

**PERBANDINGAN PEFORMA REGRESI LINIER, REGRESI
LOGISTIK, DECESION TREE, DAN RANDOM FOREST
DALAM KLASIFIKASI CUACA KOTA SEMARANG**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

AGUS NURWAHID

21.11.3988

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

**PERBANDINGAN PEFORMA REGRESI LINIER, REGRESI
LOGISTIK, DECESION TREE, DAN RANDOM FOREST
DALAM KLASIFIKASI CUACA KOTA SEMARANG**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh
AGUS NURWAHID
21.11.3988

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2025**

HALAMAN PERSETUJUAN

SKRIPSI

**PERBANDINGAN PEFORMA REGRESI LINIER, REGRESI LOGISTIK,
DECESION TREE, DAN RANDOM FOREST DALAM KLASIFIKASI
CUACA KOTA SEMARANG**

yang disusun dan diajukan oleh

Agus Nurwahid

21.11.3988

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 23 Juni 2025

Dosen Pembimbing,



Arif Akbarul Huda, S.Si., M.Eng.
NIK. 190302287

HALAMAN PENGESAHAN

SKRIPSI

**PERBANDINGAN PEFORMA REGRESI LINIER, REGRESI LOGISTIK,
DECESION TREE, DAN RANDOM FOREST DALAM KLASIFIKASI
CUACA KOTA SEMARANG**

yang disusun dan diajukan oleh

Agus Nurwahid

21.11.3988

Telah dipertahankan di depan Dewan Penguji
pada tanggal 23 Juni 2025

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Andika Agus Slameto, S.Kom., M.Kom.
NIK. 190302109

Ferian Fauzi Abdulloh, S.Kom., M.Kom.
NIK. 190302276

Arif Akbarul Huda, S.Si., M.Eng.
NIK. 190302287



Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 23 Juni 2025

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Agus Nurwahid
NIM : 21.11.3988

Menyatakan bahwa Skripsi dengan judul berikut:

Perbandingan Performa Regresi Linier, Regresi Logistik, Decision Tree, dan Random Forest dalam Klasifikasi Cuaca Kota Semarang

Dosen Pembimbing : Arif Akbarul Huda, S.Si., M.Eng.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 23 Juni 2025

Yang Menyatakan,



METERA
TEMPER
211EAMX232292307

Agus Nurwahid

HALAMAN PERSEMBAHAN

Puji syukur saya panjatkan kepada Allah SWT atas rahmat dan karunia-Nya, sehingga saya dapat menyelesaikan skripsi ini, yang berjudul " Perbandingan Performa Regresi Linier, Regresi Logistik, Decision Tree, dan Random Forest dalam Klasifikasi Cuaca Kota Semarang," dengan penuh rasa hormat dan terimakasih, skripsi saya persembahkan kepada:

1. Ayah dan Ibu tercinta, yang bernama Bapak Tukijo dan Ibu Murnilah selaku donatur tetap dalam segala aspek baik dana ataupun tenaga, yang selalu memberikan cinta dan doa karenanya saya bisa singgah dititik ini.
2. Dosen Pembimbing saya, Bapak Arif Akbarul Huda, S.Si., M.Eng. , yang telah memberikan bimbingan, saran dan dukungan yang sangat berarti sehingga skripsi ini dapat terselenggara dengan lancar. Terimakasih atas kesabaran dan dedikasi Anda.
3. Keluarga besar saya, yang selalu memberikan semangat dan cinta, Terimakasih dukungan moral dan emosional yang kalian berikan.
4. Teman-teman seperjuangan, kepada Adi, Sultan, Farhan, Galih, Hafid , Ichsan dan teman-teman IF3, yang selalu menemani dalam saya berproses baik suka maupun duka. Terimakasih atas kebersamaan dan dukungan kalian selama proses menuntut ilmu ini.
5. Almamater tercinta, Universitas Amikom Yogyakarta, ynaq telah memberikan saya fasilitas dan peluang untuk saya belajar dan berkembang.
6. Tuhan Yang Maha Esa, atas berkat dan rahmatnya yang selalu tercurah dalam setiap langkah saya menuntut ilmu.

