**AKSIOMA: Jurnal Matematika dan Pendidikan Matematika**
Vol. 14, No. 2 September 2023
e-ISSN 2579-7646

# Perbandingan Regresi Logistik dan *Random Forest* pada Klasifikasi Cuaca Wilayah Jawa Tengah

[1]Ayu Sulistiara, [2]Najmah Istikaanah, [3]Niken Larasati

[1,2,3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Jenderal Soedirman
Email: niken.larasati@unsoed.ac.id

***Abstrak***

*Cuaca merupakan salah satu aspek penting yang berpengaruh terhadap aktivitas manusia. Adanya perubahan cuaca yang dipengaruhi oleh berbagai faktor seperti suhu, kelembapan udara, kecepatan angin, arah angin, waktu, dan lokasi, menjadikan pentingnya untuk mengetahui kemungkinan cuaca yang akan terjadi guna menghindari dan mempersiapkan solusi dari dampak yang ditimbulkan. Kemungkinan cuaca yang akan terjadi dapat ditentukan dengan lebih akurat menggunakan metode klasifikasi cuaca yang baik. Pada penelitian ini metode klasifikasi yang digunakan adalah regresi logistik dan random forest. Peneliti membandingkan kedua metode tersebut menggunakan data cuaca di wilayah Jawa Tengah yang dibagi dalam tiga proporsi data latih yang berbeda, yaitu 60%, 70% dan 80%, dan evaluasi modelnya menggunakan nilai area under curve (AUC). Rata-rata AUC dari metode regresi logistik dan random forest berturut-turut adalah 0,6923 dan 0,7419. Berdasarkan hasil analisis kedua metode tersebut, nilai AUC tertinggi didapatkan dari hasil klasifikasi menggunakan metode random forest.*

**Kata kunci:** *klasifikasi cuaca; regresi logistik; random forest; AUC*

***Abstract***

*Weather is one of the important aspects that affect human activity. There are changes in weather that are influenced by various factors such as temperature, air humidity, wind speed, wind direction, time of day, and location, making it important to know the possibility of the weather that will occur in order to avoid and prepare solutions for the impacts. The possibility of weather that will occur can be determined more accurately using good weather classification methods. In this study the classification method used is logistic regression and random forest. Researchers compared the two methods using weather data in the Central Java region divided into three different proportions of training data, namely 60%, 70% and 80%, and evaluated the model using the area under curve (AUC) value. The average AUC of the logistic regression method and random forest were 0.6923 and 0.7419, respectively. Based on the results of the analysis of the two methods, the highest AUC value was obtained from the classification results using the random forest method.*

**Keywords**: *weather classification; logistic regression; random forest; AUC*

## A. Pendahuluan

Cuaca merupakan salah satu aspek penting dalam kehidupan manusia. Cuaca yang tidak menentu dan berubah-ubah berpengaruh terhadap aktivitas manusia seperti pertanian, perkebunan, dan penerbangan. Perubahan cuaca yang ekstrem dapat menyebabkan bencana alam yang menimbulkan kerugian material dan menghilangkan nyawa manusia. Oleh karena itu, cuaca menjadi penting untuk dimonitor sehingga potensi

terjadinya dampak dari perubahan cuaca dapat dihindari dan juga dipersiapkan solusinya.

Jawa Tengah merupakan salah satu wilayah di Indonesia yang sering mengalami cuaca ektrem seperti hujan lebat dan hujan angin. Menurut Badan Meteorologi, Klimatologi, dan Geofisika (BMKG, 2023), terjadinya cuaca ektrem tersebut disebabkan oleh adanya pola belokan angin dan pertemuan angin di wilayah Jawa Tengah, serta didukung oleh suhu muka laut di perairan Jawa Tengah dan kelembapan udara yang relatif tinggi. Selain itu, menurut BMKG (2020), terdapat faktor-faktor yang secara umum berpengaruh terhadap cuaca seperti suhu, kelembapan udara, angin, waktu, lokasi, dan lain sebagainya. Faktor-faktor tersebut membentuk cuaca yang kemudian terbagi ke dalam beberapa jenis, yaitu cerah, cerah berawan, berawan, berawan tebal, udara kabur, asap, kabut, hujan ringan, hujan sedang, hujan lebat, hujan lokal, dan hujan petir. Kemungkinan cuaca yang akan terjadi dapat ditentukan dengan lebih akurat menggunakan metode klasifikasi cuaca yang baik.

