2.1 Seleksi Kolom

Tahapan ini menghasilkan hanya kolom-kolom numerik yang relevan telah dipertahankan, yaitu temp, temp_min, temp_max, pressure, humidity, wind_speed, wind_deg, dan clouds_all, serta target klasifikasi weather_main.

III.2.2 Filter Kategori Target

temp	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	cloud _all	weather_main
27.1	27	27.4	1010	94	0.5	0	40	Rain
27.8	27.4	28	1010	88	2.1	240	20	Rain
29.5	29	30.4	1010	83	3.1	240	20	Clouds
30.1	30	30.4	1010	79	4.1	260	20	Clouds

30.2	30	30.4	1009	79	4.1	260	20	Clouds
30	30	30	1008	79	5.1	250	20	Clouds

Data telah difilter berdasarkan kategori target 'weather_main'. Kategori cuaca yang terlalu jarang kemunculannya dihapus dari data, sehingga hanya menyisakan kelas-kelas utama yang frekuensinya mencukupi untuk analisis, seperti Rain, Clouds, dan Clear.

III.2.3 Label Encoding

Proses label encoding telah dilakukan pada variabel target weather_main. Nilai kategorikal seperti Rain, Clouds, dan Clear diubah menjadi bentuk numerik, contohnya menjadi 2, 1, 0, dan 3. Selain itu, dipastikan bahwa variabel ini bertipe faktor biasa (bukan ordered factor), sehingga tidak ada asumsi urutan di antara kelas. Langkah ini penting agar algoritma klasifikasi seperti Multinomial Logistic Regression dapat memproses target secara numerik tanpa bias urutan.

III.2.4 Penanganan Missing Values

Penanganan missing values dilakukan dengan mengisi nilai hilang pada variabel numerik menggunakan rata-rata, sedangkan pada target, tidak digunakan tidak terdapat missing values sehingga tidak dilakukan penanganan terhadapnya.

III.2.5 Split Data

Proses split data dilakukan dengan membagi dataset menjadi dua bagian (data sebanyak 70% data (6170 observasi) digunakan sebagai data latih, sedangkan sisanya 30% (2641 observasi) digunakan sebagai data uji. Pembagian ini mempertahankan proporsi kategori dari variabel target, sehingga model yang dilatih tetap representatif terhadap distribusi kelas aslinya.

III.2.6 Balancing Kelas pada Data Train

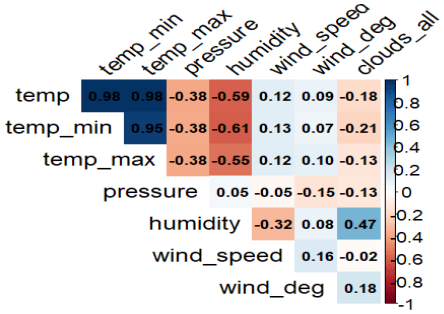
Untuk mengatasi ketidakseimbangan kelas pada variabel target, dilakukan penyeimbangan data train menggunakan metode resampling yang menggabungkan undersampling dan oversampling sederhana. Jumlah data train awal dikalikan dua untuk menentukan total data setelah penyeimbangan, kemudian nilai tersebut dibagi empat (jumlah kategori cuaca) guna memperoleh target jumlah data per kategori. Jika jumlah data suatu kategori melebihi target tersebut, dilakukan undersampling secara acak untuk menguranginya. Sebaliknya, jika jumlahnya kurang, dilakukan oversampling dengan pengambilan sampel ulang (dengan pengembalian) hingga mencapai jumlah yang ditentukan. Setelah dilakukan balancing pada train data, distribusi kelas target memiliki jumlah observasi yang sama yaitu 3.085 data, sehingga total data train setelah penyeimbangan adalah 12.340 data. Langkah ini hanya diterapkan pada data train agar model tidak bias terhadap kategori mayoritas, sedangkan data test dibiarkan dalam kondisi asli agar evaluasi performa tetap mencerminkan distribusi cuaca yang sebenarnya.

III.2.7 Normalisasi

Pada tahapan ini semua fitur numerik telah berhasil dinormalisasi ke rentang 0 hingga 1. Ini mengindikasikan bahwa proses normalisasi berjalan dengan baik dan konsisten untuk data pelatihan.