KATA PENGANTAR

Puji syukur selalu terpanjatkan atas kehadiran Tuhan Yang Maha Esa, karena berkat rahmat dan karunia-Nya, saya dapat menyelesaikan skripsi ini dengan judul "Perbandingan Peforma Regresi Linier, Regresi Logistik, Decesion Tree, dan Random Forest dalam Klasifikasi Cuaca Kota Semarang." Disusunnya skripsi ini adalah sebagai bentuk hasil proses saya menimba ilmu dan sebagai syarat sah untuk memperoleh gelar Sarjana di Prodi Informatika Universitas Amikom Yogyakarta.

Penelitian ini bertujuan untuk membangingkan peforma dari bebagai metode analisis data dalam memetakan klasifikasi cuaca, yang nantinya digunakan sebagai perencanaan dan pengambilan keputusan dari bebagai sektor diantaranya sektor pertanian, perikanan, transportasi dan berbagai aspek yang berpengaruh penting dengan cuaca. Dalam proses penyusunan skripsi ini melibatkan bebeerbagai pihak yang memberikan dukungan, bantuan, dan bimbingan yang sangat berarti.

Tidak lupa ucapan terimakasih saya haturkan kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas Amikom Yogyakarta, yang telah memberikan kesempatan dan dukungan dalam menyelesaikan pendidikan saya.
2. Bapak Arif Akbarul Huda, S.Si., M.Eng., selaku dosen pembimbing, yang selalu memberikan bantuan berupa bimbingan, arahan dan motivasi sebagai pilar penting terselasaikannya penelitian dan penyusunan skripsi ini.
3. Ibu Murnilah, selaku ibu saya tercinta, yang telah melahirkan dan membesarkan saya dengan segenap hati.
4. Bapak Tukijo, selaku bapak saya, yang telah memberikan dukungan baik saran ataupun dana.
5. Saudara dekat saya yang telah membantu saya dalam berbagai aspek secara tidak langsung.

6. Sultan Faaiz Fadlurrahman Jaghadita, M Sachib Farhan Nauvaldi, Adi Dwi Ribowo dan Galih Dwi Prayogo, selaku teman dekat saya selama berkuliah dan teman-teman IF3, yang turut membantu dalam suka maupun duka selama saya menuntut ilmu di Amikom tercinta ini.

Saya juga menyampaikan terima kasih yang sebesar-besarnya kepada seluruh pihak yang telah memberikan dukungan, baik secara langsung maupun tidak langsung, dalam penyusunan dan penyelesaian skripsi ini. Meskipun tidak dapat saya sebutkan satu per satu, setiap bantuan, doa, dan dorongan yang diberikan sangat berarti dan menjadi penyemangat selama proses penelitian ini. Semoga segala bentuk kebaikan yang telah diberikan mendapat balasan yang setimpal dari Tuhan Yang Maha Esa. Saya juga berharap, skripsi ini dapat memberikan manfaat, baik dalam meningkatkan pemahaman mengenai klasifikasi cuaca, maupun sebagai kontribusi ilmiah yang dapat mendukung pengembangan ilmu pengetahuan dan teknologi di masa yang akan datang.

