

Dian Ayu Afifah, S.Si., M.Sc. | Agiska Ria Supriyatna, S.Si., M.T.I
Ir. Nurul Qomariyah S.Kom., M.Kom. | Ir. Eko Subyantoro, S.Kom., M.Kom.
Ir. Imam Asrowadi, S.Kom., M.Kom. | Dr. Ir. Septafiansyah Dwi Putra, S.T., M.T.
Ir. Zuriati, S.Kom., M.Kom. | Ismail Puji Saputra, S.Kom., M.Kom.

Strategi Cerdas

Pengolahan Data untuk Analitik Prediktif dan Segmentasi



Strategi Cerdas

Pengolahan Data untuk Analitik Prediktif dan Segmentasi

Dian Ayu Afifah, S.Si., M.Sc. | Agiska Ria Supriyatna, S.Si., M.T.I
Ir. Nurul Qomariyah S.Kom., M.Kom. | Ir. Eko Subyantoro, S.Kom., M.Kom.
Ir. Imam Asrowardi, S.Kom., M.Kom. | Dr. Ir. Septafiansyah Dwi Putra, S.T., M.T.
Ir. Zuriati, S.Kom., M.Kom. | Ismail Puji Saputra, S.Kom., M.Kom.



Strategi Cerdas Pengolahan Data untuk Analitik Prediktif dan Segmentasi

Ditulis oleh:

Dian Ayu Afifah, S.Si., M.Sc.
Agiska Ria Supriyatna, S.Si., M.T.I.
Ir. Nurul Qomariyah, S.Kom., M.Kom.
Ir. Eko Subyantoro, S.Kom., M.Kom.
Ir. Imam Asrowardi, S.Kom., M.Kom.
Dr. Ir. Septafiansyah Dwi Putra, S.T., M.T.
Ir. Zuriati, S.Kom., M.Kom.
Ismail Puji Saputra, S.Kom., M.Kom.

Diterbitkan, dicetak, dan didistribusikan oleh
PT Literasi Nusantara Abadi Grup
Perumahan Puncak Joyo Agung Residence Blok B11 Merjosari
Kecamatan Lowokwaru Kota Malang 65144
Telp : +6285887254603, +6285841411519
Email: literasinusantaraofficial@gmail.com
Web: www.penerbitlitnus.co.id
Anggota IKAPI No. 340/JTI/2022



Hak Cipta dilindungi oleh undang-undang. Dilarang mengutip
atau memperbanyak baik sebagian ataupun keseluruhan isi buku
dengan cara apa pun tanpa izin tertulis dari penerbit.

Cetakan I, Desember 2025

Perancang sampul: D Gea Nuansa

Penata letak: D Gea Nuansa

ISBN : 978-634-234-935-9

x + 206 hlm. ; 15,5x23 cm.

©Desember 2025

Kata Pengantar

Puji syukur kehadiran Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan buku yang berjudul “Strategi Cerdas Pengolahan Data untuk Analitik Prediktif dan Segmentasi.”

Buku ini disusun sebagai upaya untuk memberikan pemahaman yang komprehensif mengenai bagaimana data dapat diolah secara efektif untuk menghasilkan wawasan yang bernilai, khususnya dalam konteks analitik prediktif dan segmentasi. Di era digital saat ini, data bukan sekadar kumpulan angka, melainkan aset strategis yang mampu menjadi dasar pengambilan keputusan cerdas di berbagai bidang, mulai dari bisnis, pemerintahan, hingga riset akademik.

Melalui buku ini, penulis berusaha menyajikan konsep, metode, serta strategi praktis yang dapat diterapkan untuk mengolah data secara efisien—mulai dari tahap pengumpulan, pembersihan, transformasi, hingga penerapan model analitik yang relevan. Pembahasan juga dilengkapi dengan contoh kasus dan pendekatan berbasis teknologi terkini agar pembaca dapat memahami penerapannya dalam dunia nyata.

Penulis menyadari sepenuhnya bahwa buku ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi penyempurnaan karya ini di masa mendatang.

Akhir kata, penulis berharap buku ini dapat memberikan manfaat nyata bagi para pembaca—baik mahasiswa, peneliti, praktisi data, maupun siapa pun yang ingin memahami dan mengoptimalkan potensi data untuk mendukung pengambilan keputusan yang lebih cerdas.

Semoga buku ini menjadi kontribusi kecil dalam memperkaya literatur ilmu data dan menjadi inspirasi bagi pengembangan strategi analitik yang inovatif di masa depan.

Politeknik Negeri Lampung
Jurusan Teknologi Informasi
Program Studi Sarjana Terapan Teknologi Rekayasa Internet

Daftar Isi

Kata Pengantar iii

Daftar Isi v

Daftar Gambar viii

CHAPTER 1

Pengantar Analitik prediktif dan segmentasi1

References:.....5

CHAPTER 2

Prediksi Kualitas Udara Menggunakan Model
Deep Learning LSTM7

 A. Ringkasan7

 B. Pendahuluan8

 C. Studi Literatur 11

 D. Metodologi penelitian.....16

 E. Eksperimen dan evaluasi.....18

 F. Diskusi23

 G. Kesimpulan.....24

References:.....25

CHAPTER 3

Prediksi Harga dan Segmentasi Mobil Bekas
Menggunakan Regresi Dan Clustering
Machine Learning Berbasis Fitur Kendaraan..... 27