Klasifikasi merupakan proses pengelompokan objek ke dalam beberapa kelas berdasarkan karakteristik yang sama. Salah satu pemanfaatan klasifikasi adalah untuk menentukan jenis cuaca. Metode yang biasa digunakan untuk klasifikasi di antaranya yaitu regresi logistik, *support vector machine, naive bayes classifier, decision tree, random forest*, dan lain sebagainya. Beberapa penelitian telah menggunakan beberapa metode sekaligus untuk memperoleh metode yang paling optimal dalam mengklasifikasikan cuaca. Siregar (2020) menggunakan *decision tree, naive bayes,* dan *random forest* untuk klasifikasi cuaca dengan variabel bebas yang digunakan yaitu suhu, kecepatan angin, tekanan, dan curah hujan. Hasil penelitian yang dilakukan Siregar menyatakan *random forest* sebagai metode dengan tingkat akurasi lebih tinggi dibandingkan dua metode lainnya. Penelitian lain dilakukan oleh Fallo (2021) yang mengklasifikasikan cuaca Bogor, Malang, dan Jakarta Utara. Penelitian Fallo tersebut menggunakan metode *support vector machine, naive bayes classifier*, dan regresi logistik, dengan akurasi tertinggi diperoleh dari klasifikasi menggunakan regresi logistik. Berdasarkan kedua penelitian yang telah disebutkan, dapat diketahui bahwa regresi logistik dan *random forest* merupakan metode klasifikasi cuaca yang paling baik pada masing-masing penelitian.

Regresi logistik dan random forest memiliki cara kerja yang berbeda dalam melakukan klasifikasi. Klasifikasi dengan regresi logistik dilakukan melalui suatu model matematika yang menghubungkan antara variabel respon dengan variabel bebas (Hosmer, dkk., 2013), sedangkan *random forest* mengklasifikasikan objek dengan mengombinasikan sejumlah pohon keputusan (Breiman, 2001). Dalam beberapa tahun terakhir, terdapat peneliti yang membandingkan hasil dan kinerja antara metode regresi logistik dengan *random forest* dalam penelitian mereka. Purwa (2019) membandingkan metode regresi logistik dan *random forest* untuk klasifikasi rumah tangga miskin di Kabupaten Karangasem, Bali tahun 2017.

Kemudian, Tanujaya, dkk. (2020) juga melakukan penelitian serupa dalam melakukan klasifikasi fitur mode pada audio *Spotify*. Hasil penelitian yang dilakukan Purwa dan juga Tanujaya yang menggunakan data biner pada penelitian mereka sama-sama menyatakan bahwa metode *random forest* lebih baik dalam mengklasifikasi objek dibandingkan dengan regresi logistik.

Berdasarkan uraian di atas, maka pada penelitian ini akan dibandingkan antara metode regresi logistik dan *random forest* dalam mengklasifikasi cuaca di wilayah Jawa Tengah, dengan variabel respon cuaca dan variabel bebas yaitu suhu udara, kelembapan udara, arah angin, kecepatan angin, waktu, dan lokasi. Dari penelitian ini, diharapkan dapat ditentukan penggunaan metode klasifikasi yang tepat untuk menentukan jenis cuaca di wilayah Jawa Tengah.

## B. Metode Penelitian

Data yang digunakan dalam penelitian ini adalah data prakiraan cuaca BMKG di wilayah Provinsi Jawa Tengah dalam kurun waktu tujuh hari (5-11 Januari 2021) yang terdiri dari 1 variabel respon dan 6 variabel bebas, sebanyak 1880 ` teori pada penelitian Alasadi dan Bhaya (2017), Kwak dan Kim (2017), Senthilnathan (2019), Damuri, dkk. (2021), Goruunescu (2011), serta Heydarian (2022), peneliti menyusun langkah-langkah yang dilakukan dalam penelitian ini sebagai berikut:

