SKRIPSI

ANALISIS KINERJA MODEL EXTREME GRADIENT BOOSTING (XGBOOST) DALAM PREDIKSI CURAH HUJAN MENGGUNAKAN TEKNIK RESAMPLING DATA DAN HYPERPARAMETER TUNNING

NAJWA LAILA ANGGRAINI

NPM 21081010191

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S.Si., M.T

Achmad Junaidi, S.Kom., M.Kom

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2025

SKRIPSI

ANALISIS KINERJA MODEL EXTREME GRADIENT BOOSTING (XGBOOST) DALAM PREDIKSI CURAH HUJAN MENGGUNAKAN TEKNIK RESAMPLING DATA DAN HYPERPARAMETER TUNNING

NAJWA LAILA ANGGRAINI

NPM 21081010191

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S.Si., M.T

Achmad Junaidi, S.Kom., M.Kom

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2025

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 LATAR BELAKANG

Curah hujan memiliki peran penting dalam siklus hidrologi dan berdampak signifikan pada berbagai sektor kehidupan, termasuk pertanian, pengelolaan sumber daya air, dan mitigasi bencana hidrometeorologi. Beberapa tahun terakhir, dilakukan pengataman pada perubahan pola curah hujan di wilayah Indonesia yang menjadi perhatian serius. Dilansir dari Badan Klimatologi dan Geofisika (BMKG), laju perubahan curah hujan tahunan di Indonesia selama 30 tahun terakhir menunjukkan peningkatan tertinggi sebesar 2.784 mm dan penurunan terendah sebesar 750 mm(BMKG, 2023a). Perubahan ini menunjukkan adanya variasi yang signifikan dalam pola curah hujan, yang tentunya memengaruhi aktivitas manusia.

Perubahan pola curah hujan juga berdampak pada intensitas bencana alam yang terjadi. Sebagai contoh, dari data yang dirilis oleh BPBD pada tahun 2023 menunjukkan bahwa lebih dari 70% bencana alam yang terjadi di Indonesia selama diakibatkan oleh curah hujan yang ekstrem(Data Bencana Di Tingkat Kabupaten/Kota 2023). Melihat kondisi ini, dibutuhkan sistem prediksi yang akurat. Penerapan klasifikasi curah hujan memerlukan metode yang mampu menganalisis dan menguraikan pola data yang memiliki kompleksitas cukup tinggi. Metode umum yang sering digunakan adalah metode pembelajaran mesin atau *machine learning,* sebuah bidang kecerdasan buatan yang bertujuan membantu manusia menganalisis data dan membuat prediksi dari pelatihan data, pola dan fitur, serta melakukan klasifikasi dan pengujian(Rahman Sya'bani et al., 2022).

Machine learning yang digunakan pada penelitian ini adalah *extreme gradient boosting* (XGBoost). XGBoost merupakan Algoritma menggunakan metodologi ansambel berdasarkan pohon keputusan, di mana beberapa pohon keputusan dihasilkan secara berurutan untuk meningkatkan kinerja model yang komprehensif. Untuk mengoptimalkan kinerja XGBoost, diimplementasikan juga optimasi *hyperparameter tunning* yaitu *Bayessian.* Namun, seringkali akurasi pada prediksi curah hujan masih sulit direalisasikan karena kompleksitas data meteorologi yang dimiliki, seperti ketidakseimbangan data antara curah hujan rendah, menengah, dan tinggi. Ketidakseimbangan data ini jelas memengaruhi akurasi model prediksi yang sering kali

bias terhadap kelas mayoritas. Untuk mengatasi hal ini, terdapat beberapa teknik yang akan digunakan pada penelitian ini diantaranya, S*ynthetic Minority Oversampling Technique* (SMOTE), merupakan teknik umum dari *oversampling* untuk menangani ketidakseimbangan data dengan membuat sampel sintetis baru dalam kelas minoritas. Kemudian ada Tomek *Link* yang mewakili pendekatan metodologis untuk filtrasi *outlier* untuk mengatasi tantangan yang ditimbulkan dari data yang tidak seimbang melalui penghapusan pasangan data yang berdekatan tetapi berada pada kelas yang berbeda. Tujuan utama dari teknik ini adalah untuk mengurasi *noise* pada data dan meningkatkan penggambaran batas antara kelas yang berbeda. Penelitian ini juga akan mencoba menggabungkan kedua teknik tersebut untuk melihat efektifitasnya dalam menangani ketidakseimbangan data.

Penelitian oleh (Wiwaha et al., 2024) yang membahas tentang cara meningkatkan nilai akurasi model XGBoost dalam memprediksi curah hujan. Parameter yang digunakan dalam klasifikasinya yaitu *rain, humidity, temperature,* dan *light level.* Dalam penelitian tersebut didapatkan nilai *accuracy* 99.981%, *precision* mencapai nilai maksimum 100%, *recall* 99.943%, dan *F1-Score* 99.971%.

Tahun sebelumnya, (Sapari et al., 2023) menggunakan algoritma XGBoost untuk memfasilitasi klasifikasi kualitas udara. Parameter yang digunakan meliputi *Particulate Matter, Particulate Matter 2.5, Carbon Monoksida, Nitrogen Dioksida, Ozon,* dan *Nitrogen Dioksida.* Analisis menghasilkan skor *precision* 97%, *recall* 100%, *F1-Score* 98%, dan *accuracy* 98.61%.

Selanjutnya penelitian yang membahas tentang penggunaan teknik SMOTE pada klasifikasi objektivitas berita online menggunakan algoritma KNN. Nilai k tetangga yang digunakan diantaranya 1,3,5,7, dan 9. Kemudian pada hasil pengujian menunjukkan teknik SMOTE dapat meningkatkan nilai *accuracy* dengan k=1 meningkat 5.00 dan untuk k=3 meningkat 1.72 (Kasanah et al., 2019)

Selain itu, pada 2024 juga dilakukan penelitian tentang efektivitas penggunaan SMOTE-Tomek untuk meningkatkan nilai *confussion matrix* pada prediksi hipertensi yang hasilnya Smote-Tomek telah berhasil dalam meningkatkan akurasi, presisi, ingatan, dan skor F1 dari berbagai model, terutama dalam ranah meta-peserta didik, yang mencapai metrik kinerja sempurna 1,0. Temuan ini menggambarkan kemanjuran metodologi

penyeimbangan data dan proses pemilihan fitur dalam menambah kemampuan prediktif model pembelajaran mesin(Odiakaose et al., 2024).

Berdasarkan penelitian-penlitian sebelumnya, penulis ingin melakukan perbandingan performance model *Extreme Gradient Boosting* (XGBoost) menggunakan beberapa teknik *resampling* data dan *hyperparameter bayesian* untuk mengklasifikasi curah hujan di Provinsi Jawa Timur. Data yang digunakan adalah data iklim harian yang diakses melalui *website* resmi Badan Klimatologi dan Geofisika (BMKG) pada tahun 2020 – 2024. Perbandingan yang digunakan mempertimbangkan faktor-faktor penting seperti *accuracy, precision, recall, F1-Score,* dan *AUC.* Pada tahap akhir dapat diketahui teknik mana yang paling akurat dalam menangani masalah ketidakseimbangan data. Penelian ini dilakukan untuk mambahas secara mendalam bagaimana 2 teknik *resampling* data dapat bekerja secara optimal jika digabungkan.

1.2 RUMUSAN MASALAH

Dari uraian latar belakang diatas, adapun rumusan masalah yang dikemukakan pada penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh penggunaan teknik *oversampling* SMOTE (*Synthetic Minority Oversampling Technique),* metode penyaringan *outlier* Tomek *Link*, dan kombinasi keduanya SMOTE-Tomek terhadap kinerja model *Extreme Gradient Boosting* (XGBoost) dalam memprediksi curah hujan di Jawa Timur?
2. Bagaimana implementasi *hyperparameter tunning* menggunakan *Bayesian* dapat meningkatkan performa model *Extreme Gradient Boosting* (XGBoost)?
3. Bagaimana hasil penggunaan teknik *resampling* data yang memberikan hasil terbaik dalam meningkatkan akurasi pada model *Extreme Gradient Boosting* (XGBoost)?