 A. Ringkasan27

 B. Pendahuluan28

 C. Studi Literatur33

D. Eksperimen dan evaluasi.....	44
E. Diskusi	54
F. Kesimpulan.....	57
References:.....	60

CHAPTER 4

Analisis Perbandingan Metode Clustering

untuk Segmentasi Daerah Berdasarkan Data COVID-19 ... 63

A. Ringkasan	63
B. Pendahuluan	64
C. Studi Literatur	70
D. Metodologi Penelitian.....	73
E. Eksperimen dan Evaluasi	78
F. Diskusi	87
G. Kesimpulan.....	92
References:.....	94

CHAPTER 5

Optimasi Klasifikasi Dini Penyakit Stroke

Menggunakan *Stacking Ensemble* Berbasis

Random Forest, XGBoost, dan LightGBM

dengan Logistic Regression sebagai Meta-Classifler 97

A. Ringkasan	97
B. Pendahuluan	98
C. Studi Literatur	100
D. Metodologi Penelitian.....	103
E. Eksperimen dan Evaluasi	108
F. Diskusi	113
G. Kesimpulan.....	116
References:.....	118

CHAPTER 6

Prediksi Customer Churn Telekomunikasi
Menggunakan Advanced Feature Engineering,
Data Balancing SMOTE, dan Stacking Multi-
Ensemble Model..... 121

A. Ringkasan 121

B. Pendahuluan 122

C. Studi Literatur 125

D. Metodologi Penelitian..... 128

E. Eksperimen dan Evaluasi 131

F. Diskusi 143

G. Kesimpulan..... 147

References:..... 150

CHAPTER 7

Analisis Komparatif Model Deteksi Anomali
Berbasis AutoEncoder, IF, LOF, dan OCSVM
pada Trafik Jaringan IoT 153

A. Ringkasan 153

B. Pendahuluan 155

C. Studi Literatur 158

D. Metode Penelitian..... 163

E. Hasil Eksperimen dan Evaluasi..... 183

F. Kesimpulan dan Rekomendasi..... 201

References:..... 203

Daftar Gambar

Gambar 1. Perbandingan Nilai Prediksi dan Nilai Aktual.....	19
Gambar 2. Visualisasi Data pada 200 Sampel Pertama	20
Gambar 3. Analisis Nilai Residual Berdasarkan Jam dalam Sehari.....	20
Gambar 4. Perbandingan Nilai <i>RMSE</i> dan <i>MAE</i>	21
Gambar 5. Perbandingan Nilai R^2 Score	22
Gambar 1. Susunan Metode Regresi dan <i>Clustering</i> untuk Analisis Mobil Bekas	39
Gambar 2. Visualisasi Heatmap Korelasi Pearson Antar Variabel Numerik	45
Gambar 3. Visualisasi Elbow Method.....	47
Gambar 4. Visualisasi Silhouette Score.....	49
Gambar 5. <i>Visualisasi Principal Component Analysis (PCA)</i>	50
Gambar 1. Diagram Alur Penelitian	73
Gambar 2. Distribusi Provinsi Berdasarkan Kategori Risiko Kasus Aktif COVID-19 di Indonesia	81
Gambar 3. Tren kasus aktif COVID-19.....	82
Gambar 4. Tren kasus aktif COVID-19.....	83
Gambar 5. Tren kasus aktif COVID-19.....	83
Gambar 6. Segmentasi Perprovinsi	85
Gambar 7. Segmentasi Perprovinsi	85
Gambar 8. Segemntasi Provinsi.....	86
Gambar 1. <i>heatmap confusion matrix default threshold.</i>	109
Gambar 2. Grafik <i>Threshold vs Precision & Recall</i>	110
Gambar 3. Grafik <i>Top 10 Feature Importance.</i>	111
Gambar 1. Confusion Matrix Random Forest.....	132
Gambar 2. ROC Curve Model Random Forest	133
Gambar 3. Confusion Matrix Random Forest + XGBoost	134
Gambar 4. ROC Curve Random Forest + XGBoost.....	134
Gambar 5. Confusion Matrix Random Forest + LightGBM	135
Gambar 6. ROC Curve Random Forest + LightGBM	135
Gambar 7. Confusion Matrix Random Forest + XGBoost + LightGBM.....	136