E-ISSN 2808-5841
P-ISSN 2808-7283

III.2.8 Visualisasi Korelasi



Gambar 5. Visualisasi Korelasi

Visualisasi korelasi di atas menunjukkan hubungan antara berbagai variabel cuaca seperti suhu (temp, temp_min, temp_max), tekanan (pressure), kelembapan (humidity), kecepatan angin (wind_speed), arah angin (wind_deg), dan tutupan awan (clouds_all). Terlihat bahwa suhu, suhu minimum, dan suhu maksimum memiliki korelasi positif yang sangat kuat satu sama lain sebesar 0.98, yang menunjukkan bahwa peningkatan satu jenis suhu cenderung diikuti oleh peningkatan jenis suhu lainnya. Secara keseluruhan, korelasi yang kuat hanya tampak pada kelompok suhu, sementara variabel lainnya menunjukkan hubungan yang lemah atau tidak signifikan.

III.2 Analisis Diskriminan

```
          Reference
Prediction    0    1    2    3
         0    6  299    2    1
         1    0 1666   29   10
         2    0  231   85   12
         3    0  226   55   19

Overall Statistics

               Accuracy : 0.6725
                 95% CI : (0.6542, 0.6904)
    No Information Rate : 0.9171
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1767
```

Gambar 6. Analisis Diskriminan

Model DA menghasilkan akurasi sebesar 67.25% dengan interval kepercayaan 95% antara 65.42% hingga 69.04%. Nilai Kappa sebesar 0.1767 menunjukkan tingkat kesepakatan yang rendah antara prediksi model dan label aktual, setelah memperhitungkan kesepakatan acak. P-Value dari McNemar's Test yang sangat kecil (< 2e-16) mengindikasikan bahwa terdapat perbedaan signifikan secara statistik antara distribusi kesalahan model, yang menyiratkan ketidakseimbangan dalam klasifikasi antar kelas.

Tabel 5. Evaluasi Analisis Diskriminan

	class : 0	**class : 1**	**class : 2**	**class : 3**
Sensitivity	1.000000	0.6879	0.49708	0.452381
Specificity	0.885389	0.8219	0.90162	0.891881
Pos Pred Value	0.019481	0.9771	0.25915	0.063333
Neg Pred Value	1.000000	0.1923	0.96282	0.990175
Prevalence	0.002272	0.9171	0.06475	0.015903
Detection Rate	0.002272	0.6308	0.03218	0.007914
Detection Prevalence	0.116622	0.6456	0.12420	0.113593
Balanced Accuracy	0.942694	0.7549	0.69935	0.672131

- Class 0 memiliki sensitivitas dan nilai prediksi negatif sempurna, namun nilai prediksi positifnya sangat rendah (0.0195), karena prevalensinya juga sangat kecil (0.0023).
- Class 1 adalah yang paling dominan (prevalensi 91.71%) dengan sensitivitas yang cukup tinggi (0.6879) dan nilai prediksi positif tinggi (0.9771).
- Class 2 dan Class 3 menunjukkan performa yang jauh lebih rendah, dengan sensitivitas masing-masing hanya 0.4971 dan 0.4524, serta nilai prediksi positif di bawah 0.3, menunjukkan bahwa model kesulitan dalam mengklasifikasi kelas-kelas minoritas ini.

III.3 Multinomial Logistic Regression Tanpa PCA

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3
         0    6    2    0    0
         1    0 1918   40   10
         2    0  284   75    7
         3    0  218   56   25