1. Mengumpulkan dan mendeskripsikan data.
2. Melakukan data *pre-processing*
   a. Mengecek keberadaan *missing value*. Apabila terdapat *missing value* kurang dari 70% maka akan dilakukan imputasi menggunakan nilai mean (data numerik) atau nilai modus (data kategorik). Namun, apabila *missing value* lebih dari 70% maka kolom/variabel yang mengandung *missing value* tersebut akan dihapus.
   b. Mendeteksi *outlier* menggunakan boxplot. Apabila terdapat *outlier* pada data penelitian maka data pengamatan yang mengandung *outlier* akan dihilangkan.
3. Melakukan analisis korelasi dengan menghitung nilai koefisien korelasi. Variabel bebas yang akan digunakan dalam membentuk model klasifikasi adalah variabel yang memiliki nilai koefisien korelasi $R \neq 0$.
4. Membagi dataset menjadi data latih dan data uji secara acak dengan perbandingan proporsi data latih dan data uji yaitu 60:40, 70:30 serta 80:20.
5. Membentuk model klasifikasi regresi logistik dan *random forest* menggunakan data latih.
6. Melakukan klasifikasi untuk data uji dengan model klasifikasi regresi logistik dan *random forest* yang telah terbentuk.
7. Melakukan evaluasi model dengan *confusion matrix* dan pengukuran AUC.
8. Membandingkan hasil evaluasi masing-masing model berdasarkan nilai AUC.

9. Menarik kesimpulan dari hasil analisis perbandingan yang telah dilakukan.

## C. Hasil dan Pembahasan

### 1. Deskripsi data

Penelitian ini menggunakan data prakiraan cuaca yang terdiri dari 6 variabel bebas dan satu variabel respon. Tipe dan jumlah data pada masing-masing variabel disajikan pada Tabel 1.

Tabel 1. Tipe dan jumlah data berdasarkan variabel penelitian

| Variabel | | Tipe data | Jumlah data |
|---|---|---|---|
| Notasi | Nama | | |
| $X_1$ | Lokasi | Kategorik | 1880 |
| $X_2$ | Waktu | Kategorik | 1880 |
| $X_3$ | Suhu udara | Numerik | 1880 |
| $X_4$ | Kelembapan udara | Numerik | 1880 |
| $X_5$ | Arah angin | Kategorik | 1880 |
| $X_6$ | Kecepatan angin | Numerik | 1880 |
| $Y$ | Cuaca | Kategorik | 1880 |

Selanjutnya, statistika deskriptif untuk data pada variabel bebas yang bertipe numerik disajikan pada Tabel 2 berikut.

Tabel 2. Statistika deskriptif variabel bebas bertipe numerik

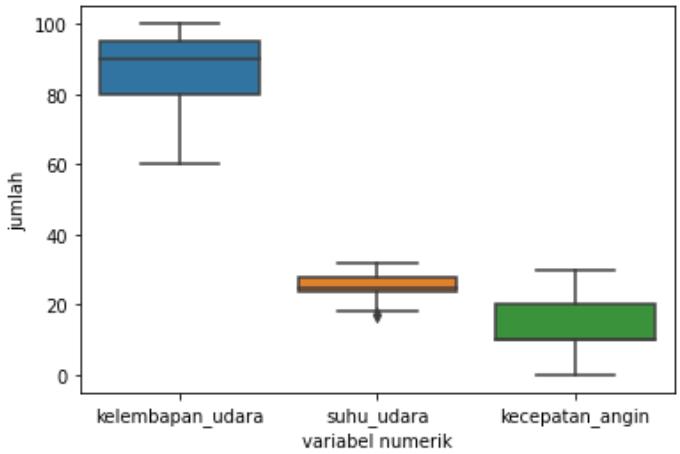
| Variabel bebas | Rata-rata | Min | Kuartil ke-1 | Median | Kuartil ke-3 | Maks |
|---|---|---|---|---|---|---|
| $X_3$ | 25,830 | 17 | 24 | 25 | 28 | 32 |
| $X_4$ | 77,173 | 60 | 65 | 80 | 90 | 95 |
| $X_6$ | 13,330 | 0 | 10 | 10 | 20 | 30 |

### 2. Data Preprocessing

Berdasarkan Tabel 1, diketahui bahwa jumlah objek pada masing-masing variabel bernilai sama sehingga tidak terdapat *missing values* pada data penelitian yang digunakan. Selanjutnya, dari nilai statistik deskriptif data yang bertipe numerik pada Tabel 2, dilakukan pengecekan *outliers* dengan melihat distribusi data menggunakan boxplot (Gambar 1). Berdasarkan Gambar 1, dapat dilihat bahwa pada variabel suhu udara terdapat *outlier* yang terletak di luar batas bawah. Adapun batas bawah dari distribusi data suhu udara adalah

$$\text{BB} = 24 - 1{,}5 \times (28 - 24) = 18.$$

Dengan demikian, untuk data dengan suhu udara yang bernilai kurang dari 18 akan dihilangkan.




