

PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATU BARA  
MENGGUNAKAN METODE *XTREME GRADIENT*  
*BOOSTING (XGBOOST)*

SKRIPSI

CYNTHIA YAPITER

201402139



PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2025

PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATU BARA  
MENGGUNAKAN METODE *XTREME GRADIENT*  
*BOOSTING (XGBOOST)*  
SKRIPSI

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah Sarjana  
Teknologi Informasi

CYNTHIA YAPITER  
201402139



PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA

MEDAN  
2025

## PERSETUJUAN

Judul : PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE *XTREME GRADIENT BOOSTING (XGBOOST)*

Kategori : SKRIPSI

Nama : CYNTHIA YAPITER

Nomor Induk Mahasiswa : 201402139

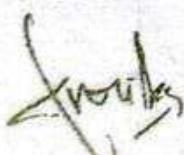
Program Studi : SARJANA (S-1) TEKNOLOGI INFORMASI

Fakultas : ILMU KOMPUTER DAN TEKNOLOGI  
INFORMASI UNIVERSITAS SUMATERA UTARA

Medan, 6 Januari 2025

Komisi Pembimbing:

Pembimbing 2,



Fahrurrozi Lubis, B.IT., M.Sc.IT.  
NIP. 198610122018052001

Pembimbing 1,



Prof. Dr. Romi Fadillah Rahmat B.Comp.Sc.,  
M.Sc  
NIP. 198603032010121004

Diketahui/disetujui oleh

Program Studi S1 Teknologi Informasi



Dedy Arisandi S.T., M.Kom.  
NIP. 197908312009121002

## **PERNYATAAN**

### **PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE XTREME GRADIENT BOOSTING (XGBOOST)**

#### **SKRIPSI**

Saya mengakui bahwa skripsi ini merupakan hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 6 Januari 2025

  
Cynthia Yapiter

201402139



## **UCAPAN TERIMA KASIH**

Penulis ingin mengucapkan rasa syukur dan terima kasih kepada Tuhan Yang Maha Esa, karena dengan rahmat dan kasih-Nya, penulis berhasil menyelesaikan skripsi dengan judul “**PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE XTREME GRADIENT BOOSTING**”. Penulisan penelitian ini sebagai salah satu persyaratan untuk meraih gelar Sarjana Komputer dalam Program Studi S1 Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.

Penulis menyadari bahwa tanpa adanya dukungan, bantuan, serta doa dari pihak-pihak terdekat selama proses penulisan skripsi ini, maka akan sangat sulit bagi penulis untuk menyelesaiannya. Oleh karena itu, penulis ingin mengungkapkan penghargaan tidak terhingga dan ucapan terima kasih kepada:

1. Diri penulis sendiri, Cynthia Yapiter, atas kesediaan melewati semua masa-masa sulit, mengorbankan banyak hal dan upaya motivasi diri sendiri dalam penulisan skripsi ini.
2. Keluarga penulis, yaitu Ayah Jardi, Ibu Polina, dan saudara-saudara penulis, Frederick Yapiter dan Filbert Alexander Yapiter, yang memberikan dorongan batin, serta doa kepada penulis.
3. Bapak Prof. Dr. Romi Fadillah Rahmat, B.Comp.Sc., M.Sc. selaku dosen pembimbing 1 atas ilmu, waktu, dan kesabaran dalam penulisan skripsi ini.
4. Bapak Fahrurrozi Lubis, B.IT., M.Sc.IT. selaku dosen pembimbing 2 atas ilmu, waktu, dan kesabaran dalam penulisan skripsi ini.
5. Bapak Dedy Arisandi, S.T., M.Kom. selaku Ketua Program Studi S1 Teknologi Informasi Universitas Sumatera Utara.
6. Bapak Ivan Jaya S.Si., M.Kom. selaku Sekretaris Program Studi S1 Teknologi Informasi Universitas Sumatera Utara
7. Seluruh Dosen Program Studi S1 Teknologi Informasi yang telah membagikan pengetahuan kepada penulis selama masa perkuliahan.
8. Seluruh Dosen Program Studi S1 Teknologi Informasi yang telah berbagi pengetahuan kepada penulis selama masa perkuliahan.