1.3 TUJUAN PENELITIAN

Merujuk pada rumusan masalah diatas, adapun tujuan utama yang ingin dicapai penulis dari penelitian ini dibedakan menjadi 2 tujuan yaitu, tujuan umum dan tujuan khusus:

1.3.1. Tujuan Umum

Penelitian ini secara umum bertujuan untuk menganalisis dan membandingkan kinerja model *Extreme Gradient Boosting* (XGBoost) baik yang dioptimasi menggunakan *hyperparameter Bayesian* maupun tidak dalam memprediksi curah hujan di Jawa Timur. Analisis dilakukan dengan membandingkan kinerja model

menggunakan data yang diolah dengan Teknik *Resampling* Data yaitu *Synthetic Minority Oversampling Technique* (SMOTE), penyaringan *outlier* Tomek *Link,* dan mengkombinasikan kedua teknik SMOTE-Tomek.

1.3.2. Tujuan Khusus

Tujuan khusus dalam penelitian ini adalah sebagai berikut:

a. Mengimplementasikan teknik *resampling* data yaitu *Synthetic Minority Oversampling Technique* (SMOTE), Tomek *Link,* dan kombinasi SMOTE-Tomek pada model *Extreme Gradient Boosting* (XGBoost) untuk memprediksi curah hujan di Provinsi Jawa Timur.
b. Mengimplementasikan optimasi *hyperparameter tunning* yaitu *bayesian* pada model XGBoost.
c. Membandingkan kinerja model menggunakan 3 teknik *resampling* data berdasarkan metrik evaluasi *accuracy, precision, recall, F-1 Score,* dan *area under the curver (AUC).*
d. Membandingkan kinerja model sebelum dan sesudah dilakukan optimasi menggunakan *bayesian* berdasarkan *confussion matrix*

1.4 MANFAAT PENELITIAN

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Bagi Akademisi

   Penelitian diharapkan memberikan kontribusi dibidang akademik dengan memperdalam pemahaman mengenai implementasi teknik *resampling* data pada kasus nyata dan pengoptimalan model *machine learning,* khususnya *Extreme Gradient Boosting* (XGBoost) dalam memprediksi curah hujan. Diharapkan nantinya hasil dari penelitian ini dapat dijadikan referensi bagi penelitian serupa dimasa depan. Penelitian juga berpotensi mengembangkan teori-teori baru terkait teknik *resampling* data dan optimasi model, memberikan landasan ilmiah untuk penelitian lebih lanjut terkait bidang yang lebih luas lagi.

2. Bagi Praktisi

   Penelitian diharapkan memberikan wawasan tentang beberapa teknik *resampling* data meliputi *synthetic minority oversampling technique* (SMOTE), tomek *link,* dan kombinasi SMOTE-Tomek serta bagaimana implementasi *hyperparameter tunning* yaitu *bayesian* dapat mempengaruhi kinerja model XGBoost dalam memprediksi curah hujan. Selain itu, penelitian ini juga berpotensi membantu

meningkatkan akurasi dan keandalan model prediksi, yang diharapkan berdampak positif pada pengolahan resiko dan perencanaan strategis diberbagai sektor.

1.5 BATASAN MASALAH

Bagian dari uraian latar belakang yang telah disajikan, dapat diidentifikasikan rumusan masalah yang menjadi fokus pembahasan sebagai berikut:

a. Dara yang digunakan dalam penelitian merupakan data sekunder (data yang diambil secara tidak langsung oleh penulis) yang diakses melalui *website* resmi BMKG (https://dataonline.bmkg.go.id/)
b. Data yang digunakan adalah laporan iklim harian di Stasiun Klimatologi Jawa Timur pada tahun 2020 – 2024.
c. Peneliti menggunakan 10 parameter antara lain temperatur minimum (C), temperatur maksimum (C), temperatur rata-rata (C), kelembapan rata-rata (%), lamanya penyinaran matahari (jam), kecepatan angin maksimum (m/s), arah angin saat kecepatan maksimum (`), kecepatan angin rata-rata (m/s), arah angin terbanyak (`), dan curah hujan (mm).
d. Luaran yang dihasilkan adalah *accuracy, precision, recall, F1-Score,* dan *area under the curve* dari masing – masing skenario pengujian yang ditentukan.

BAB II
TINJAUAN PUSTAKA

2.1 PENELITIAN TERDAHULU

Berdasarkan pada teori dan penelitian terdahulu, kajian terhadap penggunaan model *gradient boosting,* khususnya XGBoost, dalam prediksi data menjadi acuan penting untuk penelitian ini. Selain itu, optimasi *hyperparameter* menggunakan pendekatan *bayesian*

1. *Modelling and Forecasting Rainfall Patterns in India: a time series Analysis with XGBoost Algorithm* (2024)

   Penelitian yang dilakukan (Mishra et al., 2024a)menyajikan analisis menyeluruh tentang pola curah hujan di India dengan menggunakan metodologi pembelajaran mesin, khususnya XGBoost, di samping model statistik tradisional. Penelitian menggunakan analisis deret waktu untuk membangun model dan memprediksi curah hujan di berbagai periode musin di India. Studi ini membandingkan berbagai metodologi termasuk model *Autoregressive Integrated Moving Average* (ARIMA), model *state space,* serta berbagai model pendekatan mesin seperti *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), *Random Forest,* dan XGBoost. Evaluasi kinerja model pada penelitian ini dilakukan dengan menerapkan beberapa metrik evaluasi, seperti *Akaike Information Criterion* (AIC), *Bayesian Information Criteria* (BIC), *Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan Mean Absolute Persentase Error (MAPE).*

   Hasil penelitian menunjukkan bahwa XGBoost menunjukkan keunggulan secara signifikan dalam menangkap pola curah hujan nonlinier yang kompleks ketika disandingkan dengan model statistik tradisional. Hasil menunjukkan performa XGBoost dengan nilai error terkecil di seluruh musim. Secara keseluruhan XGBoost memiliki MAE 0.92, RMSE 1.05, dan MAPE 1.001. Model ARIMA menunjukkan kecenderungan untuk menyesuaikan kumpulan data latih, sedangkan model *space state* memiliki keunggulan dalam menghadapi *outlier.*

   Namun, pada jurnal disebutkan bahwa XGBoost mungkin tidak selalu bekerja optimal pada dataset yang tidak seimbang. Oleh karena itu, penelitian saat ini akan membandingkan performa model XGBoost saat bekerja menggunakan dataset mentah dan dataset yang sudah diseimbangkan terlebih dahulu untuk mengisi gap dari jurnal ini. Selain itu, jurnal juga tidak memberikan rincian tentang proses *tuning*

parameter yang digunakan pada model XGBoost. Ketiadaan rincian pada proses *tuning* ini nantinya ingin dijelaskan pada penelitian sekarang, seperti paratemer – parameter apa yang diatur untuk model XGBoost, meliputi *learning rate, max depth, subsample, n_estimators* dan lainnya yang memiliki pengaruh signifikan terhadap performa model.