Gambar 1. Boxplot untuk variabel numerik

## 3. Korelasi

Hasil perhitungan koefisien korelasi untuk masing-masing variabel $X$ terhadap variabel $Y$ sebagai variabel respon disajikan pada Tabel 3 berikut.

Tabel 3. Koefisien korelasi variabel bebas terhadap variabel cuaca

| Variabel | Koefisien korelasi terhadap $Y$ |
|---|---|
| $X_1$ | -0,003767 |
| $X_2$ | -0,065064 |
| $X_3$ | 0,140051 |
| $X_4$ | -0,120274 |
| $X_5$ | 0,074057 |
| $X_6$ | 0,091970 |

Dari Tabel 3 diketahui bahwa koefisien korelasi masing-masing variabel $X$ terhadap variabel $Y$ tidak bernilai 0 sehingga seluruh variabel bebas $X$ memiliki korelasi dengan variabel respon $Y$.

## 4. Pembagian Data

Pada penelitian ini data penelitian yang telah melalui *preprocessing* dibagi secara acak ke dalam tiga proporsi data latih yang berbeda yaitu 60%, 70%, dan 80% sehingga membentuk tiga data baru yaitu data A, B, dan C. Adapun jumlah masing-masing data yang digunakan dituliskan ke dalam Tabel 4 berikut.

Tabel 4. Pembagian data penelitian

| Data | Jumlah data latih | Jumlah data uji | Total |
|---|---|---|---|
| A | 1122 | 748 | 1870 |
| B | 1309 | 561 | 1870 |
| C | 1496 | 374 | 1870 |

## 5. Hasil Klasifikasi

Klasifikasi dilakukan pada data uji A, B, dan C menggunakan model regresi logistik dan *random forest* yang telah dibentuk dari data latih A, B, dan C.