9. Sahabat seperjuangan dalam perkuliahan yang telah membersamai seluruh perjuangan selama masa perkuliahan dan penulisan skripsi, yaitu Levina Gunawan dan Davita Candra.
10. Sahabat penulis Hiero Louis, Tiffany Liuvinia, dan sahabat-sahabat lainnya yang tidak dapat disebutkan satu per satu yang telah memberikan semangat kepada penulis.
11. Semua teman komunitas GDSC, beswan Djarum 38 dan IISMA 23 yang telah berbagi pengalaman menyenangkan selama proses perkuliahan berlangsung.

Medan, 6 Januari 2025

Penulis



Cynthia Yapiter

201402139



## ABSTRAK

Malaria merupakan penyakit mematikan yang menular melalui gigitan nyamuk betina jenis *Anopheles*. Malaria termasuk salah satu penyakit menular dengan tingkat kematian yang tinggi bahkan hingga saat ini. Menurut data dari *World Health Organization*, diperkirakan sebanyak 249 juta kasus malaria telah tercatat di 85 negara endemis malaria secara global pada tahun 2022 dan Indonesia merupakan pemberi kontribusi terbesar kedua di Asia Tenggara setelah India pada tahun tersebut dengan kasus malaria positif yang tercatat sebesar 811.636 kasus. Sementara itu, berdasarkan data dari Kementerian Kesehatan Indonesia terjadi peningkatan kasus malaria di Indonesia sebesar 32,29% pada tahun 2022 dibanding tahun sebelumnya. Meskipun 22 dari 33 kabupaten di Sumatera Utara telah tercatat bebas dari malaria, kasus malaria masih sering terjadi di area dengan kondisi geografis yang sulit terjangkau, salah satunya yakni Kabupaten Batu Bara. Kabupaten Batu Bara tergolong wilayah dengan tingkat endemis sedang di Sumatera Utara. Selain faktor mobilitas, faktor iklim seperti curah hujan, suhu, kelembapan, dan lain-lain juga memiliki pengaruh besar terhadap penyebaran malaria. Dalam rangka membantu tercapainya tujuan nasional pemerintah dalam membasmi malaria pada tahun 2030, diperlukan adanya model prediksi yang mampu secara akurat memprediksi angka penyebaran malaria sehingga penanganan dapat dilakukan secara cepat dan efektif. Penelitian ini menggunakan model *eXtreme Gradient Boosting* (XGBoost) untuk memprediksi angka penyebaran malaria di Kabupaten Batu Bara, Sumatera Utara secara mingguan. Peneliti juga melakukan sintesis data dengan menggunakan TimeGAN untuk mendapatkan data tahun 2015 hingga tahun 2018 sehingga total data yang digunakan bertambah 2 kali lipat. Model yang paling optimal menghasilkan nilai *test MAE*, *RMSE*, *MSE*, dan *R<sup>2</sup>* sebesar 2.712, 3.905, 15.251 dan 0.806 berturut-turut. Parameter model yang digunakan yaitu 0.9 *colsample\_bytree*, 0.1 *learning\_rate*, 5 *max\_depth*, 100 *estimators*, 0.6 *subsample* dan rasio split data 6:4.

Kata Kunci: Prediksi, Penyebaran Malaria, *Xtreme Gradient Boosting*, *XGBoost*, *Generative Adversarial Network*, *TimeGAN*

## **PREDICTION OF MALARIA INCIDENCE IN BATU BARA REGENCY USING THE EXTREME GRADIENT BOOSTING (XGBOOST) METHOD**

### **ABSTRACT**

*Malaria is a deadly disease transmitted through mosquito bites especially that of female Anopheles mosquitoes. It remains one of the world's infectious diseases with a high mortality rate even as of today. According to data from the World Health Organization, an estimated 249 million malaria cases were recorded across 85 malaria-endemic countries globally in 2022, with Indonesia being the second-largest contributor in Southeast Asia, following India, with a total of 811,636 positive malaria cases. Meanwhile, data from Indonesia's Ministry of Health showed a 32.29% increase in malaria cases in Indonesia in 2022 compared to the previous year. Although 22 out of 33 districts in North Sumatra have been declared malaria-free, malaria cases still frequently occur in areas with difficult-to-reach geographical conditions, one of which is Batu Bara Regency. Batu Bara Regency is classified as a region with a moderate level of endemicity in North Sumatra. In addition to mobility factors, climatic factors such as temperature, humidity, rainfall, and others also significantly influence the spread of malaria. In order to support the government's national goal of eradicating malaria by 2030, there is a need for a predictive model that can accurately forecast the spread of malaria so that swift and effective interventions can be implemented. This study employs the eXtreme Gradient Boosting (XGBoost) model to predict the weekly spread of malaria in Batu Bara Regency, North Sumatra. The researcher also performed data synthesis using TimeGAN to generate data from 2015 to 2018, effectively doubling the total dataset. The most optimal model produced test results with MAE, RMSE, MSE, and R2 values of 2.712, 3.905, 15.251, and 0.806, respectively. The model parameters used include a colsample\_bytree of 0.9, learning\_rate of 0.1, max\_depth of 5, n\_estimators of 100, subsample of 0.6, and a data split ratio of 6:4.*