Gambar 8. ROC Curve Random Forest + XGBoost + LightGBM..... 136

Gambar 9. Confusion MatrixRandom Forest + CatBoost 137

Gambar 10. ROC CurveRandom Forest + CatBoost..... 137

Gambar 11. Confusion MatrixRandom Forest + XGBoost +
LightGBM + CatBoost 138

Gambar 12. ROC CurveRandom Forest + XGBoost +
LightGBM + CatBoost 138

Gambar 12. Perbandingan Distribusi Kelas Sebelum dan
Sesudah SMOTE..... 139

Gambar 13. Confusion MatrixRandom Forest + GradientBoosting..... 139

Gambar 14. ROC CurveRandom Forest + GradientBoosting 140

Gambar 15. ROC Curve Model Final Voting Ensemble 140

Gambar 16. Distribusi Kelas Sebelum dan Sesudah SMOTEENN..... 141

Gambar 2. Formula F1-Score..... 179

Gambar 3. Pipeline Penelitian 182

Gambar 4. Distribusi Autoencoder..... 185

Gambar 5. Kurva Precision Recall Autoencoder..... 186

Gambar 6. Confusion Matrix Autoencoder..... 188

Gambar 7. Distribusi Isolation Forest..... 189

Gambar 8. DistribusiLocal Outlier Factor(LOF) 191

Gambar 9. Confusion Matrix Local Outlier Factor 193

Gambar 10. DistribusiOne-Class Support Vector Machine (OCSVM). 194

Gambar 11. Confusion Matrix One-Class SVM 195

Gambar 12. Visualisasi Perbandingan Kinerja Antar Model 197

Gambar 13. Radar Perbandingankinerja..... 198

Gambar 14. VisualisasiPerbandingan F1-Score 199



CHAPTER 1

Pengantar Analitik prediktif dan segmentasi

Analitik prediktif merupakan cabang dari analisis data yang berfokus pada penggunaan data historis, statistik, dan algoritma pembelajaran mesin untuk memperkirakan kejadian atau perilaku di masa depan. Pendekatan ini menjadi inti dari pengambilan keputusan berbasis data karena mampu mengubah pola historis menjadi wawasan prediktif yang membantu organisasi bertindak proaktif, bukan reaktif (IBM, 2024). Melalui teknik seperti regresi, pohon keputusan, random forest, dan jaringan saraf tiruan, sistem analitik prediktif dapat mempelajari hubungan kompleks antarvariabel untuk menghasilkan perkiraan dengan tingkat akurasi tinggi (Kuhn & Johnson, 2013).

Dalam praktiknya, analitik prediktif telah diadopsi secara luas di berbagai sektor. Di bidang bisnis dan keuangan, analitik ini digunakan untuk menilai risiko kredit, mendeteksi penipuan, serta mengoptimalkan strategi pemasaran dan penjualan (Provost & Fawcett, 2013). Di sektor kesehatan, model prediktif membantu memperkirakan risiko penyakit, mendukung diagnosis dini, serta meningkatkan efektivitas intervensi klinis

(Rajkomar et al., 2019). Dalam industri manufaktur dan Internet of Things (IoT), analitik prediktif berperan dalam perawatan prediktif (predictive maintenance), meminimalkan downtime peralatan, dan mengoptimalkan rantai pasok (Lee et al., 2020). Bahkan pada bidang lingkungan, prediksi berbasis data digunakan untuk memantau polusi udara, memperkirakan curah hujan, serta mengantisipasi perubahan iklim secara lebih akurat (Wang et al., 2023).

Sementara itu, segmentasi data—sering disebut juga segmentasi pasar atau segmentasi perilaku—adalah proses mengelompokkan entitas seperti pelanggan, wilayah, atau produk ke dalam kelompok yang memiliki kesamaan karakteristik. Tujuannya adalah agar strategi yang diterapkan pada tiap kelompok menjadi lebih terarah dan efisien. Pendekatan ini berakar dari analisis kluster dalam statistik dan machine learning, seperti metode K-Means, Hierarchical Clustering, atau Gaussian Mixture Model (Jain et al., 1999). Segmentasi banyak digunakan dalam pemasaran untuk mengenali pola perilaku konsumen, dalam pendidikan untuk memahami profil belajar mahasiswa, dan dalam kesehatan untuk mengidentifikasi kelompok risiko yang membutuhkan intervensi berbeda (Wedel & Kamakura, 2000).

Analitik prediktif dan segmentasi memiliki hubungan yang erat dan saling melengkapi. Segmentasi dapat menjadi langkah awal dalam memahami heterogenitas populasi sebelum membangun model prediktif—strategi yang dikenal dengan pendekatan *segment-then-predict*. Dengan mengelompokkan data terlebih dahulu, model dapat dilatih secara spesifik pada setiap segmen, sehingga menghasilkan prediksi yang lebih akurat dan relevan. Sebaliknya, hasil dari model prediktif, seperti skor risiko atau probabilitas kejadian, juga dapat digunakan untuk membentuk segmen baru yang lebih bermakna secara bisnis atau operasional (*predict-to-segment*). Pendekatan integratif ini telah terbukti efektif dalam meningkatkan personalisasi layanan dan efisiensi strategi berbasis data (Fayyad et al., 1996; Wirth & Hipp, 2000).



CHAPTER 2

Prediksi Kualitas Udara Menggunakan Model *Deep Learning* LSTM

Dian Ayu Afifah, Imam Asrowardi, Meilina,
Muhammad Hafizh Risqullah

A. Ringkasan

Penelitian ini membahas pengembangan model prediksi kualitas udara berbasis *deep learning* menggunakan arsitektur *Long Short-Term Memory* (*LSTM*). Latar belakang penelitian berangkat dari meningkatnya polusi udara, khususnya partikel halus PM2.5, yang berdampak besar terhadap kesehatan manusia dan lingkungan. Sistem pemantauan konvensional dinilai belum mampu memberikan prediksi *real-time* yang akurat dan efisien, sehingga dibutuhkan pendekatan berbasis kecerdasan buatan yang dapat memanfaatkan data historis untuk menghasilkan prediksi yang lebih presisi.