Yogyakarta, 1 Mei 2025

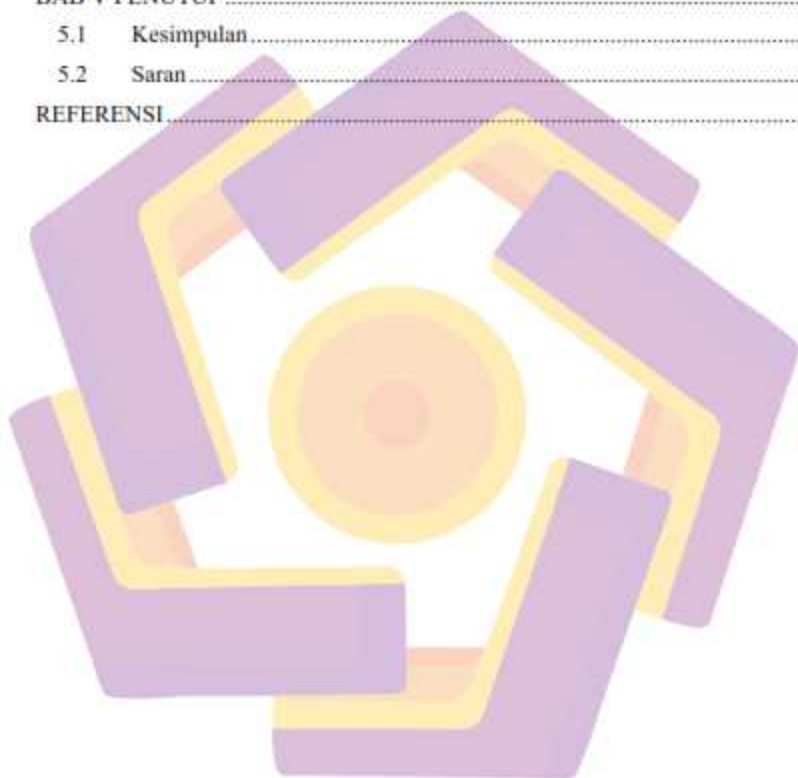
Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiv
DAFTAR LAMBANG DAN SINGKATAN	xv
DAFTAR ISTILAH	xvii
INTISARI	xxi
ABSTRACT	xxii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6
2.2 Dasar Teori	19
2.2.1 Cuaca	19
2.2.2 Machine Learning	19
2.2.3 PySpark	19
2.2.4 Klasifikasi dan Regresi	20
2.2.5 Preprocessing Data	23
2.2.6 Evaluasi Model	25
2.2.7 Hyperparameter Tuning	28

2.2.8	Deployment	28
BAB III METODE PENELITIAN		30
3.1	Objek Penelitian	30
3.2	Alur Penelitian	32
3.2.1	Perencanaan dan Inisiasi Penelitian	33
3.2.2	Identifikasi Masalah	33
3.2.3	Perumusan Masalah dan Tujuan Penelitian	33
3.2.4	Studi Literatur	33
3.2.5	Pengumpulan Data Cuaca	33
3.2.6	Proses Implementasi dan Pengkodean Program	33
3.2.7	Analisis Perbandingan Hasil Model	34
3.2.8	Penulisan Naskah	34
3.2.9	Kesimpulan dan Saran	34
3.3	Alur Proses	35
3.3.1	Data Acquisition (Pengumpulan Data)	35
3.3.2	Exploratory Data Analysis (EDA)	36
3.3.3	Data Preprocessing (Pra-pemrosesan Data)	36
3.3.4	Data Splitting	36
3.3.5	Model Selection and Training	36
3.3.6	Model Evaluation	37
3.3.7	Hyperparameter Tuning	38
3.3.8	Model Comparison and Selection	38
3.3.9	Final Model Deployment / Prediction	39
3.4	Alat dan Bahan	40
3.4.1	Data Penelitian	40
3.4.2	Alat/Instrumen Penelitian	41
BAB IV HASIL DAN PEMBAHASAN		43
4.1	Persiapan Awal	43
4.2	Pengambilan Data	44
4.3	Preprocessing Data	45
4.4	Balancing Data	48
4.5	Visualisasi Korelasi	51

4.6	Persiapan Fitur	52
4.7	Splitting Data	55
4.8	Model Klasifikasi dan Evaluasi	56
4.9	Perbandingan Kinerja Model	71
4.10	Deployment	72
BAB V PENUTUP		74
5.1	Kesimpulan	74
5.2	Saran	75
REFERENSI		77