*Keywords:* Prediction, Malaria Incidence, Xtreme Gradient Boosting, XGBoost, Generative Adversarial Network, TimeGAN

## DAFTAR ISI

PERSETUJUAN .....	iii
PERNYATAAN .....	iv
UCAPAN TERIMA KASIH.....	v
ABSTRAK.....	vii
ABSTRACT .....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR .....	xi
DAFTAR TABEL.....	xii
BAB 1 PENDAHULUAN .....	1
1.1.    Latar Belakang .....	1
1.2.    Rumusan Masalah .....	4
1.3.    Tujuan Penelitian .....	5
1.4.    Batasan Masalah .....	5
1.5.    Manfaat Penelitian .....	5
1.6.    Metodologi Penelitian .....	6
1.7.    Sistematika Penulisan .....	7
BAB 2 LANDASAN TEORI.....	9
2.1.    Malaria .....	9
2.2.    Jenis Malaria .....	10
2.3.    Meteorologi di Kabupaten Batubara .....	11
2.4.    Hubungan Faktor Meteorologi dengan Penyakit Malaria.....	12
2.5.    Generative Adversial Networks (GAN).....	13
2.6.    Extreme Gradient Boosting (XGBoost).....	14
2.7.    Penelitian Terdahulu .....	16
2.8.    Perbedaan Penelitian .....	19
BAB 3 ANALISIS DAN PERANCANGAN SISTEM.....	21
3.1.    Data .....	21
3.2.    Arsitektur Umum .....	21

3.1.1. Datasets .....	22
3.1.2. Data Preprocessing.....	22
3.1.3. Data Splitting .....	24
3.1.4. Feature Selection on The Training Dataset.....	24
3.1.5. Model Training .....	25
3.1.6. Model Evaluation dan Comparison .....	26
3.1.7. Output .....	29
3.3. Rancangan Antarmuka Sistem.....	29
<b>BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM.....</b>	<b>35</b>
4.1. Implementasi Sistem.....	35
4.1.1. Spesifikasi perangkat .....	35
4.1.2. Implementasi perancangan antarmuka.....	35
4.2. Implementasi dan Pengujian Model Data Awal Tahun 2019-2022.....	41
4.2.1. Datasets .....	41
4.2.2. Pengujian Model .....	45
4.2.3. Evaluasi Model .....	48
4.3. Implementasi dan Pengujian Model Data Gabungan Tahun 2015-2022 ....	48
4.3.1. Datasets .....	48
4.3.2. Pengujian Model .....	54
4.3.3. Evaluasi Model .....	61
4.4. Diskusi .....	61
<b>BAB 5 KESIMPULAN DAN SARAN .....</b>	<b>63</b>
5.1. Kesimpulan .....	63
5.2. Saran .....	63
<b>DAFTAR PUSTAKA .....</b>	<b>65</b>