2. *Rainfall Prediction Using Extreme Gradient Boosting* (2020)

Penelitian yang dilakukan oleh (Anwar et al., 2021)menyajikan pemeriksaan komprehensif peramalan curah hujan menggunakan metodologi *Extreme Gradient Boosting* (XGBoost) dan menjelaskan manfaatnya dibandingkan dengan kerangka prediksi konvensional. Selanjutnya, naskah ini menonjolkan keterbatasandalam metodologi tradisional, seperti ARIMA dan *Exponential Smoothing*, yang pada dasarnya bergantung pada anggapan asosiasi linier dalam data—asumsi yang sering terbukti tidak memadai untuk sifat nonlinier dan rumit pola curah hujan. Untuk memperkuat analisis ini, data meteorologi harian yang bersumber dari BMKG yang mencakup jangka waktu tujuh tahun (2013—2019) digunakan untuk pelatihan model, dengan data 2020 yang ditujukan untuk evaluasi model. Dataset mencakup 11 variabel meteorologi, termasuk suhu, kelembaban, dan tekanan; Namun, hanya delapan variabel yang dipilih berdasarkan relevansinya, dengan kelembaban relatif rata-rata (Rh_AVG) dan suhu minimum (Tn) diakui sebagai penentu paling konsekuensial dalam peramalan curah hujan.

Rejimen pelatihan menggunakan algoritma XGBoost, menggabungkan validasi silang 10 kali lipat untuk mencegah kelebihan pemasangan dan untuk menjamin generalisasi yang efektif untuk data yang tidak terlihat. Model menjalani beberapa iterasi pelatihan, dengan *Root Mean Square Error* (RMSE) yang direkam selama fase pelatihan ditetapkan pada 2,7 mm, sehingga menunjukkan penyimpangan rata-rata yang relatif kecil dari prediksi dalam kaitannya dengan pengamatan aktual. Meskipun demikian, model menunjukkan indikasi *overfitting* ketika iterasi pelatihan melebihi putaran kelima, di mana kesalahan pengujian mulai meningkat. Pengamatan ini menggarisbawahi kekritisan penghentian pelatihan pada titik optimal untuk menjaga keseimbangan antara presisi pelatihan dan kemampuan generalisasi. Selain itu, XGBoost disandingkan dengan berbagai metodologi alternatif, dan temuan menggambarkan bahwa XGBoost mengungguli yang lain, meskipun penulis mengakui adanya peluang untuk meningkatkan akurasi melalui pengoptimalan parameter lebih lanjut.

Melihat hal tersebut, pada penelitian sekarang akan dilakukan eksperimen pengoptimalan parameter lebih lanjut menggunakan *hyperparameter tuning* melalui pendekatan *bayesian optimization.* Diharapkan melalui implementasi *hyperparameter tuning* dapat mengatasi masalah *overfitting* yang terjadi pada jurnal ini, sehingga akurasi dari model juga dapat meningkat.

3. *A heart Disease Prediction Model Based on Features Optimization and Smote-XGBoost Algorithm* (2022)

Penelitian berjudul "Model Prediksi Penyakit Jantung Berdasarkan Optimasi Fitur dan Algoritma SMOTE-XGBoost"(Yang & Guan, 2022) menggambarkan metodologi baru untuk meramalkan penyakit jantung melalui penerapan teknik pembelajaran mesin. Investigasi ini terutama berkonsentrasi pada peningkatan presisi dan kemanjuran diagnostik klinis, terutama sehubungan dengan *Major Adverse Cardiovascular Events* (MACCE), dengan menggunakan data patologis dari pasien jantung. Metodologi yang digunakan mencakup pemilihan fitur melalui teknik akuisisi informasi untuk mengurangi risiko *overfitting*, penyeimbangan kembali data melalui penerapan SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan kelas dalam kumpulan data, dan evaluasi model dengan menyandingkan algoritma XGBoost terhadap lima algoritma fundamental alternatif. Temuan mengungkapkan bahwa model SMOTE-XGBoost mencapai tingkat akurasi 93,44%, sehingga memperkuat keunggulannya dibandingkan dengan pendekatan konvensional. Analisis data eksplorasi menggunakan koefisien korelasi *Pearson* dan peta panas memfasilitasi identifikasi fitur yang paling relevan, sehingga meningkatkan ketahanan model dalam prediksi MACCE. Penelitian ini menggarisbawahi signifikansi kritis dari kesetimbangan data dan pemilihan fitur yang cermat dalam pembangunan model prediktif yang manjur. Melalui integrasi optimasi fitur dan algoritma pembelajaran mesin canggih, penelitian ini memberikan kontribusi besar terhadap evolusi teknologi prediksi penyakit jantung, sambil menggambarkan implikasi mendalam dari pembelajaran mesin dalam domain medis.

Meskipun SMOTE telah terbukti efektif dalam menangani masalah ketidakseimbangan data, terdapat beberapa area yang perlu dieksplorasi untuk meningkatkan pemahaman dan penerapan teknik dalam konteks prediksi penyakit jantung dan yang lainnya. Penelitian yang ada sering kali tidak membahas pengaruh

dari parameter SMOTE, seperti jumlah sampel sintesis yang dihasilkan dan pemilihan tetangga terdekat (k-*nearest neighbors),* terhadap kinerja model. Oleh karena itu, pada penelitian saat ini, akan dibahas bagaimana pengaturan parameter dapat dioptimalkan untuk mencapai hasil akurasi yang lebih baik.

Meskipun beberapa penelitian telah mulai mengeksplorasi kombinasi dari MOTE dan teknik *balancing* lain seperti SMOTE-ENN, masih terdapat beberapa potensi kombinasi SMOTE dengan teknik *balancing* lain seperti tomek *link* untuk menciptakan dataset yang lebih seimbang dan meningkatkan kinerja model. Dengan mengisi gap ini, penelitian saat ini diharapkan dapat memberikan kontribusi terhadap pengembangan metodologi yang lebih efektif.

4. *Bagging of XGBoost Classifiers with Random Undersampling dan Tomek Link for Noisy Label-Inbalance Data*

Jurnal berjudul *"Bagging of XGBoost Classifier with Random Undersampling and Tomek Link for Noisy Label-Imbalance Data"* menggunakan algoritma baru dengan menggabungkan teknik *bagging* dengan *classifier* XGBoost dan metode *under sampling,* khususnya Tomek *Link*, untuk menangani masalah klasifikasi data yang tidak seimbang dan *noise.* Metode yang digunakan bertujuan untuk meningkatkan nilai akurasi dan efektivitas dalam pengenalan pola, terutama dalam konteks *imbalance dataset*. Metode yang digunakan antara lain *random sampling with replacement* digunakan untuk melakukan pengambilan sampel acak dengan penggantian pada data mayoritas untuk menghasilkan subset data yang seimbang. Teknik ini bertujuan untuk mengurangi varians dalam model yang dihasilkan. Kemudian metode selanjutnya adalah Tomek *Link elimination* digunakan untuk menghapus pasangan data yang tumpang tindih antara kelas mayoritas dan kelas minoritas. Teknik ini dilakukan untuk mengurangi *noise* pada data dan meningkatkan kualitas dataset yang digunakan untuk pelatihan. Hasil dari pengolahan dataset yang dilakukan kemudian dilatih menggunakan model XGBoost.

Namun, pada jurnal tidak dijelaskan bagaimana potensi teknik Tomek *Link* jika dikombinasikan dengan teknik balancing lain seperti SMOTE atau teknik *oversampling* lain untuk meningkatkan kinerja dalam konteks data yang tidak seimbang. Selain itu, penelitian saat ini juga akan mengeksplorasi optimasi parameter pada Tomek *Link,* termasuk pemilihan metriks jarak dan lain-lain yang diterapkan pada dataset lain, yaitu dataset milik BMKG untuk memprediksi curah hujan. Diharapkan dengan melakukan penelitian ini, dengan menerapkan beberapa

teknik tambahan, dapat memberikan hasil akurasi model yang lebih tinggi dan membuktikan bahwa model XGBoost dapat bekerja optimal meskipun dengan kompleksitas dataset yang berbeda pada setiap bidang.