DAFTAR TABEL

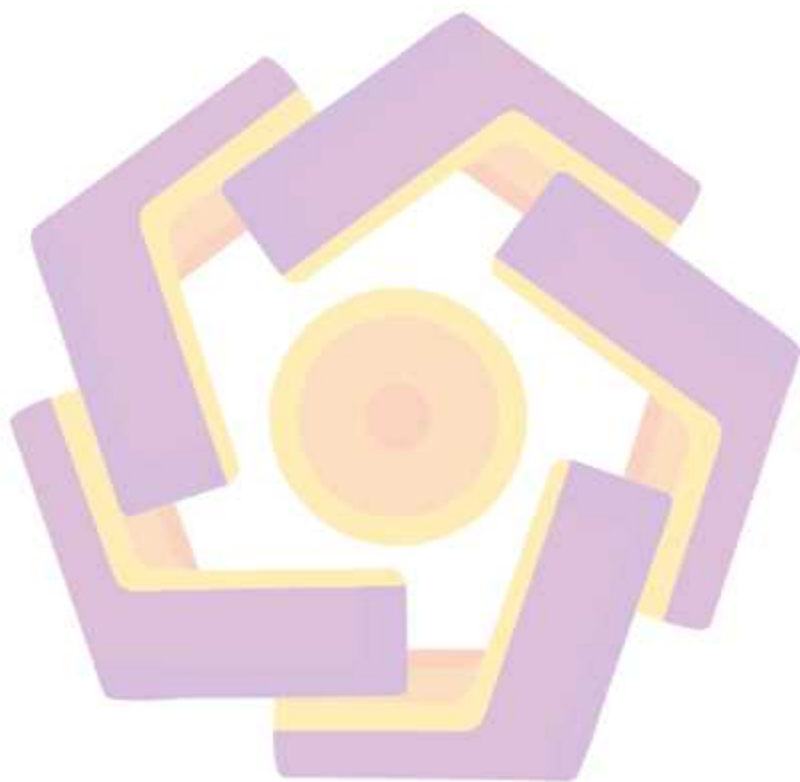
Tabel 2. 1 Keaslian Penelitian	1
Tabel 3. 1 Alat/Instrumen Penelitian	1
Tabel 4. 1 Output DataFrame	1
Tabel 4. 2 Jumlah Sampel Perkelas	1
Tabel 4. 3 Jumlah Setelah Balancing	1
Tabel 4. 4 AUC-ROC Linear Regression	1
Tabel 4. 5 Linear Regression Metrics	1
Tabel 4. 6 AUC ROC Logistic Regretions	1
Tabel 4. 7 Logistic Regretions Metrics	1
Tabel 4. 8 AUC ROC Decision Tree	1
Tabel 4. 9 Decision Tree Metrics	1
Tabel 4. 10 AUC ROC Random Forest	1
Tabel 4. 11 Random Forest Metrics	1
Tabel 4. 12 Perbandingan Model	1

DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian	1
Gambar 3. 2 Alur Proses	1
Gambar 3. 3 Wireframe	1
Gambar 4. 1 Instal dan Import Library	43
Gambar 4. 2 Pengambilan Data	44
Gambar 4. 3 Perubahan Nama Kolom	45
Gambar 4. 4 Hasil Perubahan Nama Kolom	46
Gambar 4. 5 Label Hujan	46
Gambar 4. 6 Hasil Labelisasi Hujan	47
Gambar 4. 7 Perhitungan Missing Value	47
Gambar 4. 8 Hasil Perhitungan Missing Value	47
Gambar 4. 9 Perhitungan Sampel	48
Gambar 4. 10 Visualisasi Distribusi Kelas	49
Gambar 4. 11 Undersampling	49
Gambar 4. 12 Visualisasi Setelah Balancing	51
Gambar 4. 13 Visualisasi Korelasi	52
Gambar 4. 14 Identifikasi Kolom	53
Gambar 4. 15 String Indexing	53
Gambar 4. 16 VectorAssembler	53
Gambar 4. 17 Pipeline	54

Gambar 4. 18 DataFrame Akhir	55
Gambar 4. 19 Splitting Data	55
Gambar 4. 20 Grafik ROC Linear Regression	57
Gambar 4. 21 Evaluasi BinaryClassificationEvaluator	58
Gambar 4. 22 Perbandingan Metrik Linear Regression	59
Gambar 4. 23 Grafik ROC Logistic Regretions	61
Gambar 4. 24 AUC ROC Logistic Regretions	63
Gambar 4. 25 Grafik ROC Decision Tree	64
Gambar 4. 26 Perbandingan Metrik Decision Tree	66
Gambar 4. 27 Grafik ROC Random Forest	67
Gambar 4. 28 Perbandingan Metrik Random Forest	70
Gambar 4. 30 Antamuka Prediksi Cuaca	73