## DAFTAR GAMBAR

Gambar 2.1	Peta Wilayah Kabupaten Batu Bara	12
Gambar 2.2	Struktur Generative Adversarial Networks	13
Gambar 2.3	Ilustrasi XGBoost	15
Gambar 3.1	Arsitektur Umum Penelitian	21
Gambar 3.2	Rancangan tampilan <i>home page</i>	30
Gambar 3.3	Rancangan tampilan <i>testing page</i>	30
Gambar 3.4	Rancangan tampilan <i>result page</i> dengan tabel perbandingan	31
Gambar 3.5	Rancangan tampilan <i>result page</i> dengan plot perbandingan	32
Gambar 3.6	Rancangan tampilan <i>result page</i> dengan tabel evaluasi	32
Gambar 3.7	Rancangan tampilan <i>predicting page</i>	33
Gambar 3.8	Rancangan tampilan <i>predict result page</i> dengan tabel hasil	34
Gambar 3.9	Rancangan tampilan <i>result page</i> dengan plot prediksi	34
Gambar 4.1.1	Implementasi tampilan <i>home page</i>	36
Gambar 4.1.2	Implementasi tampilan <i>testing page</i>	36
Gambar 4.1.3	Implementasi tampilan <i>result page comparison table</i>	37
Gambar 4.1.4	Implementasi tampilan <i>result page comparison plot</i>	38
Gambar 4.1.5	Implementasi tampilan <i>result page evaluation metrics</i>	38
Gambar 4.1.6	Implementasi tampilan <i>predicting page</i>	39
Gambar 4.1.7	Implementasi tampilan <i>result predict page table</i>	40
Gambar 4.1.8	Implementasi tampilan <i>result predict page plot</i>	40
Gambar 4.2.1	<i>Correlation Heatmap</i> pada data awal	41
Gambar 4.2.2	<i>Grid Search</i> pada data awal	43
Gambar 4.2.3	<i>Plot</i> perbandingan nilai asli dengan prediksi <i>test data 2019-2021</i>	46
Gambar 4.2.4	<i>Plot</i> perbandingan nilai asli dengan prediksi <i>test data 2019-2021(2)</i>	47
Gambar 4.3.1	<i>Plot</i> perbandingan nilai sintesis <i>seq_len 52</i> dan <i>batch_size 64</i>	49
Gambar 4.3.2	hasil PCA dan TSNE <i>seq_len 52</i> dan <i>batch_size 64</i>	49
Gambar 4.3.3	<i>Plot</i> perbandingan nilai sintesis <i>seq_len 52</i> dan <i>batch_size 128</i>	50
Gambar 4.3.4	hasil PCA dan TSNE <i>seq_len 52</i> dan <i>batch_size 128</i>	51
Gambar 4.3.5	<i>Correlation Heatmap</i> data 2015-2022	52
Gambar 4.3.6	<i>Plot</i> perbandingan nilai asli dengan prediksi <i>test data 2015-2022(1)</i>	55
Gambar 4.3.6	<i>Plot</i> perbandingan nilai asli dengan prediksi <i>test data 2015-2022(2)</i>	56

## DAFTAR TABEL

Tabel 3.2.1 Tabel variabel indeks curah hujan kumulatif dan total kasus	25
Tabel 3.2.2 Tabel perbandingan nilai aktual dan prediksi RMSE	26
Tabel 3.2.3 Tabel perbandingan nilai aktual dan prediksi MAE	27
Tabel 3.2.4 Tabel perbandingan nilai aktual dan prediksi MSE	28
Tabel 3.2.5 Tabel perbandingan nilai aktual dan prediksi R <sup>2</sup>	28
Tabel 4.2.1 Eksperimen optimalisasi rasio pembagian data tahun 2019-2022	42
Tabel 4.2.2 Eksperimen optimalisasi parameter model pada data tahun 2019-2022	44
Tabel 4.2.3 Parameter hasil optimalisasi GridSearch dengan data <i>training</i> awal	46
Tabel 4.2.4 Parameter hasil optimalisasi GridSearch dengan data <i>testing</i> awal	46
Tabel 4.2.5 Perbandingan nilai asli dan prediksi model pada <i>test data</i> 2019-2022	47
Tabel 4.3.1 Hasil evaluasi <i>seq_len</i> 52 dan <i>batch_size</i> 64	50
Tabel 4.3.2 Hasil evaluasi <i>seq_len</i> 52 dan <i>batch_size</i> 128	51
Tabel 4.3.3 Eksperimen optimalisasi rasio pembagian data tahun 2015-2022	52
Tabel 4.3.4 Eksperimen optimalisasi parameter model pada data tahun 2015-2022	53
Tabel 4.3.5 Parameter optimalisasi GridSearch dengan data <i>training</i> gabungan	54
Tabel 4.3.5 Parameter optimalisasi GridSearch dengan data <i>testing</i> gabungan	55
Tabel 4.3.5 Perbandingan nilai asli dan prediksi model pada <i>test data</i> 2015-2022	56

## BAB 1

### PENDAHULUAN

#### 1.1. Latar Belakang

Malaria adalah penyakit mematikan yang disebabkan oleh infeksi parasit *plasmodium*. Parasit tersebut menyebar melalui gigitan nyamuk betina jenis *Anopheles* (WHO, 2023). Malaria masih termasuk salah satu penyakit menular dengan tingkat kematian yang tinggi di dunia hingga saat ini(Garrido-Cardenas et al., 2019). Menurut *World Health Organization* pada tahun 2022, terdapat sekitar 249 juta kasus malaria dari 85 negara endemis malaria. Angka tersebut membuktikan adanya peningkatan kasus malaria dari tahun sebelumnya yakni sebesar lima juta kasus jika dibandingkan dengan tahun 2019. India dan Indonesia merupakan *contributor* dari sebanyak 94% angka kematian kasus malaria di Kawasan Asia Tenggara pada tahun 2022 (World Malaria Report, 2022). Berdasarkan tabel jumlah kasus malaria pada negara-negara di Kawasan Asia Tenggara, ditampilkan bahwa Indonesia merupakan pemberi kontribusi terbesar kedua setelah India pada tahun tersebut dengan kasus malaria positif yang tercatat sebesar 811.636 kasus. (WHO, 2023)