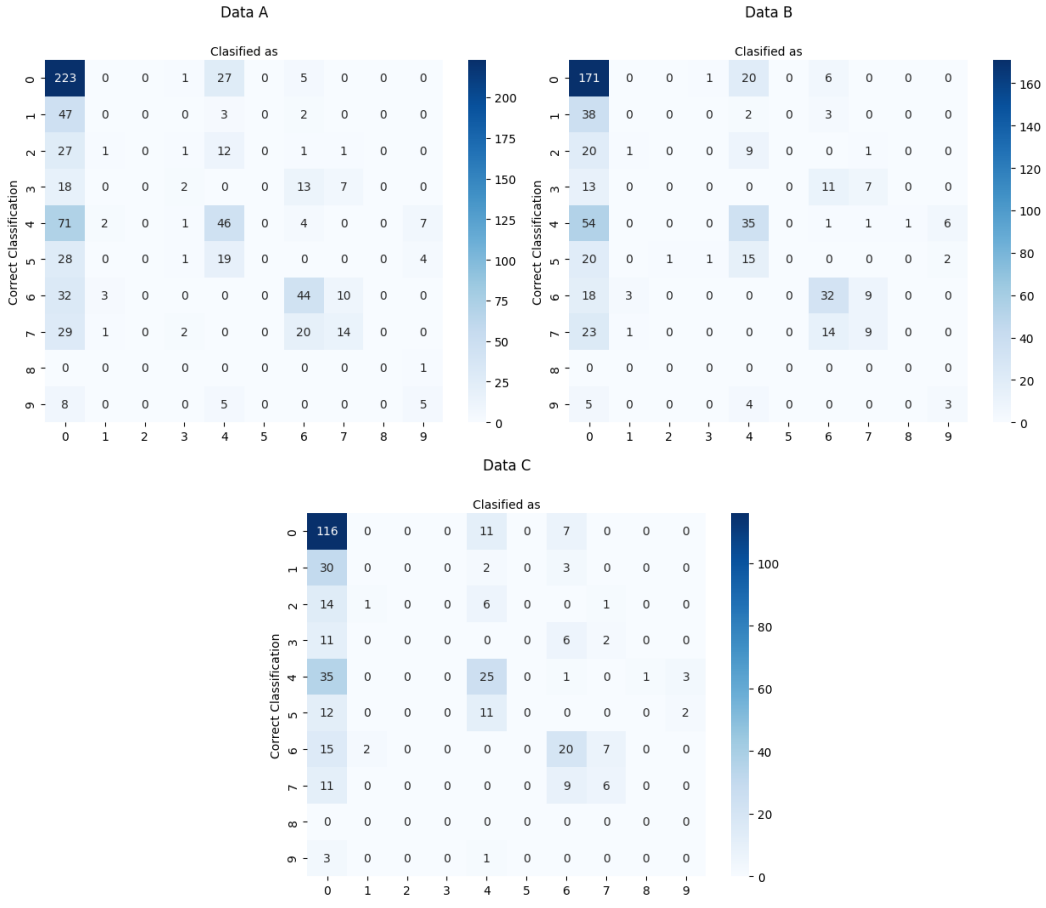
a. Hasil klasifikasi regresi logistik

Estimasi parameter regresi logistik dari data A, B, dan C dilakukan dengan menggunakan bantuan *software* Ananconda Python. Dari hasil estimasi parameter tersebut, didapatkan model regresi logistik yang digunakan untuk mengklasifikasi data uji. Hasil klasifikasi data uji A, B, dan C menggunakan model regresi logistik disajikan ke dalam bentuk *confusion matrix* multi kelas pada Gambar 3.

Selanjutnya, dari *confusion matrix* hasil klasifikasi akan dihitung nilai dari *true positive rate* (TPR) dan *false positive rate* (FPR). Perhitungan TPR dan FPR dilakukan dengan menggunakan perhitungan rata-rata mikro. Nilai rata-rata mikro dari TPR dan FPR pada data latih A, B, dan C disajikan pada Tabel 5 berikut.

Tabel 5. Rata-rata mikro klasifikasi dengan model regresi logistik

| Model | Data | $TPR_{micro}$ | $FPR_{micro}$ |
|---|---|---|---|
| Regresi logistik | A | 0,4465 | 0,0615 |
| | B | 0,4456 | 0,0616 |
| | C | 0,4465 | 0,0615 |




Gambar 3. *Confusion matrix* klasifikasi regresi logistik pada data A, B, dan C

b. Hasil klasifikasi *random forest*

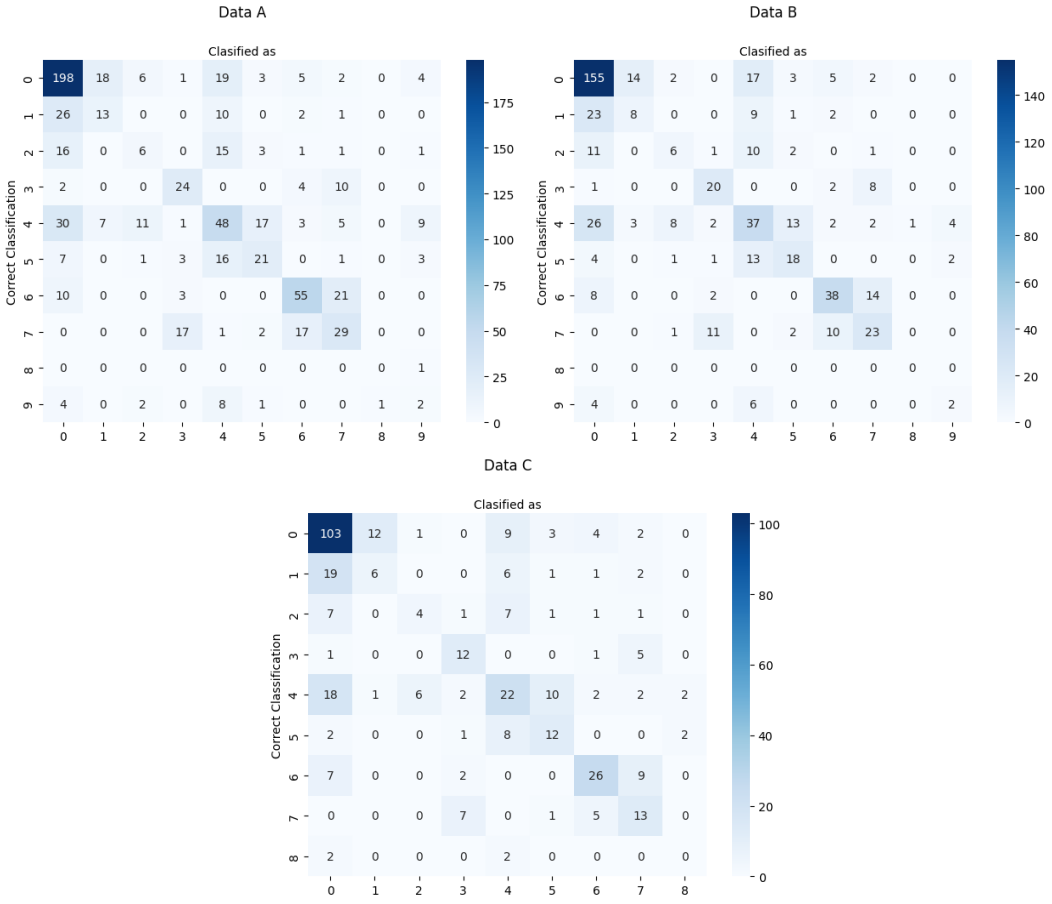
Pembentukan dan klasifikasi *random forest* dilakukan menggunakan *software* Anaconda Python. Jumlah pohon keputusan yang digunakan