5. Tomek *Link* and SMOTE *Approaches for Machine Fault Classification with an Imbalanced Dataset.*

Penelitian oleh (Swana et al., 2022) membahas penerapan teknik *resampling,* khususnya Tomek *Link* dan *Synthetic Minority Over Sampling* (SMOTE), dalam konteks pemantauan kondisi mesin dan deteksi kesalahan. Penelitian berfokus pada metode berbasis data untuk mendiagnosis kesalahan mesin, yang memanfaatkan teknik pembelajaran mesin untuk menganalisis data yang dihasilkan dalam berbagai kondisi. Metodologi yang digunakan antara lain pemodelan mesin menggunakan SVM dan KNN, ekstraksi fitur, dan penerapan teknik *resampling* data. Tomek *Link* digunakan sebagai pembersih data pasca-proses, sedangkan SMOTE digunakan untuk menambah jumlah sampel sintesis untuk meningkatkan jumlah data pada kelas minoritas dengan mencari nilai tetangga terdekat. Penerapan Tomek *Link* pasca-proses SMOTE dimaksudkan untuk mengurangi *noise* pada dataset yang dapat mempengaruhi kinerja model.

Hasil eksperimen menunjukkan bahwa penerapan Tomek *Link* dan SMOTE secara signifikan meningkatkan akurasi kinerja model klasifikasi yaitu SVM dan k-NN, dibandingkan dengan data yang tidak seimbang. Hasil dari kombinasi kedua teknik ini menunjukkan akurasi minimum sebesar 0.63 dan maksimum sebesar 0.78. Kedua teknik *resampling* ini berhasil memberikan hasil yang lebih baik dibandingkan penggunaan masing-masing teknik secara terpisah.

Jurnal merekomendasikan penelitian lebih lanjut untuk mengeksplorasi teknik *resampling* dengan menggabungkan dengan model lain seperti XGBoost. Melihat XGBoost memiliki parameter yag dapat disesuaikan untuk menangani masalah *imbalance class,* seperti penyesuaian bobot kelas. Selain itu, XGBoost juga dikenal karena kinerjanya yang superior dalam berbagai kompetisi *machine learning* dan aplikasi di dunia nyata. XGBoost menawarkan fleksibilitas dalam hal oemilihan fitur dan dapat memberikan interpretasi yang lebih baik. XGBoost juga memungkinkan pengoptimalan *hyperparameter* yang lebih mendalam, yang dapat meningkatkan kinerja model secara signifikan. Diharapkan penggunaan XGBoost dengan teknik *resampling* data yaitu SMOTE dan Tomek *Link* dapat mengisi areap gap dari jurnal ini dan memberikan potensi peningkatan dalam *machine learning.*

6. *A Bayessian Optimization Based LSTM Model for Wind Power Forecasting in the Adama District, Eithiopia* (2023)

Jurnal berjudul *"A Bayessian Optimization Based LSTM Model for Wind Power Forecasting in the Adama District, Eithiopia"* oleh (Habtemariam et al., 2023)membahas tentang pengusulan model yang bertujuan untuk meningkatkan akurasi pembangkit energi angin dengan menggunakan *Long Short Term Memory* (LSTM) yang dioptimalkan menggunakan *bayessian optimization.* Metodologi yang digunakan pada penelitian ini dijelaskan dengan rinci mulai dari pengumpulan data hingga membangun model LSTM menggunakan *bayessian optimization.* Hasil dari model LSTM ini dibandingkan dengan model *baseline* XGBoost dan ARIMA. Hasil eksperimen menunjukkan bahwa model LSTM mengungguli model *baseline* dalam hal metrik evaluasi seperti *Mean Absolute Error (MAE), Root Mean Square Error (RMSE),* dan *Mean Absolute Percentage Error (MAPE)* menggunakan data pembangkit energi angin di Ethiopia.

Meskipun hasil akhir menunjukkan keunggulan model LSTM dibandingkan dengan XGBoost, pada penelitian ini ingin mencoba eksperimen menggunakan dataset dari BMKG untuk memprediksi curah hujan di provinsi Jawa Timur menggunakan *bayessian optimization.* Dataset juga terlebih dahulu dilakukan *balancing* menggunakan teknik SMOTE dan Tomek *Link.* Penelitian saat ini diharapkan dapat memberikan wawasan baru tentang penggunaan *hyperparameter tuning* yaitu *bayessian optimization* jika dijalankan dengan XGBoost dengan dataset yang berbeda apakah mendapatkan hasil yang optimal.

2.2 LANDASAN TEORI

2.2.1. *Extreme Gradient Boosting* (XGBoost)

XGBoost merupakan algoritma pembelajaran mesin yang efisien, fleksibel, dan ringan, yang sering diterapkan dalam data mining dan sistem rekomendasi. Algoritma ini beroperasi dengan membangun model prediksi yang terdiri dari beberapa pohon keputusan (*decision trees*)(Ke et al., 2022). Setiap pohon keputusan dilatih secara bertahap, dan hasil dari semua pohon tersebut digabungkan untuk menghasilkan prediksi akhir. XGBoost tidak membuat asumsi tentang bentuk hubungan antara prediktor dan target, yang membuatnya lebih fleksibel dibanding model-model linear. Ini memungkinkan XGBoost untuk menangkap hubungan yang kompleks dan non-linear seperti dalam dataset curah hujan(Mishra et al., 2024b).

Model berbasis pohon, seperti XGBoost juga lebih tahan terhadap *outliers* dibandingkan dengan model linear, sehingga lebih stabil dalam kondisi data yang bervariasi. Dalam XGBoost terdapat 2 fungsi objektif yang dibangun untuk meminimalkan kesalahan prediksi, yaitu *loss function* untuk mengukur kesalahan antar nilai yang diprediksi dan nilai sebenarnya dan *regularization term* digunakan untuk mengontrol kompleksitas model untuk mencegah *overfitting(Ke et al., 2022)*. Berikut ini adalah komponen kunci dari algoritma XGBoost:

- Fungsi Loss: XGBoost melakukan minimalisir fungsi loss untuk meningkatkan akurasi model. Fungsi loss umum yang digunakan adalah:

$$L(y, \hat{y}) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

- Model Prediksi: XGBoost dibangun sebagai beberapa dari pohon keputusan dengan rumus:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i)$$

- Gradient Boosting: XGBoost menggunakan pendekatan gradient boosting, dimana setiap pohon baru dibangun untuk mengurangi kesalahan dari pohon sebelumnya. Gradien dari fungsi loss dihitung untuk setiap prediksi:

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \widehat{y_i^2}}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \widehat{y_i^2}}$$

- Pembentukan Pohon: XGBoost membangun pohon keputusan dengan meminimalkan fungsi loss yang diperbarui dengan menambahkan pohon baru dengan rumus:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \mu f_t(x_i)$$

- Regulasi: XGBoost juga menerapkan regulasi untuk menghindari overfitting. Fungsi loss dimodifikasi menjadi:

$$L(y, \hat{y}) + \sum_{k=1}^{K} \cap (f_k)$$

Dimana: $\cap (f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$

2.2.2. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan metode yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, terutama dalam konteks klasifikasi(Alamri & Ykhlef, 2024). SMOTE sendiri adalah teknik oversampling yang menciptakan contoh baru dalam kelas minoritas dengan cara mensintesis data. Alih-alih hanya menduplikasi contoh yang ada, SMOTE menghasilkan contoh baru dengan interpolasi antara contoh minoritas yang ada. Ini membantu dalam memperkaya informasi yang tersedia untuk model klasifikasi(Alamri & Ykhlef, 2024).

Teknik SMOTE menawarkan ketahanan maksimal terhadap variasi rasio ketidakseimbangan. Hal ini menunjukkan bahwa SMOTE dapat secara efektif bekerja meskipun terdapat perubahan dalam proporsi kelas mayoritas dan minoritas yang sering terjadi dalam pengaplikasiannya. Selain itu, SMOTE juga memiliki desain yang sederhana namun efektif, terdiri dari komponen – komponen utama yang bertanggung jawab untuk representasi data, *resampling,* dan klasifikasi. Ini membuatnya mudah diterapkan dalam berbagai konteks(Dablain et al., 2023).