Overall Statistics

               Accuracy : 0.7664
                 95% CI : (0.7498, 0.7824)
    No Information Rate : 0.9171
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2361
```

Gambar 7. Multinomial Logistic Regression Tanpa PCA

Model Multinomial Logistic Regression (MLR) tanpa PCA menunjukkan akurasi sebesar 76.64%, lebih tinggi dibandingkan model LDA sebelumnya. Nilai Kappa sebesar 0.2361 menunjukkan adanya peningkatan kesesuaian prediksi dibanding LDA, meskipun masih tergolong rendah.

Tabel 6. Evaluasi Multinomial Logistic Regression Tanpa PCA

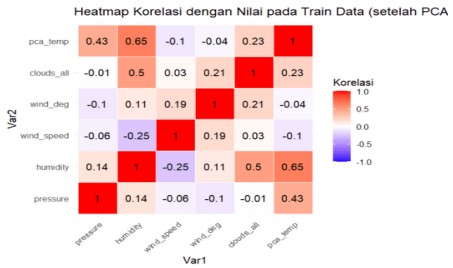
	class : 0	class : 1	class : 2	class : 3
Sensitivity	1.000000	0.7919	0.43860	0.595238
Specificity	0.999241	0.7717	0.88219	0.894575
Pos Pred Value	0.750000	0.9746	0.20492	0.083612
Neg Pred Value	1.000000	0.2511	0.95780	0.992741
Prevalence	0.002272	0.9171	0.06475	0.009466
Detection Rate	0.002272	0.7262	0.02840	0.009466
Detection Prevalence	0.003029	0.7452	0.13858	0.113215
Balanced Accuracy	0.999620	0.7818	0.66039	0.744906

- Class 0, model mampu mengklasifikasikan dengan sangat baik, dengan sensitivitas dan spesifisitas mendekati 1. Namun, prevalensi kelas ini sangat rendah (0.002272), sehingga keberhasilan tersebut tidak terlalu berdampak besar secara keseluruhan.
- Class 1 sebagai kelas mayoritas memiliki sensitivitas tinggi (0.7919) dan nilai prediksi positif sangat baik (0.9746), meskipun nilai negatif prediksinya rendah (0.2511), menunjukkan bahwa banyak prediksi negatif yang salah.
- Untuk Class 2 dan 3, performa model masih kurang optimal. Sensitivitasnya cukup rendah, serta nilai prediksi positifnya juga sangat rendah (0.20492 dan 0.083612), mengindikasikan banyak kesalahan dalam mengidentifikasi instance positif dari kelas-kelas ini. Hal ini mengindikasikan bias terhadap kelas mayoritas, yang terlihat juga dari nilai Detection Prevalence yang jauh lebih tinggi pada Class 1.

III.4 Multinomial Logistic Regression menggunakan PCA

Matriks korelasi pada visualisasi korelasi menunjukkan adanya korelasi yang sangat tinggi antara variabel suhu, yaitu antara temp dan temp_min (0.98), temp dan temp_max (0.98), serta temp_min dan temp_max (0.95). Korelasi tinggi ini mengindikasikan adanya redundansi informasi, yang dapat menyebabkan multikolinearitas jika digunakan secara langsung dalam model prediktif. Dengan struktur korelasi seperti ini, penerapan PCA menjadi relevan karena dapat mereduksi dimensi. Principal Component Analysis (PCA) merupakan suatu teknik yang

menganalisis suatu tabel data observasi menjadi tabel data baru yang memiliki korelasi serupa [13].



Gambar 8. HeatMap Korelasi Train Data (Setelah PCA)

Setelah PCA diterapkan, ketiga variabel suhu tersebut direduksi menjadi satu komponen utama, yang dinamakan pca_temp. Komponen ini berhasil merangkum informasi utama dari ketiga variabel suhu awal, sekaligus mengeliminasi redundansi yang ada. Korelasi pca_temp dengan variabel lain seperti humidity dan pressure menunjukkan distribusi yang lebih merata dan tidak terlalu dominan, yang menandakan pengurangan multikolinearitas. Dengan demikian, penggunaan pca_temp sebagai variabel tunggal tidak hanya menyederhanakan model, tetapi juga meningkatkan kestabilan dan kemampuan generalisasi model prediktif, serta mengurangi risiko overfitting.

```
Prediction    0    1    2    3
         0    6    1    0    0
         1    0 1930   37    8
         2    0  275   78   11
         3    0  216   56   23