Berdasarkan data dari Kementerian Kesehatan Indonesia, diketahui bahwa terdapat peningkatan kasus malaria di Indonesia sebesar 32,29% pada tahun 2022 dibanding tahun sebelumnya dengan jumlah kasus sebanyak 415.140 (Widi, 2022). Peningkatan ini menunjukkan bahwa upaya pencegahan malaria belum mencapai kemajuan yang signifikan dalam mencapai target poin 3.3 dari *Sustainable Development Goals* (SDGs) yang ditetapkan oleh Perserikatan Bangsa-Bangsa (PBB) terkait penyakit menular. Pada poin tersebut, target global pada penyakit malaria adalah mengurangi kasus infeksi dan jumlah kematian yang diakibatkan oleh malaria setidaknya 90% pada tahun 2030 (World Health Organization, 2022). Target-target tersebut didukung dengan pembangunan sebuah kerangka kerja komprehensif dalam memandu negara-negara bersangkutan yang disebut dengan *Global Technical Strategy* (GTS) dan telah diadopsi mulai dari tahun 2015. Namun pada kenyataannya, jumlah kasus malaria pada tahun 2022 sudah melebihi 55% dari target *Global Technical Strategy for Malaria*(GTS) di tahun 2025. Jika hal ini terus berlanjut, maka penyebaran malaria secara global

diperkirakan akan meningkat hingga 89% pada tahun 2030. Selain peningkatan jumlah kasus yang melebihi target, tingkat kematian akibat malaria juga telah melebihi target GTS sebesar 53%, yang dimana berpeluang untuk meningkat hingga sebesar 88% di luar jalur pada tahun 2030. (WHO, 2023).

Indonesia merupakan satu-satunya negara di wilayah Asia Tenggara yang tidak mengalami penurunan kasus sejak tahun 2015 dan juga mengalami peningkatan angka kematian akibat penyakit malaria. Hal ini tentunya tidak memenuhi target GTS (WHO, 2023). Sehubungan dengan target Indonesia dalam eliminasi malaria pada tahun 2030, Kementerian Kesehatan telah menentukan 5 provinsi sebagai target eliminasi yang dimana salah satunya merupakan Provinsi Sumatera Utara (Rokom, 2022). Meskipun 21 dari 33 kabupaten/kota di Sumatera Utara telah dinyatakan bebas malaria, masih terdapat kerentanan akan terjadinya wabah malaria di daerah-daerah yang masih sulit dijangkau baik dari segi geografis yang kurang mendukung maupun segi mobilitas tinggi. Hal ini yang menjadi tantangan dalam upaya mengeliminasi malaria secara penuh dari Provinsi Sumatera Utara (Fahmi et al., 2022). Oleh karena itu, diperlukan sebuah model prediksi yang mampu secara akurat memprediksi angka kejadian malaria sehingga memungkinkan tindakan pencegahan dilakukan secara tepat waktu dan efektif.

Menurut laporan yang disusun oleh Dinas Kesehatan Sumatera Utara pada tahun 2022, terdapat sebanyak 5.133 kasus malaria yang tercatat di wilayah tersebut. Pada bulan September 2022, tiga daerah di Sumatera Utara dikategorikan sebagai daerah endemis sedang. Ketiga daerah tersebut mencakup Asahan, Batubara, dan Labuhanbatu Utara (Kementerian Kesehatan, 2022). Faktor-faktor iklim seperti curah hujan, suhu, dan kelembaban relatif telah terbukti memengaruhi perkembangan morfologi nyamuk. Suhu lingkungan yang tinggi dapat mempercepat perkembangan parasit malaria, yang berpotensi meningkatkan insiden malaria (Fischer et al., 2020). Namun, dampak dari faktor iklim terhadap kasus malaria dapat berbeda-beda antar negara dan bahkan di dalam negara yang sama. Penelitian menunjukkan bahwa suhu memiliki hubungan linier yang signifikan dengan variabilitas malaria di beberapa negara yang telah diteliti. Selain suhu, curah hujan dan radiasi permukaan juga telah terbukti memengaruhi variasi dalam kasus malaria (Nkiruka et al., 2021).