Terdapat langkah – langkah dalam implementasi teknik SMOTE. Langkah pertama adalah mengidentifikasi kelas minoritas dalam dataset. Kelas ini merupakan kategori yang memiliki jumlah sampel yang lebih sedikit dibandingkan dengan kelas minoritas. Kedua, memilih secara acak sampel x_i dari kelas minoritas. Nilai ini akan menjadi titik awal untuk menghasilkan sampel sintesis. Ketiga menentukan nilai *k*, yang merupakan jumlah tetangga terdekat untuk dipertimbangkan. Nilai *k* ini biasanya ditentukan berdasarkan eksperimen sebelumnya atau menggunakan metode *cross validation.* Keempat, setelah menentukan nilai *k*, selanjutnya adalam menghitung jarak antara sampel x_i yang dipilih secara acak dan semua sampel lain dalam kelas minoritas. Pada langkah ini digunakan metrik jarak seperti Euclidean *distance* untuk menemukan *k* tetangga dari *x.* Rumus menghitung jarak Euclidan antara dua titik x_i dan x_j adalah:

$$dx_{i,}x_j = \sqrt{\sum_{m=1}^{M}(x_{i,m} - x_{j,m})^2}$$

Kelima menghitung interpolasi untuk membuat sampel sintetis. Untuk setiap tetangga terdekat xj yang ditemukan, dibuat sampel sintetis baru xnew menggunakan rumus:

$$x_{new} = x_i + \lambda \,.\, (x_j - x_i)$$

Dimana λ adalah nilai acak yang diambil dari distribusi uniform antara 0 dan 1, yaitu λ ∼U (0,1). Langkah keenam adalah mengulagi proses kedua hingga kelima untuk setiap sampel dalam kelas minoritas hingga jumlah sampel sintetis yang diinginkan tercapai. Terakhir menggabungkan sampel sintesis dengan dataset asli untuk menciptakan dataset yang lebih seimbang antara kelas minoritas dan mayoritas.

2.2.3. Tomek *Link*

Tomek Links adalah metode yang diperkenalkan oleh Ivan Tomek untuk membersihkan data dalam konteks klasifikasi. Metode ini berfokus pada menemukan pasangan instance dari kelas yang berlawanan yang berdekatan satu sama lain dalam ruang fitur. Sebuah pasangan disebut sebagai Tomek Link jika dua instance, satu dari kelas minoritas dan satu dari kelas mayoritas, memiliki jarak Euclidean yang minimal dan tidak ada instance lain yang lebih dekat ke salah satu dari mereka dibandingkan dengan jarak antara keduanya(Ainaa Hanis Zuhairi et al., 2024a). Tujuan utama dari penggunaan Tomek Links adalah untuk memperbaiki batas keputusan antara kelas-kelas yang berbeda. Dengan menghapus instance yang merupakan Tomek Links, kita dapat membersihkan data dari noise dan memperjelas batas antara kelas mayoritas dan minoritas(Le et al., 2024). Tomek Links sering digunakan sebagai metode undersampling, di mana contoh dari kelas mayoritas dihapus untuk menyeimbangkan dataset. Dengan fokus pada penghapusan contoh mayoritas yang membentuk Tomek Links, kita dapat mengurangi ukuran kelas mayoritas tanpa kehilangan informasi penting dari kelas minoritas(Alamri & Ykhlef, 2024).

Proses implementasi Tomek *Link* yang pertama setiap instance dalam dataset, identifikasi tetangga terdekat (nearest neighbors) dari instance tersebut. Jika terdapat pasangan Tomek Links, hapus instance dari kelas mayoritas yang berdekatan dengan instance dari kelas minoritas(Ainaa Hanis Zuhairi et al., 2024a). Proses ini membantu dalam mengurangi kebisingan dalam data dan meningkatkan kemampuan model untuk belajar dari data yang lebih bersih. Rumus dari proses Tomek *Link* secara umum adalah:

Jika x_i, y_i dan x_j, y_j adalah Tomek *Link,* maka:

hapus x_i, y_i jika y_i adalah kelas mayoritas.

2.2.4. *Bayesian Optimazion*

Bayesian Optimization (BO) dijelaskan melalui dua komponen utama, yaitu Gaussian Processes (GPs) dan Acquisition Functions (AFs). GPs berfungsi sebagai model probabilistik yang digunakan untuk memperkirakan fungsi objektif yang tidak diketahui, dengan asumsi bahwa data yang diamati mengikuti distribusi Gaussian. Model ini ditandai oleh fungsi rata-rata dan fungsi kovarians yang memungkinkan prediksi yang efisien meskipun dengan jumlah evaluasi yang terbatas. Sementara itu, AFs digunakan untuk menentukan titik-titik evaluasi berikutnya dengan memaksimalkan ekspektasi perbaikan atau informasi yang diperoleh dari model GP. Jurnal ini juga mengkategorikan berbagai algoritma BO yang ada berdasarkan sifat masalah optimisasi yang dihadapi, serta membahas tantangan dan arah penelitian masa depan dalam bidang ini, termasuk isu-isu seperti heterogenitas, privasi, dan keadilan dalam sistem optimisasi terdistribusi(Wang et al., 2023).

Kelebihan Bayesian optimization terletak pada kemampuannya untuk secara efisien menemukan kombinasi hyperparameter optimal dalam ruang parameter yang kompleks dan multimodal, yang sering kali muncul dalam pelatihan model pembelajaran mendalam. Metode ini menggunakan model probabilistik untuk memperkirakan fungsi objektif, sehingga dapat meminimalkan jumlah evaluasi yang diperlukan untuk mencapai hasil yang baik. Dengan pendekatan ini, Bayesian optimization tidak hanya mengurangi waktu komputasi yang dibutuhkan, tetapi juga mengurangi risiko overfitting dengan memilih set hyperparameter yang menjanjikan berdasarkan informasi yang diperoleh dari iterasi sebelumnya. Selain itu, metode ini dapat menangani fungsi objektif yang mahal untuk dievaluasi, menjadikannya sangat cocok untuk aplikasi dalam pelatihan model yang memerlukan sumber daya komputasi yang besar, seperti jaringan saraf dalam(Habtemariam et al., 2023).

Rumus dasar dari Bayesian Optimization dapat dijelaskan melalui Teorema Bayes, yang merupakan fondasi dari pendekatan ini. Teorema Bayes dinyatakan sebagai berikut:

$$P(A|B) = P(B)P(B|A).P(A)$$

Di mana:

- P(A|B) adalah probabilitas posterior, yaitu probabilitas dari hipotesis A setelah melihat bukti B.
- P(B|A) adalah probabilitas likelihood, yaitu probabilitas dari bukti B jika hipotesis A benar.
- P(A) adalah probabilitas prior, yaitu probabilitas awal dari hipotesis A sebelum melihat bukti B.
- P(B) adalah probabilitas marginal, yaitu probabilitas dari bukti B di seluruh ruang hipotesis.

2.2.5. *Imbalanced Dataset*

Imbalanced dataset merupakan masalah umum dalam tugas klasifikasi yang terjadi ketika distribusi instance antara kelas mayoritas dan kelas minoritas sangat tidak seimbang(Ainaa Hanis Zuhairi et al., 2024). Situasi ini dapat menyebabkan penurunan kinerja yang signifikan pada model klasifikasi, karena algoritma cenderung memprioritaskan akurasi pada kelas mayoritas dan mengabaikan kelas minoritas(Ruisen et al., 2018). Untuk mengatasi masalah ini, diperlukan berbagai strategi seperti *resampling* data.