untuk klasifikasi yaitu sebanyak 50 pohon. Adapun untuk hasil klasifikasi disajikan ke dalam bentuk *confusion matrix* multi kelas pada Gambar 4.

Selanjutnya, nilai rata-rata mikro dari TPR dan FPR untuk klasifikasi dengan *random forest* diperoleh sebagai berikut.

Tabel 6. Rata-rata mikro klasifikasi dengan *random forest*

| Model | Data | $TPR_{micro}$ | $FPR_{micro}$ |
|---|---|---|---|
| *Random forest* | A | 0,5294 | 0,0523 |
| | B | 0,5472 | 0,0503 |
| | C | 0,5294 | 0,523 |




Gambar 4. *Confusion matrix* klasifikasi *random forest* pada data A, B, dan C

## 6. Evaluasi Hasil Klasifikasi

Setelah dilakukan klasifikasi, model akan dievaluasi menggunakan ukuran nilai AUC. Nilai AUC diperoleh dengan menghitung luas daerah di bawah kurva ROC yang merupakan hasil plot antara nilai TPR dan FPR yang telah diperoleh pada bagian 5.

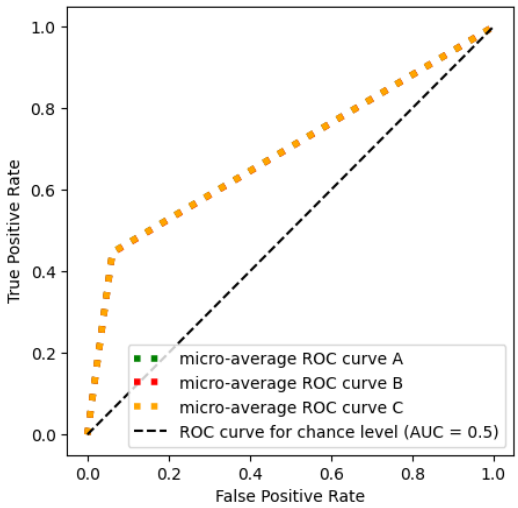
a. Regresi logistik

Kuva ROC dari klasifikasi regresi logistik disajikan pada Gambar 5. Nilai AUC dari kurva ROC pada Gambar 5 disajikan pada Tabel 7.

Tabel 7. Nilai AUC klasifikasi regresi logistik

| Model | Data | AUC |
|---|---|---|
| Regresi logistik | A | 0,6925 |
| | B | 0,6920 |
| | C | 0,6925 |
| | Rata-rata | 0,6923 |

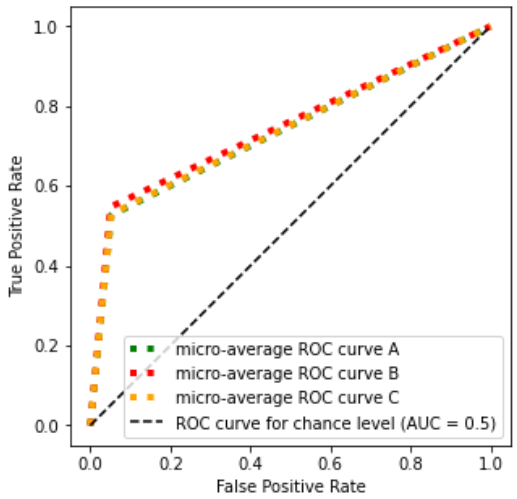
Dari Tabel 7, diketahui bahwa rata-rata nilai AUC dari klasifikasi regresi logistik yaitu 0,6923. Berdasarkan rata-rata nilai AUC yang diperoleh, model klasifikasi regresi logistik tersebut dikategorikan sebagai klasifikasi yang buruk.