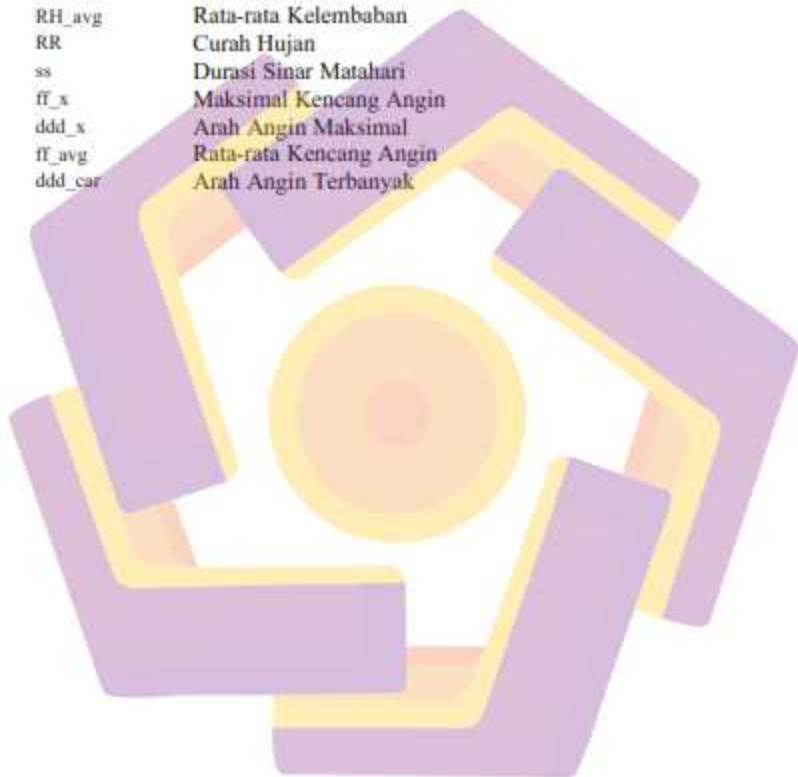
DAFTAR LAMPIRAN



DAFTAR LAMBANG DAN SINGKATAN

$\hat{y}$	Nilai prediksi
β_0	Intersep (bias)
$\beta_1, \beta_2, \dots, \beta_n$	Koefisien regresi
$x_1, x_2, \dots, x_n$	Variabel fitur
MSE	Mean Squared Error
y_i	Nilai aktual
$\hat{y}_i$	Nilai prediksi
n	Jumlah data
$P(y = 1 x)$	Probabilitas bahwa label adalah 1 (misal: hujan)
p_i	Proporsi sampel pada kelas ke-i
c	Jumlah kelas
$h_T(x)$	Prediksi dari pohon ke-t terhadap data input x
T	Jumlah total pohon dalam Random Forest
mode	Nilai yang sering muncul dari hasil prediksi
SMOTE	Synthetic Minority Over-sampling Technique
x_i	Vektor fitur sampel minoritas yang ada.
x_{zi}	Vektor fitur dari k-nearest dari x_i dalam kelas minoritas.
δ	Nilai acak antara 0 dan 1
x_{new}	Sampel data sintetis baru yang dihasilkan.
ML	Machine Learning
AUC-ROC	Area Under the Curve – Receiver Operating Characteristic
EDA	Exploratory Data Analysis
CSV	Comma-Separated Values
MLlib	Machine Learning Library (Apache Spark)
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate

FPR	False Positive Rate
°C	Celcius
%	Persen
m/s	Meter per detik
BMKG	Badan Meteorologi, Klimatologi, dan Geofisika
Tn	Minimal Suhu
Tx	Maksimal Suhu
Tavg	Rata-rata Suhu
RH_avg	Rata-rata Kelembaban
RR	Curah Hujan
ss	Durasi Sinar Matahari
ff_x	Maksimal Kencang Angin
ddd_x	Arah Angin Maksimal
ff_avg	Rata-rata Kencang Angin
ddd_car	Arah Angin Terbanyak



DAFTAR ISTILAH

Akurasi	Persentase prediksi yang benar dari seluruh prediksi yang dilakukan.
Algoritma Machine Learning	Metode matematis untuk membuat model komputer belajar dari data dan memprediksi hasil.
Apache Spark	Framework/platform komputasi terdistribusi untuk pemrosesan data besar secara cepat.
Artificial Neural Network (ANN)	Model komputasi yang meniru cara kerja otak manusia.
AUC-ROC	Metrik untuk mengukur kemampuan model membedakan kelas positif dan negatif.
Balancing Data	Proses menyeimbangkan jumlah data antara kelas mayoritas dan minoritas.
Balanced_df	Data frame yang sudah diseimbangkan untuk menghindari bias kelas dalam pemodelan.
BinaryClassificationEvaluator	Evaluator untuk menilai performa model klasifikasi biner seperti AUC-ROC.
BMKG	Badan Meteorologi, Klimatologi, dan Geofisika — penyedia data cuaca di Indonesia.
C4.5	Algoritma pohon keputusan yang mengembangkan ID3 dengan pemangkasan dan fitur tambahan.
CategoricalColumns	Kolom data yang berisi kategori atau kelas, bukan nilai numerik.
Confusion Matrix	Matriks evaluasi performa model klasifikasi dengan menghitung TP, TN, FP, dan FN.
CrossValidator	Metode validasi silang untuk memilih model terbaik dengan membagi data menjadi beberapa fold.
CSV	Format file data tabular yang dipisahkan dengan koma.
Data Balancing	Proses menyeimbangkan jumlah data antar kelas untuk mengatasi ketidakseimbangan.
Data Frame	Struktur data tabel yang digunakan di Spark untuk menyimpan dan mengelola data.
Data Historis	Data masa lalu yang digunakan sebagai dasar analisis atau pelatihan model.
Data Mining	Proses menemukan pola dan informasi berguna dari kumpulan data besar.
Data Preprocessing	Tahap pengolahan data sebelum pelatihan model, seperti pembersihan dan normalisasi.
Decision Tree	Algoritma klasifikasi berbasis struktur pohon

Decision Tree Classifier	Model klasifikasi yang menggunakan pohon keputusan berbasis aturan if-then.
Durasi Penyinaran Surya	Lama waktu penyinaran matahari dalam satu hari, diukur dalam jam.
ElasticNetParam	Parameter yang mengatur kombinasi regularisasi L1 dan L2 pada model regresi.
Encoding	Mengubah data menjadi format yang dapat diproses oleh algoritma.
Exploratory Data Analysis (EDA)	Analisis awal data untuk memahami pola dan hubungan antar fitur menggunakan statistik dan visualisasi.
Feature (Fitur)	Variabel input dalam dataset yang digunakan dalam proses pelatihan model.
Feature Engineering	Membuat atau memilih fitur baru dari data untuk meningkatkan performa model.
FeaturesCol	Kolom yang berisi vektor fitur numerik sebagai input model machine learning.
F1 Score	Rata-rata harmonis dari precision dan recall, berguna pada data yang tidak seimbang.
Gaussian Naïve Bayes	Varian Naïve Bayes untuk data kontinu dengan distribusi normal.
Gradio	Library Python untuk membuat antarmuka pengguna (UI) dari model machine learning.
Hyperparameter Tuning	Proses mencari parameter terbaik untuk meningkatkan performa model ML.
InferSchema	Opsi otomatis mendeteksi tipe data saat membaca file CSV di Spark.
Kaggle	Platform penyedia dataset dan kompetisi data science.
Kecepatan Angin Maksimum	Nilai tertinggi kecepatan angin dalam satu hari, diukur dalam m/s.
Kelembaban Rata-rata	Rata-rata kelembaban udara dalam satu hari dalam persen.
Klasifikasi Cuaca	Proses mengelompokkan kondisi cuaca ke dalam kategori, seperti hujan atau tidak hujan.
Koefisien Determinasi	Nilai yang menunjukkan seberapa baik variabel bebas menjelaskan variabel terikat.
Label	Target output (kelas) dalam supervised learning. Dalam hal ini: hujan atau tidak hujan.
Label Hujan	Target klasifikasi, 0 – tidak hujan, 1 – hujan.
LabelCol	Kolom target yang berisi label kelas untuk klasifikasi.
Levenberg-Marquardt	Metode optimisasi yang sering digunakan dalam pelatihan jaringan saraf.