Overall Statistics

               Accuracy : 0.7713
                 95% CI : (0.7548, 0.7872)
    No Information Rate : 0.9171
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2464
```

Gambar 9. Multinomial Logistic Regression dengan PCA

Hasil evaluasi model klasifikasi menunjukkan bahwa akurasi keseluruhan adalah 77.13% dengan interval kepercayaan 95% antara 75.48% hingga 78.72%. Namun, nilai No Information Rate (NIR) sebesar 91.71% menunjukkan bahwa model belum melampaui prediksi dasar secara signifikan (P-Value = 1). Nilai Kappa sebesar 0.2464 mengindikasikan tingkat kesepakatan rendah antara prediksi dan referensi setelah mengoreksi kemungkinan kesepakatan acak.

Tabel 7. Evaluasi Multinomial Logistic Regression dengan PCA

	class : 0	**class : 1**	**class : 2**	**class : 3**
Sensitivity	1.000000	0.7969	0.45614	0.547619
Specificity	0.999620	0.7945	0.88421	0.895344
Pos Pred Value	0.857143	0.9772	0.21429	0.077966
Neg Pred Value	1.000000	0.2613	0.95916	0.991901
Prevalence	0.002272	0.9171	0.06475	0.015903
Detection Rate	0.002272	0.7308	0.02953	0.008709
Detection Prevalence	0.002651	0.7478	0.13783	0.111700
Balanced Accuracy	0.999810	0.7957	0.67018	0.721482

- Untuk kelas 0, sensitivitas dan spesifisitas hampir sempurna (1.000 dan 0.9996), namun ini bisa disebabkan oleh prevalensi yang sangat rendah (0.22%).
- Kelas 1 menunjukkan performa terbaik.

- Sebaliknya, kelas 2 dan 3 memiliki sensitivitas rendah (45.61% dan 54.76%) serta nilai prediksi positif yang juga rendah, menandakan bahwa model mengalami kesulitan membedakan kedua kelas ini.
- Secara umum, meskipun akurasi cukup tinggi, model cenderung bias terhadap kelas mayoritas dan memiliki performa yang kurang optimal pada kelas-kelas minoritas.

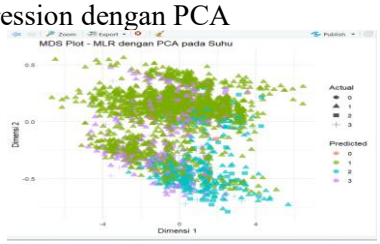
III.5 MDS 2 Metode Terbaik

III.6.1 Multinomial Logistic Regression Tanpa PCA

Gambar 10. MDS Multinomial Logistic Regression Tanpa PCA

Plot MDS menunjukkan bahwa model Multinomial Logistic Regression sudah mampu mengenali pola umum dalam data, meskipun masih terdapat tumpang tindih antar kelas. Prediksi model cukup konsisten, terutama pada kelas 1 yang banyak teridentifikasi. Ini menunjukkan potensi model dalam mengklasifikasi cuaca, meski masih perlu peningkatan untuk membedakan kelas-kelas yang lebih mirip secara karakteristik.

III.6.2 Multinomial Logistic Regression dengan PCA



Gambar 11. MDS Multinomial Logistic Regression dengan PCA

Plot MDS untuk MLR dengan PCA pada suhu menunjukkan pola klasifikasi yang cukup terarah. PCA membantu menyederhanakan informasi suhu, yang terlihat dari distribusi prediksi yang lebih menyebar dan tetap konsisten, khususnya pada kelas 1 dan 2. Ini menandakan bahwa penerapan PCA mendukung kinerja model dalam mengenali struktur data secara lebih efisien, meskipun masih ada tumpang tindih antar kelas.

III.7 Perbandingan

Tabel 8. Perbandingan

metode	akurasi
Analisis Diskriminan	0.6687
MLR tanpa PCA	0.7664
MLR dengan PCA	0.7713

Berdasarkan tabel perbandingan, Multinomial Logistic Regression (MLR) dengan PCA menunjukkan akurasi tertinggi sebesar 0.7713, disusul oleh MLR tanpa PCA sebesar 0.7664, dan yang terendah adalah Linear Discriminant Analysis (LDA) dengan 0.6687. Hasil ini menunjukkan bahwa MLR lebih unggul dalam memodelkan data dibandingkan LDA, baik dengan maupun tanpa PCA. Penerapan PCA sedikit meningkatkan kinerja MLR, sehingga

dapat diasumsikan bahwa PCA berhasil mereduksi dimensi dan mengurangi multikolinearitas tanpa menghilangkan informasi penting.

IV. KESIMPULAN