Metodologi penelitian menggunakan dataset PRSA dari UCI yang mencatat kualitas udara Beijing tahun 2010–2014, dengan 13 fitur utama meliputi parameter meteorologi dan konsentrasi PM2.5. Data diolah melalui tahap *preprocessing* seperti *one-hot encoding*, normalisasi, dan pembentukan jendela waktu (*sliding window*) sebelum dimasukkan ke dalam model LSTM. Model dievaluasi menggunakan metrik MAE, RMSE, dan R^2 , dan dibandingkan dengan model CNN serta CNN-LSTM. Hasilnya menunjukkan bahwa LSTM murni memberikan performa terbaik dengan nilai RMSE dan MAE terendah serta R^2 mencapai 0.957, menandakan kemampuan tinggi dalam menangkap pola temporal pada data polusi udara.

Penelitian ini menyimpulkan bahwa model LSTM sangat efektif dalam memprediksi konsentrasi PM2.5 secara akurat dan stabil, sehingga berpotensi digunakan dalam sistem pemantauan kualitas udara berbasis AI secara *real-time*. Rekomendasi pengembangan selanjutnya mencakup integrasi data eksternal seperti kondisi cuaca dan aktivitas lalu lintas, serta eksplorasi model lanjutan seperti *Attention-LSTM* atau Transformer untuk peningkatan akurasi dan efisiensi prediksi.

B. Pendahuluan

Latar Belakang

Kemajuan teknologi informasi (TI) dalam beberapa dekade terakhir telah mengubah industri secara signifikan melalui teknologi digital. Di Indonesia, sektor TI berkembang pesat dengan meningkatnya kebutuhan digitalisasi di berbagai sektor, termasuk bisnis, pemerintahan, dan pendidikan. Transformasi ini difokuskan pada pengembangan infrastruktur, pengelolaan data, dan keamanan informasi melalui teknologi seperti komputasi awan, big data, dan keamanan siber. Namun, di balik pertumbuhan ini, tantangan besar muncul akibat kesenjangan antara keterampilan dan tanggung jawab profesional TI. Hal ini terutama terlihat ketika mereka menangani proyek di luar kompetensinya [1].



CHAPTER 3

Prediksi Harga dan Segmentasi Mobil Bekas Menggunakan Regresi Dan Clustering Machine Learning Berbasis Fitur Kendaraan

Nurul Qomariyah, Eko Subyantoro, Amanda Bulan
Nayla, Gilang Rizki Ramadhan, Syahreza Riatma

A. Ringkasan

Penelitian ini membahas pengembangan model prediksi harga dan segmentasi mobil bekas menggunakan pendekatan *machine learning* berbasis fitur kendaraan. Latar belakang penelitian berangkat dari pesatnya pertumbuhan pasar mobil bekas, yang menuntut metode analisis cerdas untuk menentukan harga dan segmentasi pasar secara akurat. Untuk menjawab tantangan tersebut, penelitian ini memanfaatkan dua pendekatan utama: **regresi** untuk memprediksi harga jual kendaraan berdasarkan fitur seperti tahun pembuatan, jarak tempuh, dan jenis bahan

bakar, serta **clustering** untuk mengelompokkan mobil ke dalam segmen pasar dengan karakteristik yang serupa.

Metodologi penelitian menggunakan dataset mobil bekas dari India yang terdiri dari 14.993 entri dan 11 fitur utama. Pra-pemrosesan dilakukan melalui pembersihan data, penanganan nilai kosong, encoding fitur kategorikal, serta normalisasi data numerik. Proses segmentasi dilakukan dengan algoritma K-Means, dan hasil evaluasi menunjukkan dua kluster optimal dengan nilai *Silhouette Score* tertinggi (0,717), yang membedakan segmen mobil harga menengah dan premium. Selanjutnya, model Random Forest Regressor digunakan untuk prediksi harga, menghasilkan performa tinggi dengan nilai $R^2 = 0,9477$, MAE = 92.439, dan RMSE = 146.838, menunjukkan tingkat akurasi yang sangat baik.

Hasil penelitian menunjukkan bahwa kombinasi metode regresi dan klusterisasi secara signifikan meningkatkan keandalan prediksi harga mobil bekas. Segmentasi pasar berbasis data juga memberikan wawasan strategis bagi pelaku industri otomotif dalam menentukan strategi harga dan pemasaran yang lebih terarah. Penelitian ini merekomendasikan pengembangan lebih lanjut dengan integrasi algoritma *ensemble* dan klusterisasi adaptif seperti DBSCAN atau Gaussian Mixture Model, serta penerapan sistem prediktif berbasis web agar dapat dimanfaatkan secara langsung oleh dealer, platform jual beli, dan konsumen.

B. Pendahuluan

Latar Belakang

Dalam beberapa tahun terakhir, pasar mobil bekas telah menunjukkan pertumbuhan yang signifikan, terutama di negara-negara berkembang. Perkembangan ini mendorong pentingnya pengembangan metode yang efektif untuk memprediksi harga serta mengelompokkan segmen pasar berdasarkan fitur kendaraan. Penerapan teknik *machine learning*, termasuk regresi dan *clustering*, telah menjadi fokus penelitian yang intensif karena



CHAPTER 4

Analisis Perbandingan Metode Clustering untuk Segmentasi Daerah Berdasarkan Data COVID-19

Agiska Ria Supriyatna, Zuriati, Dheril Pratama
Prasetya, Fadil Risa Oktariano, Riky Alvian

A. Ringkasan

Penelitian ini menganalisis segmentasi wilayah di Indonesia berdasarkan pola penyebaran COVID-19 menggunakan berbagai metode *clustering*, yaitu K-Means, DBSCAN, Hierarchical Clustering, dan DBA-KMeans. Tujuannya adalah mengelompokkan provinsi-provinsi yang memiliki kesamaan tren penyebaran kasus aktif harian untuk mendukung pengambilan keputusan kebijakan publik berbasis data. Data yang digunakan berupa deret waktu kasus COVID-19 harian dari Maret 2020 hingga Maret 2022 untuk 34 provinsi di Indonesia. Melalui pendekatan *Dynamic Time Warping Barycenter Averaging K-Means* (DBA-KMeans),

penelitian ini berhasil menangkap variasi temporal antar wilayah yang tidak selalu sinkron secara waktu namun serupa dalam pola penyebarannya.