Logistic Regression	Model statistik untuk memprediksi probabilitas kelas biner.
Machine Learning	Metode komputasi untuk membuat model prediktif dari data.
Max Depth	Parameter yang mengatur kedalaman maksimum pohon keputusan.
Max Iter	Jumlah maksimum iterasi yang dilakukan selama pelatihan model.
Max Suhu	Suhu maksimum dalam satu hari (°C).
Min Instances Per Node	Jumlah minimum data yang harus ada pada setiap node dalam pohon keputusan.
Min Suhu	Suhu minimum dalam satu hari (°C).
Model Deployment	Tahap penerapan model ke lingkungan nyata atau produksi untuk melakukan prediksi.
MulticlassClassificationEvaluator	Evaluator untuk mengukur performa model klasifikasi dengan beberapa kelas.
OneHotEncoder	Metode mengubah data kategorikal menjadi representasi numerik biner.
ParamGridBuilder	Fitur untuk membuat kombinasi parameter hyperparameter tuning.
Pipeline	Rangkaian tahapan transformasi data dan pelatihan model secara berurutan.
Precision	Ukuran seberapa banyak prediksi positif yang benar dari semua yang diprediksi positif.
Predict	Proses menghasilkan prediksi model terhadap data baru.
Recall	Ukuran seberapa banyak prediksi positif yang benar dari semua data aktual positif.
Random Forest	Teknik ensemble berbasis kumpulan pohon keputusan untuk meningkatkan akurasi.
Random Forest Classifier	Model ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi.
Random Forest Classifier	Algoritma klasifikasi yang menggunakan banyak pohon keputusan.
RandomSplit	Metode membagi data secara acak menjadi bagian pelatihan dan pengujian.
RandomizedSearchCV	Metode pencarian parameter optimal dengan pengambilan nilai secara acak.
Recall	Proporsi data positif yang berhasil dikenali oleh model.
RegParam	Parameter regulasi untuk mencegah overfitting pada model.
Regresi Linear Berganda	Model regresi dengan lebih dari satu variabel bebas.
Regresi Linier	Algoritma prediktif untuk hubungan linear

Regresi Logistik	antar variabel. Digunakan sebagai baseline.
ROC Curve	Algoritma klasifikasi biner yang memodelkan probabilitas suatu kejadian.
Scaling/Normalization	Kurva yang menunjukkan hubungan True Positive Rate dan False Positive Rate.
Spark MLlib	Mengubah skala fitur numerik agar nilainya seragam dan tidak bias.
SparkSession	Library machine learning pada Apache Spark untuk pemodelan skala besar.
StringIndexer	Objek untuk menjalankan perintah dan operasi Spark SQL.
Studi Literatur	Transformasi mengubah nilai kategori menjadi indeks numerik.
Supervised Learning	Penelaahan terhadap sumber-sumber ilmiah sebelumnya untuk mendukung penelitian.
SVMS	Metode machine learning yang menggunakan data berlabel untuk pelatihan model.
Timestamp	Support Vector Machines, algoritma klasifikasi dan regresi (tidak ada dalam daftar awal).
Train Set	Format waktu pencatatan data berupa tanggal dan jam.
Training Set	Data yang digunakan untuk melatih model.
Test Set	Data yang digunakan untuk melatih model.
Testing Set	Data yang digunakan untuk menguji performa model setelah pelatihan.
TP (True Positive)	Data yang digunakan untuk menguji performa model.
TPR (True Positive Rate)	Jumlah prediksi benar untuk kelas positif.
Undersampling	Rasio data positif yang diprediksi benar.
Vector Assembler	Teknik untuk menangani data tidak seimbang dengan mengurangi sampel dari kelas mayoritas.
VectorAssembler	Fitur dalam PySpark untuk menggabungkan berbagai kolom menjadi satu vektor fitur.
	Transformasi yang menggabungkan kolom fitur menjadi satu kolom vektor fitur.