Pada tahun 2023, terdapat penelitian yang dilakukan untuk membandingkan metode pembelajaran mesin dalam memprediksi malaria dengan menggunakan dataset klinis dari Kaggle, dan gambar X-ray yang diambil dengan jumlah pasien sebanyak 1079 dan atribut yang digunakan berjumlah 23. Atribut-atribut tersebut meliputi umur, gender, gejala sakit kepala, gejala kesulitan bernapas, mata yang membengkak dan sebagainya. Metode-metode yang dipakai dalam penelitian ini adalah *Gaussian NB*, *XGBoost*, *Bagging Classifier*, *Logistic Regression*, *Decision Tree Classifier*, *Gradient Boosting Classifier*, *Hist Gradient Boosting Classifier*, *Random Forest Classifier*, *Extra Trees Classifier*, *LGBM Classifier*, *Ada Boost Classifier*, *SGD Classifier*, dan *K Nearest Neighbors (KNN) Classifier*. Dari semua metode tersebut, model *Gaussian NB* memiliki tingkat performa akurasi yang tertinggi yakni sebesar 97,66% dan nilai AUC sebesar 98%. Diikuti oleh performa model *Logistic Regression* dan *XGBoost Classifier* dengan tingkat akurasi 97.42% dan 97.31% dan nilai AUC sebesar 97.7% dan 97.73% secara berurutan (Islam et al., 2023).

Selain itu, terdapat juga penelitian yang dilakukan untuk memprediksi malaria dengan menerapkan teknik *Explainable Artificial Intelligence* (XAI), yaitu *Shapley Additive Explanation* (SHAP) dan *Local Interpretable Model-agnostic Explanation* (LIME), untuk memberikan hasil model yang lebih mudah dimengerti. Berbagai model, termasuk *Xtreme Gradient Boosting*, *Decision Tree*, *Logistic Regression* (LR), *AdaBoost*, *Support Vector Machine* (SVM), *K-means*, *K-Nearest Neighbor*, *Random Forest*, *Naive Bayes*, dan *Explainable Boosting Machines* (EBMs) digunakan dalam penelitian ini. Hasil studi menunjukkan bahwa Random Forest dan Explainable Boosting Machines mencapai akurasi tertinggi sebesar 84%. Model XGBoost juga mencapai tingkat akurasi sebesar 83%. Meskipun model XGBoost disini digunakan untuk tugas classification yakni secara spesifik memprediksi apakah pasien terinfeksi malaria atau tidak, penelitian ini menunjukkan bahwa XGBoost memiliki kinerja yang bagus dalam memprediksi malaria (Rajab et al. 2023).

Terdapat penelitian sebelumnya yang telah menerapkan algoritma XGBoost dalam memprediksi malaria di wilayah Prancis dan Thailand pada tahun 2022. Data yang digunakan untuk melatih model penelitian ini berupa data meteorologi saja sehingga diperlukan tambahan fitur input yang efektif untuk mencoba meningkatkan akurasi model prediksi (Methiyothin et al., 2022). Penelitian terdahulu telah menunjukkan bahwa meningkatnya kasus malaria selain berhubungan dengan factor meteorologis

juga berhubungan erat dengan faktor iklim, jenis pekerjaan, dan juga dapat dipengaruhi oleh kasus impor.

Penelitian oleh Latief, A.M et al (2020) menyatakan bahwa XGBoost dapat secara akurat memprediksi klasifikasi tumor pada data *hepatocellular carcinoma gene expression* yang memiliki *missing value* sebesar 20% tanpa amputasi data dan penggunaan *grid search* pada proses *hyperparameter tuning* model dapat meningkatkan performa evaluasi model tersebut. Terdapat juga penelitian oleh Zou, M et al.(2022) yang mengajukan penggunaan model XGBoost yang dioptimalisasi pada dataset kecil dalam memprediksi relative density of SLMed Ti-6Al-4V parts. Optimalisasi yang dimaksud yakni dengan penerapan metode GridsearchCV. Hasil prediksi model XGBoost ini lebih superior dibandingkan dengan model SVR dan DNN. Hasil kedua penelitian ini menunjukkan bahwa XGBoost merupakan model machine learning yang dapat diimplementasikan pada bidang Kesehatan yang cenderung memiliki masalah pengumpulan data yang susah dan terbatas jumlahnya.