Model Multinomial Logistic Regression (MLR) dengan PCA memberikan hasil paling baik dibanding dua model lainnya, dengan akurasi tertinggi sebesar 77,13%. Hal ini menunjukkan bahwa penggunaan PCA sedikit membantu meningkatkan performa model MLR, meskipun selisihnya tidak terlalu signifikan dibanding MLR tanpa PCA (76,64%). Kedua model MLR secara konsisten menunjukkan performa yang lebih unggul dibandingkan Linear Discriminant Analysis (LDA), yang hanya mencapai akurasi 66,87%. Visualisasi Plot MDS juga menunjukkan bahwa model MLR mampu mengenali pola umum dalam data, terutama pada kelas 1 yang teridentifikasi cukup konsisten, dan penerapan PCA membantu memperjelas sebaran data meskipun masih terdapat tumpang tindih antar kelas. Dengan demikian, MLR, baik dengan maupun tanpa PCA, dapat dijadikan sebagai pendekatan yang andal dalam membangun model klasifikasi cuaca, terutama karena kestabilan dan kemampuannya dalam mengenali pola pada data.

V. DAFTAR PUSTAKA

[1] N. Anusha, M. Sai Chaithanya, and G. Jithendranath Reddy, "Weather prediction using multi linear regression algorithm," in _IOP Conference Series: Materials Science and Engineering_, vol. 590, 2019, Art. no. 012034, doi: 10.1088/1757-899X/590/1/012034.

[2] D. U. Wustqa, E. Listyani, R. Subekti, R. Kusumawati, M. Susanti, and Kismiantini, "Analisis data multivariat dengan program R," _J. Pengabdi. Masy. MIPA Pendidik. MIPA_, vol. 2, no. 2, pp. 83–86, 2018. [Online]. Available: http://journal.uny.ac.id/index.php/jpmmp

[3] M. S. Tjahaya, Raupong, and G. M. Tinungki, "Analisis diskriminan linear robust dengan metode Winsorized modified one-step M-estimator," _Estimasi_, vol. 3, no. 1, pp. 1–13, 2022, doi: 10.20956/ejsa.vi.11302.

[4] E. Novitasari and A. Sofro, "Analisis regresi multinomial untuk pemodelan faktor penyebab kekerasan dalam rumah tangga," _MATHunesa_, vol. 11, no. 1, 2023. [Online]. Available: https://journal.unesa.ac.id/index.php/mathunesa

[5] I. Qutab, K. I. Malik, and H. Arooj, "Sentiment classification using multinomial logistic regression on Roman Urdu text," _Int. J. Innov. Sci. Technol._, vol. 4, no. 2, pp. 323–335, 2022, doi: 10.33411/IJIST/2022040204.*(references)*

[6] P. J. Muhammad Ali, "Investigating the impact of min-max data normalization on the regression performance of k-nearest neighbor with different similarity measurements," _ARO Sci. J. Koya Univ._, vol. 10, pp. 85–91, 2022, doi: 10.14500/aro.10955.

[7] A. Quraisy, "Data Normality Using Kolmogorov-Smirnov and Shapiro-Wilk Tests," J-HEST Journal of Health Education Economics Science and Technology, vol. 3, no. 1, 2020.

[8] K. Jiamwattanapong, N. Ingadapa, and B. Plubin, "On Testing Homogeneity of Covariance Matrices with Box's M and the Approximate Tests for Multivariate Data," *European Journal of Applied Sciences*, vol. 9, no. 5, pp. 426–436, Oct. 2021, doi: 10.14738/aivp.95.11115..

[9] W. Anis, Kuntoro, and S. Melaniani, "DIFFERENCE OF POWER TEST AND TYPE II ERROR (β) ON MARDIA MVN TEST, HENZE ZIKLER'S MVN TEST, AND ROYSTON'S MVN TEST USING MULTIVARIATE DATA ANALYSIS," Jurnal Biometrika dan Kependudukan, vol. 10, no. 2, 2021, doi: 10.20473/jbk.v10i2.2021.153-161.

[10] J. S. Sura, R. Panchal, and A. Lather, "Economic value-added (EVA) myths and realities: evidence from the Indian manufacturing sector," IIM Ranchi journal of management studies, vol. 2, no. 1, 2023, doi: 10.1108/irjms-03-2022-0037.

[11] J. Cuneen and D. Tobar, "Chi-square Tests," in Sport Industry Research and Analysis, 2021. doi: 10.4324/9781315212944-28.

[12] S. Fadilah, Y. Rosdiana, M. Maemunah, N. Hernawati, E. Sukarmanto, and R. Hartanto, "Multidimensional scaling (Mds): Sustainability assessment model of community economic empowerment," Polish Journal of Management Studies, vol. 24, no. 2, 2021, doi: 10.17512/pjms.2021.24.2.08.

[13] D. Sartika B. Ginting et al., "Contribution Principal Component Analysis to Optimizing Data by Reducing Product Data on Transaction," *Journal of Physics: Conference Series*, vol. 1898, no. 1, 2021, doi: 10.1088/1742-6596/1898/1/012034.