INTISARI

Klasifikasi cuaca merupakan aspek penting dalam pengambilan keputusan di berbagai sektor seperti pertanian, transportasi, dan mitigasi bencana. Penelitian ini membandingkan performa empat metode machine learning, yaitu Regresi Linier, Regresi Logistik, Decision Tree, dan Random Forest, dalam mengklasifikasikan kondisi cuaca di Kota Semarang. Dataset yang digunakan mencakup variabel meteorologi seperti tanggal pengamatan, suhu minimum (min_suhu), suhu maksimum (max_suhu), rata-rata suhu (rata2_suhu), rata-rata kelembaban udara (rata2_kelembaban), durasi penyinaran matahari (dur_surya), kecepatan angin maksimum (max_kecangin), arah angin maksimum (arah_angin_max), rata-rata kecepatan angin (rata2_kecangin), serta label hujan sebagai variabel target klasifikasi. Data diproses dan dianalisis menggunakan PySpark dengan pendekatan supervised learning. Evaluasi performa model dilakukan menggunakan metrik AUC-ROC untuk menilai ketepatan dan kemampuan masing-masing model dalam membedakan kelas cuaca, serta akurasi, presisi, recall, dan f1-score sebagai evaluasi lanjutan. Hasil penelitian menunjukkan bahwa metode Random Forest memberikan performa terbaik, dengan nilai AUC, akurasi, presisi, recall, dan f1-score tertinggi dibandingkan dengan Regresi Linier, Regresi Logistik, dan Decision Tree. Temuan ini menegaskan bahwa teknik ensemble seperti Random Forest lebih efektif dalam menangani kompleksitas data meteorologi dan dapat menjadi solusi andal untuk klasifikasi cuaca di wilayah perkotaan seperti Semarang. Penelitian ini diharapkan dapat memberikan kontribusi bagi pengembangan sistem prediksi cuaca yang lebih akurat di Indonesia.

Kata kunci: Klasifikasi cuaca, Regresi Linier, Regresi Logistik, Decision Tree, Random Forest, PySpark, data meteorologi, Kota Semarang.

ABSTRACT

Weather classification is a crucial aspect in decision-making across various sectors such as agriculture, transportation, and disaster mitigation. This study compares the performance of four machine learning methods—Linear Regression, Logistic Regression, Decision Tree, and Random Forest—in classifying weather conditions in Semarang City. The dataset includes meteorological variables such as observation date, minimum temperature (min_suhu), maximum temperature (max_suhu), average temperature (rata2_suhu), average humidity (rata2_kelembaban), duration of sunlight (dur_surya), maximum wind speed (max_kecangin), maximum wind direction (arah_angin_max), and average wind speed (rata2_kecangin), with rainfall serving as the classification target. The data was processed and analyzed using PySpark with a supervised learning approach. Model performance was evaluated using the AUC-ROC metric to assess the accuracy and ability of each model to distinguish weather classes, along with accuracy, precision, recall, and F1-score as additional evaluation metrics. The results showed that the Random Forest method outperformed the others, achieving the highest AUC, accuracy, precision, recall, and F1-score. These findings confirm that ensemble techniques like Random Forest are more effective in handling the complexity of meteorological data and can serve as a reliable solution for weather classification in urban areas such as Semarang. This study is expected to contribute to the development of more accurate weather prediction systems in Indonesia.

Keywords: *weather classification, Linear Regression, Logistic Regression, Decision Tree, Random Forest, PySpark, meteorological data, Semarang City.*