Sehubungan dengan fakta bahwa penelitian malaria di Indonesia masih sangat terbatas terutama pada wilayah tertentu yang susah dijangkau secara geografis, penulis mengusulkan untuk melakukan penelitian dengan judul “Prediksi Penyebaran Malaria di Kabupaten Batubara Menggunakan Metode Xtreme Gradient Boosting (XGBoost)”. Model XGBoost yang digunakan pada penelitian ini dioptimalisasikan dengan penggunaan metode *Gridsearch* pada proses *hyperparameter tuning* model. Terkait masalah jumlah data yang sangat terbatas, penelitian ini juga menerapkan generasi data sintesis dalam upaya menghasilkan model prediksi yang optimal. Adapun penggunaan optimalisasi model pada penelitian ini diharapkan dapat menunjang penelitian selanjutnya dalam bidang Kesehatan maupun bidang lainnya yang menghadapi masalah pengumpulan data yang serupa.

## 1.2. Rumusan Masalah

Sehubungan dengan tujuan nasional pemerintah tahun 2030 yakni membasmi penyakit malaria di Indonesia, salah satu tindakan yang dapat dilakukan adalah antisipasi. Oleh karena itu, diperlukan adanya model prediksi untuk meramalkan kejadian kasus positif malaria. Penggunaan model XGBoost masih belum digunakan pada prediksi kejadian malaria di wilayah Indonesia yang cenderung susah untuk

didapatkan datanya. Penggunaan metode ini diharapkan dapat menghasilkan model prediksi dengan tingkat akurasi dan performa yang bagus.

### **1.3. Tujuan Penelitian**

Tujuan penelitian ini adalah untuk memprediksi penyebaran penyakit malaria di Kabupaten Batu Bara dengan menggunakan metode *Xtreme Gradient Boosting* (XGBoost) dalam bidang regression berdasarkan data iklim wilayah tersebut.

### **1.4. Batasan Masalah**

Berikut adalah batasan-batasan masalah jelas yang telah ditetapkan pada penelitian ini:

1. Penelitian berfokus pada wilayah Provinsi Sumatera Utara, tepatnya di Kabupaten Batu Bara.
2. Data meteorologi dan hubungannya dengan jumlah kasus positif digunakan untuk memprediksi penyebaran malaria.
3. Variabel-variabel iklim yang digunakan yakni suhu maksimum, curah hujan rata-rata harian, curah hujan kumulatif, indeks kelembapan, indeks tingkat genangan air ataupun banjir, indeks kepadatan penduduk, dan total kasus positif yang diperoleh secara mingguan

### **1.5. Manfaat Penelitian**

Berikut adalah beberapa manfaat penelitian ini:

1. Memberikan dukungan kepada Dinas Kesehatan Provinsi Sumatera Utara dalam memprediksi jumlah kasus positif penyakit malaria pada Kabupaten BatuBara.
2. Mengkontribusikan pada upaya menjadikan Kabupaten Batu Bara di provinsi Sumatera Utara sebagai wilayah yang terbebas dari penyakit malaria
3. Menjadi rujukan penelitian selanjutnya dalam menggunakan metode *Xtreme Gradient Boosting*.

## 1.6. Metodologi Penelitian

Tahapan-tahapan yang akan dilakukan pada penelitian ini adalah sebagai berikut:

### 1. Studi Literatur

Studi literatur yang merupakan tahap pertama dalam penelitian ini dilakukan penulis dengan mengumpulkan sumber referensi dari buku, jurnal, *conference paper*, artikel, *report*, skripsi, dan sumber-sumber lainnya yang membahas mengenai malaria dan metode *Xtreme Gradient Boosting*.

### 2. Analisis Permasalahan

Pada tahapan ini, penulis akan melakukan analisa permasalahan penyebaran kasus malaria serta mencari tahu faktor-faktor penting yang mempengaruhinya, dengan merujuk pada berbagai sumber referensi yang telah dikumpulkan sebelumnya. Hal ini bertujuan untuk mendapatkan pemahaman secara mendalam mengenai penyakit malaria dan konsep algoritma *Xtreme Gradient Boosting* (XGBoost).

### 3. Perancangan Sistem

Dalam tahap ini, penulis merancang sistem yakni berupa arsitektur umum untuk memprediksi penyebaran kasus malaria berdasarkan hasil dari studi literatur dan analis permasalahan yang telah dilakukan.