Dataset tidak seimbang ditandai oleh perbedaan signifikan dalam jumlah contoh antara kelas-kelas yang ada, di mana satu kelas memiliki jumlah yang jauh lebih banyak dibandingkan kelas lainnya. Ciri ini dapat terlihat melalui visualisasi distribusi kelas yang menunjukkan dominasi kelas mayoritas, misalnya, dalam prediksi curah hujan, di mana jumlah hari hujan jauh lebih sedikit dibandingkan dengan hari tidak hujan. Model yang dilatih pada dataset semacam ini sering kali menunjukkan kinerja yang buruk dalam mengidentifikasi kelas minoritas, dengan metrik evaluasi seperti akurasi yang dapat memberikan gambaran menyesatkan tentang kinerja model. Hal ini disebabkan oleh fakta bahwa model dapat mencapai akurasi tinggi hanya dengan memprediksi kelas mayoritas, sementara gagal mengenali kelas minoritas. Selain itu, model tersebut mungkin mengalami kesulitan dalam generalisasi pada data baru yang memiliki proporsi kelas berbeda, sehingga memerlukan teknik penanganan khusus seperti oversampling atau undersampling untuk meningkatkan akurasi dan keandalan prediksi.

2.2.6. Prediksi Curah Hujan

Curah hujan merupakan salah satu komponen penting dalam sistem iklim yang dipengaruhi oleh berbagai faktor atmosferik dan geofisik(Bimaprawira & Rejeki, 2021). Curah hujan adalah jumlah air yang jatuh ke permukaan bumi dalam bentuk hujan, salju, atau bentuk presipitasi lainnya selama periode tertentu, biasanya diukur dalam milimeter (mm) atau inci. Curah hujan merupakan salah satu komponen penting dalam siklus hidrologi dan memiliki dampak signifikan terhadap berbagai aspek kehidupan, termasuk pertanian, pengelolaan sumber daya air, dan ekosistem(Anwar et al., 2021).

Pola curah hujan di Provinsi Jawa Timur sendiri secara signifikan dibentuk oleh fitur topografi, dinamika angin monsun, dan fenomena iklim global yang menyeluruh, terutama El Niño dan La Niña. Topografi yang beragam, mencakup dataran rendah dan medan pegunungan, secara kritis mempengaruhi distribusi curah hujan; khususnya, daerah yang ditinggikan seperti Semeru dan Arjuno cenderung mengalami peningkatan curah hujan karena pengangkatan orografis (Suryadi et al., 2020). Dinamika pola angin monsun juga memberikan dampak yang cukup besar, di mana angin barat mengangkut udara yang sarat kelembaban dari Samudra Hindia, sehingga menambah curah hujan selama musim hujan, sementara angin timur yang berasal dari Australia yang gersang berkontribusi pada pengurangan curah hujan selama musim kemarau (BMKG, 2023)

Selain itu, fenomena atmosfer global seperti El Niño dan La Niña sangat mempengaruhi intensitas curah hujan. Fenomena El Niño, yang ditandai dengan peningkatan suhu permukaan laut di Samudra Pasifik, sering mengakibatkan penurunan curah hujan dan meningkatkan kerentanan terhadap kondisi kekeringan. Sebaliknya, peristiwa La Niña dikaitkan dengan penurunan suhu permukaan laut, yang menyebabkan peningkatan tingkat curah hujan, yang dapat memicu banjir di daerah tertentu(Fais Putra, 2024). Selain itu, distribusi spasial presipitasi ini lebih lanjut dimodulasi oleh berbagai dinamika atmosfer, termasuk Gelombang Khatulistiwa Rossby dan Osilasi Madden-Julian (BMKG, 2020). Variabilitas multifaset ini biasanya menghasilkan puncak curah hujan yang terjadi antara bulan Januari dan Maret, dengan perbedaan intensitas yang nyata yang dapat diamati antara sektor utara dan selatan Jawa Timur.

Dalam dekade kontemporer, metodologi yang didasarkan pada kecerdasan buatan telah mendapatkan keunggulan yang cukup besar karena kemahiran mereka dalam mengelola kumpulan data yang rumit dan banyak. Algoritma komputasi, termasuk Random Forest, Support Vector Machines, dan XGBoost, menunjukkan peningkatan kapasitas untuk meramalkan curah hujan dengan tingkat presisi yang lebih tinggi dibandingkan dengan teknik konvensional. Meskipun demikian, pendekatan ini memerlukan kumpulan data pelatihan yang ekstensif dan sumber daya komputasi yang substansif. Fenomena curah hujan tunduk pada serangkaian faktor yang mempengaruhi, meliputi fitur topografi, lintasan angin, tekanan atmosfer, kondisi termal, dan fenomena global seperti El Niño dan La Niña. Fenomena El Niño sering mengakibatkan penurunan curah hujan karena suhu permukaan laut yang meningkat di Samudra Pasifik, sedangkan La Niña cenderung menambah curah hujan sebagai konsekuensi dari penurunan suhu samudera. Selanjutnya, dinamika atmosfer, dicontohkan oleh Madden-Julian Osilation (MJO), memberikan pengaruh terhadap siklus curah hujan setiap minggu hingga bulanan, sehingga membuat pola curah hujan sangat bervariasi

Namun demikian, kendala utama dalam ranah peramalan curah hujan terletak pada ketidakpastian data yang melekat dan sifat rumit dinamika atmosfer. Implikasi perubahan iklim semakin memperburuk variabilitas pola curah hujan, sehingga memerlukan pengembangan model prediksi yang lebih adaptif dan tepat. Dalam konteks praktis, prakiraan curah hujan memainkan peran penting dalam mengoptimalkan jadwal penanaman dalam domain pertanian, mengantisipasi risiko yang terkait dengan banjir dan kekeringan, dan memfasilitasi pengelolaan sumber daya air. Melalui penggunaan metodologi yang tepat dan kumpulan data yang akurat, peramalan curah hujan dapat secara signifikan berkontribusi pada kemajuan inisiatif pembangunan berkelanjutan .

BAB III
DESAIN DAN IMPLEMENTASI SISTEM

3.1 METODE PENELITIAN

Metodologi mencakup kumpulan teknik, prinsip, dan prosedur komprehensif yang digunakan dalam bidang atau disiplin ilmu tertentu untuk memfasilitasi penelitian yang cermat dan sistematis. Metodologi menyediakan struktur atau arahan yang mendukung peneliti atau praktisi dalam merencanakan, menjalankan, dan mengevaluasi penelitian atau tindakan tertentu. Tujuan utama metodologi adalah memastikan bahwa penelitian atau tindakan dilaksanakan dengan sistematis, terorganisir, dan dapat dipercaya, sehingga dapat meningkatkan pemahaman tentang fenomena yang sedang diteliti. Metodologi penelitian ini sebagai berikut:

3.1.1 Kebutuhan Perangkat

Dalam menunjang penulis dalam melakukan penelitian ini, terdapat dua kebutuhan perangkat yang digunakan yaitu perangkat keras dan perangkat lunak. Berikut kebutuhan perangkat yang digunakan:

3.1.1.1 Perangkat Keras (Hardware)

- Laptop Acer Aspire 5
- Processor Intel ® Core™ I3
- RAM 12 GB

3.1.1.2 Perangkat Lunak (Software)

- Sistem Operasi Windows 10 Home
- Google Colaboratory
- Python
- File CSV yang didapat dari website BMKG (https://dataonline.bmkg.go.id/)

3.1.2 Sumber Data

Sumber Data Penelitian Terdapat dua jenis data yang sering digunakan dalam penelitian, yaitu data primer dan data sekunder. Data primer adalah data yang diperoleh secara langsung melalui proses observasi, interaksi langsung dengan subjek, wawancara, distribusi kuesioner, atau hasil dari pengujian yang dilakukan. Data sekunder merujuk pada informasi yang diperoleh dari sumber-sumber yang sudah ada sebelumnya, seperti jurnal, buku, atau dokumen resmi lainnya. Pada penelitian ini, penulis menggunakan data sekunder.