Gambar 5. Kurva ROC regresi logistik

b. *Random forest*

Kuva ROC dari klasifikasi *random forest* disajikan pada Gambar 6. Selanjutnya nilai AUC dari kurva ROC klasifikasi *random forest* tersebut disajikan pada Tabel 8.




Gambar 6. Kurva ROC *random forest*

Tabel 8. Nilai AUC klasifikasi *random forest*

| Model | Data | AUC |
|---|---|---|
| *Random forest* | A | 0,7386 |
| | B | 0,7485 |
| | C | 0,7386 |
| | Rata-rata | 0,7419 |

Dari Tabel 8 diketahui bahwa rata-rata nilai AUC dari klasifikasi *random forest* yaitu 0,7419, sehingga klasifikasi *random forest* tersebut termasuk ke dalam model klasifikasi yang cukup baik.

## D. Simpulan

Klasifikasi cuaca di wilayah Jawa Tengah dengan metode regresi logistik dan *random forest* memperoleh rata-rata nilai AUC berturut-turut yaitu 0,6923 dan 0,7419. Berdasarkan nilai tersebut, diperoleh bahwa metode *random forest* lebih baik dari regresi logistik dalam mengklasifikasi cuaca di wilayah Jawa Tengah karena memiliki nilai AUC yang lebih tinggi.

Penelitian selanjutnya disarankan untuk menambahkan variabel lain yang mungkin berpengaruh terhadap cuaca, serta menggunakan data dengan interval waktu 1-3 tahun. Peneliti juga menyarankan untuk melakukan peninjauan kembali terhadap teknik pengolahan data lainnya agar dapat meningkatkan ketepatan hasil klasifikasi.

## E. Daftar Pustaka

Alasadi, S. A., dan Bhaya, W. S. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102–4107.

Badan Meteorologi, Klimatologi, dan Geofisika. *Data Prakiraan Cuaca Terbuka BMKG (Cuaca Jalur Transportasi Darat)*. http://diseminasi.meteo.bmkg.go.id/posko-jalur-darat, diakses pada 16 Oktober 2022.

Badan Meteorologi, Klimatologi, dan Geofisika. (2020). *Pengertian Cuaca Menurut BMKG*. https://maritim.kalbar.bmkg.go.id/konten/pengertian-cuaca/, diakses pada 15 Januari 2023.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Damuri, A., Riyanto, U., Rusdianto, H., dan Aminudin, M. (2021). Implementasi *Data Mining* dengan Algoritma *Naïve Bayes* untuk Klasifikasi Kelayakan Penerima Bantuan Sembako. *Jurnal Riset Komputer*, 8(6), 219–225.

Fallo, S. I. (2021). *Support Vector Machine, Naive Bayes Classifier,* dan Regresi Logistik Ordinal dalam Prediksi Cuaca. Universitas Gadjah Mada.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques* (Vol. 12). Springer Science dan Business Media.

Heydarian, M., Doyle, T. E., dan Samavi, R. (2022). MLCM: Multi-label Confusion Matrix. *IEEE Access*, 10, 19083–19095.

Hosmer Jr, D. W., Lemeshow, S., dan Sturdivant, R. X. (2013). *Applied Logistic Regression*. 398. John Wiley dan Sons.

Kwak, S. K., dan Kim, J. H. (2017). Statistical Data Preparation: Management of Missing Values and Outliers. *Korean Journal of Anesthesiology*, 70(4), 407–411.

Purwa, T. (2019). Perbandingan Metode Regresi Logistik dan *Random Forest* untuk Klasifikasi *Data Imbalanced* (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika dan Komputasi*, 16(1), 58–73.

Senthilnathan, S. (2019). Usefulness of Correlation Analysis. *Available at SSRN 3416918*.

Siregar, A. M. (2020). Klasifikasi untuk Prediksi Cuaca Menggunakan *Esemble Learning*.

Tanujaya, L. B. C., Susanto, B., dan Saragih, A. (2020). The Comparison of Logistic Regression Methods and Random Forest for Spotify Audio Mode Featurre Classification. *Indonesian Journal of Data and Science*, 1(3), 68–78.