Hasil eksperimen menunjukkan bahwa metode DBA-KMeans mampu membentuk tiga klaster utama: risiko rendah, sedang, dan tinggi. Provinsi seperti DKI Jakarta, Jawa Barat, dan Jawa Timur termasuk dalam klaster risiko tinggi dengan pola lonjakan tajam pada gelombang varian Delta dan Omicron, sedangkan wilayah seperti Gorontalo dan Maluku Utara tergolong risiko rendah dengan tren kasus yang stabil. Evaluasi kuantitatif memperlihatkan bahwa DBA-KMeans unggul dalam menangani data deret waktu dengan nilai *Inertia* terendah dan *Silhouette Score* tinggi (0,668), menunjukkan kualitas klasterisasi yang baik. Visualisasi hasil klaster juga memperlihatkan perbedaan pola penyebaran antar kelompok, sekaligus mempermudah interpretasi spasial-temporal pandemi di Indonesia.

Secara keseluruhan, penelitian ini menegaskan efektivitas DBA-KMeans berbasis DTW sebagai metode klasterisasi spasial-temporal dalam menganalisis pandemi. Pendekatan ini tidak hanya deskriptif tetapi juga aplikatif, karena dapat membantu pemerintah dalam merancang kebijakan mitigasi yang lebih kontekstual berdasarkan pola risiko wilayah. Pengembangan lebih lanjut disarankan dengan menambahkan variabel seperti tingkat vaksinasi, mobilitas sosial, dan kapasitas layanan kesehatan, serta menerapkan model prediktif berbasis *deep learning* atau metode *shape-based clustering* untuk mendukung sistem peringatan dini yang lebih adaptif dan presisi.

B. Pendahuluan

Latar Belakang

Pandemi COVID-19 menjadi peristiwa global yang mengguncang tatanan sosial, ekonomi, dan kesehatan dalam skala yang belum pernah terjadi sebelumnya pada abad ke-21. Sejak kemunculan kasus pertama di Wuhan pada akhir 2019, virus ini dengan cepat menyebar ke seluruh penjuru dunia, menyebabkan krisis multidimensi yang melampaui sekadar isu



CHAPTER 5

Optimasi Klasifikasi Dini Penyakit Stroke Menggunakan *Stacking Ensemble* Berbasis Random Forest, XGBoost, dan LightGBM dengan Logistic Regression sebagai Meta-Classifier

Nurul Qomariyah, Ismail Puji Saputra, Eja Prastio,
Farhan Febri Aryaputra

A. Ringkasan

Penelitian ini berfokus pada optimasi klasifikasi dini penyakit stroke menggunakan pendekatan stacking ensemble yang menggabungkan tiga model pembelajaran mesin — Random Forest, XGBoost, dan LightGBM — dengan Logistic Regression sebagai meta-classifier. Latar belakang penelitian ini berangkat dari urgensi deteksi dini stroke sebagai penyakit dengan angka kematian tinggi di dunia, di mana model tunggal sering kali gagal mendeteksi kasus minoritas akibat ketidakseimbangan data

medis. Dengan menggunakan dataset publik Stroke Prediction Dataset yang berisi lebih dari 40 ribu data pasien, penelitian ini menerapkan pipeline preprocessing yang meliputi imputasi data hilang, normalisasi, dan encoding fitur kategorikal agar model dapat belajar secara optimal tanpa oversampling.

Hasil eksperimen menunjukkan bahwa model stacking ensemble memberikan performa terbaik dengan akurasi 92%, ROC AUC 0,972, dan F1-score rata-rata 0,91, mengungguli model tunggal. Model mampu mendeteksi sebagian besar kasus stroke dengan recall 0,92 pada kelas positif (stroke) dan 0,91 pada kelas negatif (non-stroke), menunjukkan keseimbangan antara sensitivitas dan ketepatan prediksi. Analisis feature importance mengidentifikasi lima faktor dominan terhadap risiko stroke, yaitu kadar glukosa darah, hipertensi, penyakit jantung, indeks massa tubuh, dan status pernikahan.

Penelitian ini membuktikan bahwa pendekatan stacking ensemble efektif dalam meningkatkan akurasi dan stabilitas model prediksi medis tanpa harus melakukan penyeimbangan data buatan. Temuan ini berpotensi diterapkan dalam sistem deteksi dini berbasis AI untuk membantu tenaga medis melakukan skrining pasien berisiko tinggi secara lebih cepat dan akurat. Rekomendasi ke depan mencakup validasi eksternal dengan data rumah sakit nyata, penerapan teknik interpretabilitas seperti SHAP/LIME, serta integrasi model ke dalam dashboard klinis real-time guna mempercepat pengambilan keputusan medis.