Data yang penulis gunakan adalah data meteorologi yang mencakup informasi tentang curah hujan harian di Provinsi Jawa Timur dari tahun 2020 hingga 2024, yang diakses melalui website resmi BMKG (https://dataonline.bmkg.go.id/). Dataset ini berisi berbagai parameter seperti temperatur maksimum, temperatur minimum, kelembapan rata-rata, kecepatan angin, dan curah hujan. Data diunduh dalam format CSV dan telah diproses lebih lanjut untuk keperluan analisis. Total data yang digunakan dalam penelitian ini mencakup lebih dari 2000 baris data dengan 12 variabel utama.

Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
21,8	30,8	27,1	78	8,6	6,4	4	180	2	C
22,8	31,1	27,1	77	0	6,1	4	80	1	C
21,9	30,8	26,4	79	0	6,5	6	70	3	C
21,9	31	26,1	77	0	4,6	6	90	2	C
21,4	29,8	25,2	83	0	5,1	3	160	1	C
20,9	28,6	25,1	84	7,8	4,8	4	230	1	C
21,3	29	25	84	6,3	0,9	5	60	2	C
21,2	29,6	25,6	82	5,9	1,8	3	40	1	C
21,4	31	25,9	81	0	2,6	4	110	1	C
21	30,9	26,4	77	2,6	6	4	100	2	C
22,4	31,5	27,4	71	0	6,5	3	150	2	E

Table 1. Dataset BMKG

Keterangan:

8888: data tidak terukur
9999: Tidak Ada Data (tidak dilakukan pengukuran)
Tn: Temperatur minimum (°C)
Tx: Temperatur maksimum (°C)
Tavg: Temperatur rata-rata (°C)
RH_avg: Kelembapan rata-rata (%)
RR: Curah hujan (mm)

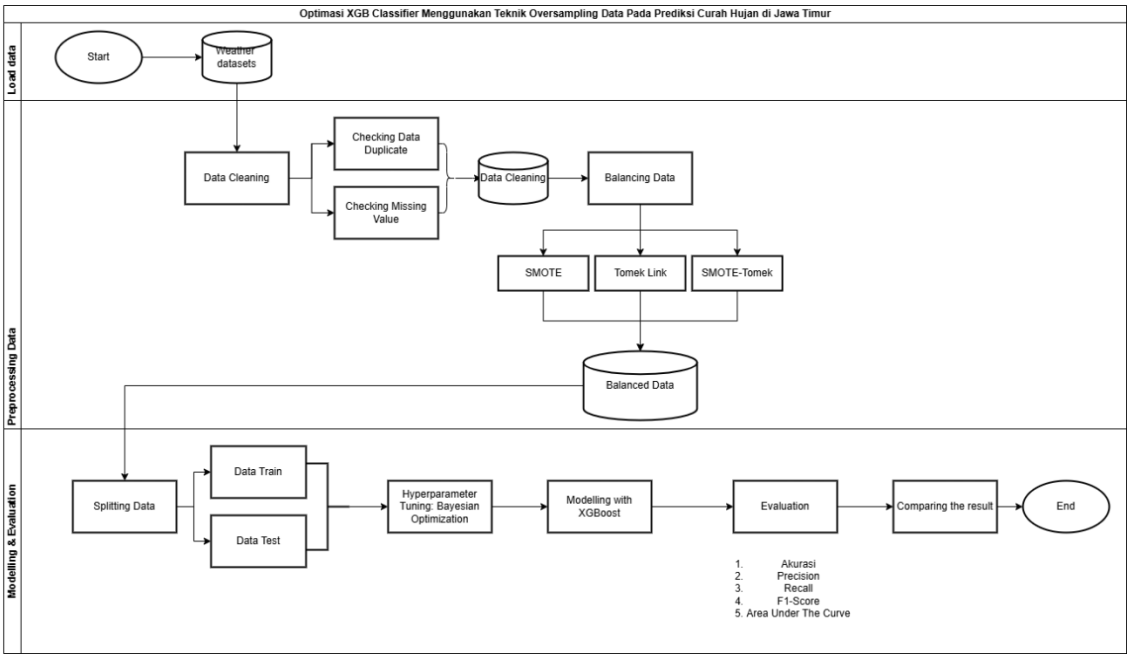
ss: Lamanya penyinaran matahari (jam)
ff_x: Kecepatan angin maksimum (m/s)
ddd_x: Arah angin saat kecepatan maksimum (°)
ff_avg: Kecepatan angin rata-rata (m/s)
ddd_car: Arah angin terbanyak (°)

3.1.3 Studi Pustaka

Studi pustaka dilakukan untuk memahami informasi yang telah ada terkait topik penelitian. Penulis mengumpulkan data dengan membaca artikel, jurnal, dan sumber lain yang relevan tentang prediksi curah hujan, algoritma XGBoost, teknik SMOTE, machine learning, serta isu meteorologi. Literatur yang digunakan dicantumkan dalam daftar pustaka.

3.1.4 Tahapan Penelitian

Tahapan penelitian terdiri dari beberapa langkah yang terstruktur untuk memastikan proses penelitian dilakukan secara sistematis. Diagram alur tahapan penelitian ditunjukkan pada gambar 3.1.



Gambar 3.1. Diagram Alur

3.1.5 Identifikasi Masalah

Langkah pertama adalah mengidentifikasi masalah utama yang menjadi dasar penelitian ini. Permasalahan utama melibatkan tantangan dalam mengklasifikasikan tingkat curah hujan berdasarkan data meteorologi yang memiliki kompleksitas tinggi. Dari identifikasi ini, tujuan dan batasan penelitian ditentukan.

3.1.6 Pengumpulan Data

Data meteorologi yang mencakup parameter seperti curah hujan, kelembapan, dan kecepatan angin diakses dari website BMKG. Data ini diolah untuk menjadi input dalam model prediksi yang menggunakan algoritma XGBoost dan teknik resampling SMOTE.

3.1.7 Preprocessing Data

Proses preprocessing melibatkan langkah-langkah seperti:

- **Penggabungan Data:** Dataset dari berbagai tahun digabung menggunakan fungsi pd.concat().
- **Transformasi Data:** Nilai kategori diubah menjadi tipe numerik menggunakan labelencoder.fit_transform.
- **Pembersihan Data:** Mengisi nilai hilang dengan median menggunakan df.fillna(df.median()).
- **Normalisasi Data:** Normalisasi dilakukan menggunakan metode Min-Max Scaling.
- **Feature Selection:** Variabel dengan korelasi rendah dihapus untuk meningkatkan efisiensi analisis.
- **Balancing Data:**

Skenario balancing data dilakukan dengan tiga pendekatan utama:

1. **Balancing Menggunakan SMOTE:** Teknik Synthetic Minority Oversampling Technique digunakan untuk meningkatkan jumlah sampel pada kelas minoritas. Teknik ini membuat data sintetis baru berdasarkan tetangga terdekat untuk memperkaya kelas minoritas.
2. **Balancing Menggunakan Tomek Link:** Digunakan untuk menghilangkan pasangan data yang tumpang tindih antara kelas mayoritas dan minoritas. Teknik ini membantu mengurangi noise pada dataset dan meningkatkan kejelasan batas antar kelas.
3. **Balancing Menggunakan SMOTE-Tomek:** Kombinasi SMOTE dan Tomek Link diterapkan untuk menciptakan dataset yang seimbang dan bersih. SMOTE digunakan untuk menambah data pada kelas minoritas, sedangkan Tomek Link menghapus noise yang dihasilkan.

3.1.8 Pembagian Dataset

Pembagian Data Dataset dibagi menjadi 80% data latih dan 20% data uji. Data uji dinormalisasi dengan parameter yang sama seperti data latih untuk menjaga konsistensi.