B. Pendahuluan

Latar Belakang

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence / AI) telah membawa transformasi signifikan dalam bidang kesehatan, termasuk dalam diagnosis penyakit kronis seperti stroke. Stroke merupakan salah satu penyebab kematian utama di dunia, termasuk di Indonesia, dengan angka kematian mencapai lebih dari 6 juta kasus



CHAPTER 6

Prediksi Customer Churn Telekomunikasi Menggunakan Advanced Feature Engineering, Data Balancing SMOTE, dan Stacking Multi-Ensemble Model

Agiska Ria Supriyatna, Irfan Syah Utama,
Muhammad Abiyyu Utomo, Muhammad Hibban
Anvirhesa

A. Ringkasan

Penelitian ini mengembangkan sistem prediksi customer churn pada industri telekomunikasi menggunakan kombinasi teknik *advanced feature engineering*, penyeimbangan data (*SMOTE* dan *SMOTEENN*), serta pendekatan *stacking multi-ensemble model*. Tujuan utamanya adalah untuk meningkatkan akurasi prediksi pelanggan yang berpotensi berhenti berlangganan dengan memanfaatkan kekuatan gabungan beberapa algoritma *machine learning* seperti Random Forest,

XGBoost, LightGBM, CatBoost, dan Gradient Boosting. Dataset yang digunakan adalah IBM Telco Customer Churn Dataset dengan tingkat churn 26,5%. Melalui tahapan *preprocessing*, *feature engineering*, dan balancing data, model dibangun untuk menangani tantangan utama seperti ketidakseimbangan kelas, keterbatasan representasi fitur, dan overfitting pada model tunggal.

Hasil eksperimen menunjukkan bahwa model stacking multi-ensemble secara konsisten melampaui performa baseline Random Forest, meningkatkan akurasi dari 84,49% menjadi 86,62% dan nilai ROC AUC dari 0,92 menjadi 0,95. Kombinasi model terbaik (RF + CatBoost) menghasilkan keseimbangan optimal antara *precision* dan *recall*, dengan penurunan signifikan pada *false positive* dan *false negative*. Selain itu, penggunaan *Voting Ensemble* berbasis stacking meningkatkan performa secara drastis dengan akurasi 97,05% dan ROC AUC 0,9944, menjadikannya konfigurasi paling stabil dan presisi. Penelitian ini menegaskan bahwa integrasi *ensemble learning* dan *feature engineering* mampu menciptakan model prediksi churn yang akurat, interpretatif, dan adaptif — memberikan nilai strategis tinggi bagi perusahaan dalam merancang kebijakan retensi pelanggan yang berbasis data.

B. Pendahuluan

Latar Belakang

Dalam era digital saat ini, industri telekomunikasi menjadi salah satu sektor yang sangat kompetitif dan cepat berubah. Penyedia layanan seluler dan internet berlomba untuk mempertahankan pelanggan mereka di tengah persaingan tarif, kualitas layanan, dan inovasi produk. Salah satu tantangan utama yang dihadapi perusahaan telekomunikasi adalah customer churn, yaitu fenomena ketika pelanggan menghentikan layanannya dan berpindah ke penyedia lain.

Menjaga pelanggan yang ada memiliki biaya yang secara signifikan lebih rendah dibandingkan dengan memperoleh pelanggan baru. Menurut



CHAPTER 7

Analisis Komparatif Model Deteksi Anomali Berbasis AutoEncoder, IF, LOF, dan OCSVM pada Trafik Jaringan IoT

Dian Ayu Afifah, Septafiansyah Dwi Putra, Rahmat Hadinata, Hafish Arrusal Isfalana, Adam Fikri Nawibowo

A. Ringkasan

Penelitian ini menganalisis perbandingan empat model deteksi anomali berbasis *unsupervised learning* Autoencoder, Isolation Forest (IF), Local Outlier Factor (LOF), dan One-Class Support Vector Machine (OCSVM) dalam mengidentifikasi aktivitas abnormal pada trafik jaringan *Internet of Things* (IoT) menggunakan dataset terbaru CIC IoT-IDAD 2024. Latar belakang penelitian berangkat dari meningkatnya ancaman keamanan IoT, seperti serangan *botnet Mirai*, yang menuntut sistem deteksi anomali adaptif tanpa ketergantungan pada data berlabel. Dengan pendekatan

novelty detection, model dilatih menggunakan data trafik normal (*benign*) untuk mengenali pola perilaku sah, kemudian diuji terhadap data anomali yang dihasilkan oleh serangan *Mirai*. Pra-pemrosesan mencakup pembersihan data, imputasi median, normalisasi *MinMaxScaler*, serta pembagian data secara strategis agar model hanya belajar dari data *benign* selama pelatihan.

Keempat algoritma diuji secara empiris menggunakan metrik kinerja utama seperti *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *ROC-AUC*. Hasil eksperimen menunjukkan bahwa *Local Outlier Factor (LOF)* memberikan performa terbaik dengan *Accuracy* 0.6999, *Precision* 0.9217, *Recall* 0.5404, *F1-Score* 0.6813, dan *ROC-AUC* 0.7798, menandakan keseimbangan tinggi antara kemampuan mendeteksi serangan dan menghindari alarm palsu. *Autoencoder* menempati posisi kedua dengan *precision* baik (0.8553) namun *recall* rendah (0.3024), menunjukkan bahwa model ini akurat tetapi sering melewatkan anomali. Sementara *Isolation Forest* dan *One-Class SVM* menunjukkan performa yang jauh lebih lemah, masing-masing memiliki nilai *F1* di bawah 0.1 dan *recall* sangat rendah (<0.05), menandakan keterbatasan mereka dalam mengenali pola serangan kompleks pada trafik IoT.

Secara keseluruhan, penelitian ini menegaskan bahwa *LOF* merupakan metode paling efektif untuk deteksi anomali pada jaringan IoT berbasis *data flow*, karena kemampuannya mengadaptasi variasi lokal dan menangani ketidakseimbangan data dengan baik. Sementara *Autoencoder* menunjukkan potensi untuk dikembangkan lebih lanjut dengan optimasi threshold dan tuning arsitektur, metode *IF* dan *OCSVM* dinilai kurang cocok untuk skenario IoT tanpa pelabelan eksplisit. Temuan ini memberikan landasan empiris bagi pengembangan sistem deteksi anomali *IoT* yang lebih adaptif, akurat, dan efisien dalam menghadapi ancaman siber yang terus berevolusi.

References:

- Agrawal, S., & Agrawal, J. (2015). Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science*, 60, 708–713. <https://doi.org/10.1016/j.procs.2015.08.220>
- Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., & Guizani, M. (2020). A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security. *IEEE Communications Surveys & Tutorials*, 22(3), 1646–1685. <https://doi.org/10.1109/COMST.2020.2988293>
- Anam, K. (2024). Early Stopping on CNN-LSTM Development to Improve Classification Performance. *Journal of Applied Data Sciences*, 5(3), 1175–1188. <https://doi.org/10.47738/jads.v5i3.312>
- Ayad, A. G., El-Gayar, M. M., Hikal, N. A., & Sakr, N. A. (2024). Efficient Real-Time Anomaly Detection in IoT Networks Using One-Class Autoencoder and Deep Neural Network. *Electronics*, 14(1), 104. <https://doi.org/10.3390/electronics14010104>
- Canadian Institute for Cybersecurity. (2024). *IoT-DIAD 2024 UNB*. <https://www.unb.ca/cic/datasets/iot-diad-2024.html>
- Ceselli, A. (2023). *Information Management course*.
- Churcher, A., Ullah, R., Ahmad, J., Ur Rehman, S., Masood, F., Gogate, M., Alqahtani, F., Nour, B., & Buchanan, W. J. (2021a). An Experimental Analysis of Attack Classification Using Machine Learning in IoT Networks. *Sensors*, 21(2), 446. <https://doi.org/10.3390/s21020446>
- Churcher, A., Ullah, R., Ahmad, J., Ur Rehman, S., Masood, F., Gogate, M., Alqahtani, F., Nour, B., & Buchanan, W. J. (2021b). An Experimental Analysis of Attack Classification Using Machine Learning in IoT Networks. *Sensors*, 21(2), 446. <https://doi.org/10.3390/s21020446>
- CISA. (2017). <https://www.cisa.gov/news-events/alerts/2016/10/14/heightened-ddos-threat-posed-mirai-and-other-botnets>

- Dee, R., &Zincir-Heywood, N. (2025). *Preliminary Study of Learning Models to Understand the Effect of Unlabelled Data for IoT Attack Detection*.
- Gao, X., Xie, D., Zhang, Y., Wang, Z., He, C., Yin, H., & Zhang, W. (2025). *A Comprehensive Survey on Imbalanced Data Learning* (No. arXiv:2502.08960). arXiv. <https://doi.org/10.48550/arXiv.2502.08960>
- Garcia, S., &Parmisano, A. (2020a). *IoT-23 Dataset: A labeled dataset of Malware and Benign IoT Traffic*. Stratosphere IPS. <https://www.stratosphereips.org/datasets-iot23>
- Garcia, S., &Parmisano, A. (2020b). *IoT-23 Dataset: A labeled dataset of Malware and Benign IoT Traffic*. Stratosphere IPS. <https://www.stratosphereips.org/datasets-iot23>
- Goh, K. W., Surono, S., Afiatin, M. Y. F., Mahmudah, K. R., Irsalinda, N., Chaimanee, M., & Onn, C. W. (2024). Comparison of Activation Functions in Convolutional Neural Network for Poisson Noisy Image Classification. *Emerging Science Journal*, 8(2), 592–602. <https://doi.org/10.28991/ESJ-2024-08-02-014>
- Imad, M., Abul Hassan, M., Hussain Bangash, S., &Naimullah. (2022). A Comparative Analysis of Intrusion Detection in IoT Network Using Machine Learning. In M. Ouaisa, Z. Boulouard, M. Ouaisa, I. U. Khan, & M. Kaosar (Eds.), *Big Data Analytics and Computational Intelligence for Cybersecurity* (Vol. 111, pp. 149–163). Springer International Publishing. https://doi.org/10.1007/978-3-031-05752-6_10
- Kerimov, K., Kurbanov, S., & Azizova, Z. (2025). РОБОТИ ТА СИСТЕМИ ШТУЧНОГО ІНТЕЛЕКТУ.
- Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (No. arXiv:1412.6980). arXiv. <https://doi.org/10.48550/arXiv.1412.6980>