3.1.9 Proses Klasifikasi

XGBoost (Extreme Gradient Boosting) adalah algoritma pembelajaran mesin berbasis pohon keputusan yang dirancang untuk menangani data dalam jumlah besar dan kompleks. Algoritma ini bekerja dengan memanfaatkan teknik boosting, yaitu menggabungkan beberapa model sederhana (weak learners) untuk membentuk model yang lebih kuat. Pada penelitian ini, XGBoost digunakan untuk memprediksi curah hujan berdasarkan variabel input dari dataset meteorologi.

Langkah-langkah klasifikasi menggunakan XGBoost adalah sebagai berikut:

1. **Persiapan Data:** Data yang telah melalui tahap preprocessing dibagi menjadi data latih dan data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi performa model.

2. **Pelatihan Model:** Model XGBoost dilatih menggunakan data latih. Proses pelatihan melibatkan optimasi parameter seperti learning rate, max depth, subsample, dan n_estimators untuk mendapatkan kinerja terbaik.
3. **Prediksi:** Setelah model dilatih, data uji digunakan untuk membuat prediksi. Prediksi ini dibandingkan dengan nilai aktual untuk mengevaluasi akurasi model.
4. **Evaluasi Model:** Evaluasi dilakukan menggunakan metrik seperti akurasi, precision, recall, dan F1-score. Selain itu, matriks konfusi digunakan untuk menganalisis performa model dalam mengklasifikasikan data ke dalam kategori yang sesuai.

XGBoost dipilih karena kecepatan dan efisiensinya dalam menangani dataset besar, serta kemampuannya dalam menangani data dengan ketidakseimbangan kelas. Teknik ini memastikan bahwa model dapat menangkap pola yang kompleks dalam data meteorologi untuk prediksi yang akurat.

3.1.10 Evaluasi Model

Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Kinerja model divalidasi menggunakan teknik k-fold cross-validation dengan lima lipatan. Matriks konfusi digunakan untuk membandingkan hasil prediksi dan data aktual, memberikan analisis mendalam tentang performa model.

BAB VI

PENGUJIAN DAN ANALISA

4.1 METODE PENGUJIAN

4.2 HASIL PENGUJIAN

BAB V
PENUTUP

5.1. KESIMPULAN

5.2. SARAN PENGEMBANGAN

DAFTAR PUSTAKA

Ainaa Hanis Zuhairi, Fitri Yakub, Mas Omar, Muhammad Sharifuddin, Khamarrul Azahari Razak, & Amrul Faruq. (2024a). Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link. *Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES)*, *10*, 845–850. https://doi.org/10.5109/7323359

Ainaa Hanis Zuhairi, Fitri Yakub, Mas Omar, Muhammad Sharifuddin, Khamarrul Azahari Razak, & Amrul Faruq. (2024b). Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link. *Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES)*, *10*, 845–850. https://doi.org/10.5109/7323359

Alamri, M., & Ykhlef, M. (2024). Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data. *IEEE Access*, *12*, 14050–14060. https://doi.org/10.1109/ACCESS.2024.3357091

Anwar, M. T., Winarno, E., Hadikurniawati, W., & Novita, M. (2021). Rainfall prediction using Extreme Gradient Boosting. *Journal of Physics: Conference Series*, *1869*(1). https://doi.org/10.1088/1742-6596/1869/1/012078

Bimaprawira, A. K., & Rejeki, H. A. (2021). KETERKAITAN PERIODISITAS CURAH HUJAN DI DAERAH PESISIR DAN PEGUNUNGAN PROVINSI JAWA TIMUR DENGAN VARIABILITAS CUACA SKALA GLOBAL DAN REGIONAL Relationship of Rainfall Periodicity in Coastal dan Mountain Areas of East Java Province with Global and Regional Scale Weather Variability. In *Jurnal Sains & Teknologi Modifikasi Cuaca* (Vol. 22, Issue 2).

BMKG. (2023a). *Analisis Laju Perubahan Curah Hujan Tahunan*. https://www.bmkg.go.id/iklim/analisis-laju-perubahan-curah-hujan

BMKG. (2023b). *Distribusi Curah Hujan di Jawa Timur Tahun 2023*. Staklim-Jatim.Bmkg.Co.Id.

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Data Bencana di Tingkat Kabupaten/Kota dapat dilihat file pdf Buku Data Bencana Indonesia 2023 pada qris berikut ini. (n.d.).

Fais Putra. (2024). *Gelombang Ekuatorial Rossby Penyebab Peningkatan Curah Hujan di Jatim*. Rri.Co.Id.

Habtemariam, E. T., Kekeba, K., Martínez-Ballesteros, M., & Martínez-Álvarez, F. (2023). A Bayesian Optimization-Based LSTM Model for Wind Power Forecasting in the Adama District, Ethiopia. *Energies*, *16*(5). https://doi.org/10.3390/en16052317

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

Ke, H., Gong, S., He, J., Zhang, L., & Mo, J. (2022). A hybrid XGBoost-SMOTE model for optimization of operational air quality numerical model forecasts. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.1007530

Le, T. T. H., Shin, Y., Kim, M., & Kim, H. (2024). Towards unbalanced multiclass intrusion detection with hybrid sampling methods and ensemble classification. *Applied Soft Computing*, *157*. https://doi.org/10.1016/j.asoc.2024.111517

Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., & Yadav, R. (2024a). Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. *Environmental Earth Sciences*, *83*(6). https://doi.org/10.1007/s12665-024-11481-w

Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., & Yadav, R. (2024b). Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. *Environmental Earth Sciences*, *83*(6). https://doi.org/10.1007/s12665-024-11481-w

Odiakaose, C. C., Aghware, F. O., Okpor, M. D., Eboka, A. O., Binitie, A. P., Ojugo, A. A., Setiadi, D. R. I. M., Ibor, A. E., Ako, R. E., Geteloma, V. O., Ugbotu, E. V., & Aghaunor, T. C. (2024). Hypertension Detection via Tree-Based Stack Ensemble with SMOTE-Tomek Data Balance and XGBoost Meta-Learner. *Journal of Future Artificial Intelligence and Technologies*, *1*(3), 269–283. https://doi.org/10.62411/faith.3048-3719-43

Rahman Sya'bani, D., Hamzah, A., & Susanti, E. (2022). *KLASIFIKASI BUAH SEGAR DAN BUSUK MENGGUNAKAN ALGORITMA CONVOLUTIONAL NEURAL NETWORK DENGAN TFLITE SEBAGAI MEDIA PENERAPAN MODEL MACHINE LEARNING*.

Ruisen, L., Songyi, D., Chen, W., Peng, C., Zuodong, T., Yanmei, Y., & Shixiong, W. (2018). Bagging of Xgboost Classifiers with Random Under-sampling and Tomek Link for Noisy Label-imbalanced Data. *IOP Conference Series: Materials Science and Engineering*, *428*(1). https://doi.org/10.1088/1757-899X/428/1/012004

Sapari, A. M., Id Hadiana, A., & Umbara, F. R. (2023). *Air Quality Classification Using Extreme Gradient Boosting (XGBOOST) Algorithm ARTICLE INFORMATION ABSTRACT* (Vol. 5, Issue 2). http://innovatics.unsil.ac.id

Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors*, *22*(9). https://doi.org/10.3390/s22093246

Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent Advances in Bayesian Optimization. *ACM Computing Surveys*, *55*(13s). https://doi.org/10.1145/3582078

Wiwaha, D. D., Gafyunedi, D. A., Mahdi, Z. M., Putro, I. W., Pramudita, B. A., & Setiawan, D. P. (2024). Enhancing Rainfall Prediction Accuracy through XGBoost Model with Data Balancing Techniques. *2024 20th IEEE International Colloquium on Signal Processing and Its Applications, CSPA 2024 - Conference Proceedings*, 120–125. https://doi.org/10.1109/CSPA60979.2024.10525558

Yang, J., & Guan, J. (2022). A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm. *Information (Switzerland)*, *13*(10). https://doi.org/10.3390/info13100475