- Krzysztoń, E., Rojek, I., & Mikołajewski, D. (2024). A Comparative Analysis of Anomaly Detection Methods in IoT Networks: An Experimental Study. *Applied Sciences*, 14(24), 11545. <https://doi.org/10.3390/app142411545>
- Lim, K., & Prum, S. (2023). *Machine Learning Approaches for Anomaly Detection in IoT Networks: A Survey*.
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. <https://doi.org/10.1109/ICDM.2008.17>
- ma,arif, A. S., -, S. S., -, M. N. A., & -, G. S. (2024). Deep Learning for Anomaly Detection in IoT Systems: Techniques, Applications, and Future Directions. *International Journal For Multidisciplinary Research*, 6(4), 24601. <https://doi.org/10.36948/ijfmr.2024.v06i04.24601>
- Moustafa, N., Turnbull, B., & Choo, K.-K. R. (2019). An Ensemble Intrusion Detection Technique Based on Proposed Statistical Flow Features for Protecting Network Traffic of Internet of Things. *IEEE Internet of Things Journal*, 6(3), 4815–4830. <https://doi.org/10.1109/JIOT.2018.2871719>
- NCCoE. (2025). <https://www.nccoe.nist.gov/iot>
- Neto. (2023). *IoT Dataset 2023 | Datasets | Research | Canadian Institute for Cybersecurity | UNB*. <https://www.unb.ca/cic/datasets/iotdataset-2023.html>
- Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A Survey of Network-based Intrusion Detection Data Sets. *Computers & Security*, 86, 147–167. <https://doi.org/10.1016/j.cose.2019.06.005>
- Russell, P., Elsayed, M. A., Nandy, B., Seddigh, N., & Zincir-Heywood, N. (2023). On the Fence: Anomaly Detection in IoT Networks. *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, 1–4. <https://doi.org/10.1109/NOMS56928.2023.10154271>

- Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11(1), 105. <https://doi.org/10.1186/s40537-024-00957-y>
- Saranya, K., & Valarmathi, A. (2025). A multilayer deep autoencoder approach for cross layer IoT attack detection using deep learning algorithms. *Scientific Reports*, 15(1), 10246. <https://doi.org/10.1038/s41598-025-93473-9>
- Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (n.d.). *Support Vector Method for Novelty Detection*.
- Sharafaldin, I., Canadian Institute for Cybersecurity (CIC), UNB, Fredericton, Canada, Gharib, A., Canadian Institute for Cybersecurity (CIC), UNB, Fredericton, Canada, Lashkari, A. H., Canadian Institute for Cybersecurity (CIC), UNB, Fredericton, Canada, Ghorbani, A. A., & Canadian Institute for Cybersecurity (CIC), UNB, Fredericton, Canada. (2017). Towards a Reliable Intrusion Detection Benchmark Dataset. *Software Networking*, 2017(1), 177–200. <https://doi.org/10.13052/jsn2445-9739.2017.009>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (n.d.). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*.
- Tran, Q.-A., Duan, H., & Li, X. (2020). *One-class Support Vector Machine for Anomaly Network Traffic Detection*.
- Zhukabayeva, T., Adamova, A., Zholshiyeva, L., Mardenov, Y., Karabayev, N., & Baumuratova, D. (2025). Tackling the anomaly detection challenge in large-scale wireless sensor networks. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(2), 2479. <https://doi.org/10.11591/ijece.v15i2.pp2479-2490>



Strategi Cerdas

Pengolahan Data untuk Analitik Prediktif dan Segmentasi

Buku Strategi Cerdas Pengolahan Data untuk Analitik Prediktif dan Segmentasi membahas secara komprehensif bagaimana data diolah dan dimanfaatkan untuk mendukung pengambilan keputusan berbasis kecerdasan buatan. Dimulai dengan pengantar analitik prediktif dan segmentasi, buku ini menjelaskan berbagai teknik seperti regresi, decision tree, deep learning, hingga metode clustering seperti K-Means, Hierarchical, DBSCAN, dan DBA-KMeans yang digunakan untuk memahami pola serta memprediksi kejadian masa depan. Melalui sejumlah studi kasus, pembaca diajak mengeksplorasi penerapan nyata analitik data di berbagai bidang. Penelitian prediksi kualitas udara menggunakan LSTM menunjukkan kemampuan deep learning dalam memproses data deret waktu. Studi mengenai prediksi harga dan segmentasi mobil bekas memanfaatkan Random Forest dan K-Means untuk menghasilkan model penentuan harga yang akurat dan segmentasi pasar yang jelas. Pada konteks kesehatan masyarakat, analisis clustering penyebaran COVID-19 menampilkan bagaimana metode spasial-temporal mampu membagi wilayah Indonesia ke dalam kluster risiko yang berbeda. Bab selanjutnya membahas optimasi klasifikasi dini penyakit stroke dengan pendekatan stacking ensemble, disusul penelitian prediksi churn pelanggan telekomunikasi menggunakan rekayasa fitur lanjutan dan teknik penyeimbangan data.



✉ literasinusantaraofficial@gmail.com
🌐 www.penerbitlitnus.co.id
📖 Literasi Nusantara
📞 literasinusantara_085755971589

Pendidikan

+17

ISBN 978-634-234-935-9



9 786342 349359