### 4. Implementasi

Setelah sistem dibuat dan dievaluasi, maka pada tahap ini sistem yang telah dirancang akan diimplementasikan untuk mendapatkan tujuan yang diharapkan. Penulis akan membangun model yang menerapkan *Xtreme Gradient Boosting* regression dengan menggunakan bahasa pemrograman *Python* beserta *library-library* nya.

### 5. Pengujian

Tahap ini merupakan momen pengujian atau evaluasi pada sistem yang telah dirancang terhadap data saat ini dengan tujuan mengetahui kualitas dari pengaplikasian XGBoost dalam model prediksi.

### 6. Dokumentasi dan Penyusunan Laporan

Pada tahap ini, penulis akan melaksanakan dokumentasi dan penyusunan laporan. Laporan dan dokumentasi tersebut akan berisikan proses penelitian dari tahap awal hingga hasil akhir.

## 1.7. Sistematika Penulisan

Pada skripsi ini terdapat lima bagian utama dalam sistematika penulisan, yaitu:

### BAB 1 : Pendahuluan

Pada bab ini, terdapat gambaran umum mengenai latar belakang masalah yang menjadi dasar penelitian, rumusan masalah yang ingin diselesaikan, tujuan penelitian yang hendak dicapai, serta manfaat yang diharapkan dari penelitian tersebut. Pendahuluan ini memberikan konteks dan arah penelitian, membantu pembaca memahami pentingnya topik yang dibahas serta alasan dilakukannya penelitian tersebut.

### BAB 2 : Landasan Teori

Bab ini berisi penjelasan mengenai konsep, teori, dan kajian pustaka yang mendasari penelitian. Bagian ini mencakup teori-teori relevan yang dijadikan dasar dalam analisis serta referensi dari penelitian terdahulu yang berkaitan dengan topik yang diangkat. Adapun teori-teori yang digunakan yakni pengertian malaria, jenis malaria, meteorologi di Kabupaten Batu bara, hubungan faktor meteorologi dengan penyakit malaria, *Generative Adversarial Network* (GAN), *Xtreme Gradient Boosting* (XGBoost). Terdapat juga rangkuman beberapa penelitian-penelitian relevan terdahulu. Bab ini bertujuan untuk memberikan kerangka berpikir yang jelas dan memperkuat argumen penelitian, sehingga hasil penelitian dapat dipertanggungjawabkan secara ilmiah. Penjelasan dalam bab ini juga digunakan untuk mendefinisikan istilah-istilah penting yang digunakan dalam penelitian.

### BAB 3 : Analisis dan Perancangan Sistem

Pada bab ini, tahapan-tahapan pada perancangan sistem dijelaskan secara rinci. Dataset yang digunakan, *data preprocessing*, *model training*, model *output* hingga rancangan *interface* yang akan digunakan.

### BAB 4 : Impementasi dan Pengujian Sistem

Sesuai judul bab ini, penjelasan implementasi dan pengujian sistem dilakukan secara rinci disini. Proses penerapan arsitektur sistem, pembangunan *interface* sesuai

rancangan serta spesifikasi *hardware* dan *software* yang digunakan dalam implementasi tertulis disini. Hasil evaluasi dan pengujian juga termasuk.

#### BAB 5 : Kesimpulan dan Saran

Peneliti meringkas hasil penelitian yang telah dilakukan, menyampaikan kesimpulan utama berdasarkan analisis data serta jawaban atas rumusan masalah. Selain itu, bagian ini juga memuat saran-saran yang diberikan peneliti terkait implementasi hasil penelitian, serta rekomendasi untuk penelitian lebih lanjut agar dapat mengembangkan atau memperbaiki aspek-aspek yang belum terjangkau dalam penelitian ini.



## **BAB 2**

### **LANDASAN TEORI**

#### **2.1.Malaria**

Malaria adalah penyakit mematikan yang disebabkan oleh infeksi parasit *plasmodium*. Parasit tersebut menyebar melalui gigitan nyamuk betina jenis *Anopheles*. Pada tahun 2019, WHO telah menyatakan malaria sebagai ancaman terberbahaya ke-9 bagi kesehatan dunia. Adapun beberapa jenis parasit malaria yang menginfeksi manusia, yaitu *Plasmodium falciparum*, *P. vivax*, *P. ovale*, dan *P. malariae*. Penularan parasit ini adalah dari nyamuk yang berkembang biak di banyak daerah tropis dan subtropis. Selain itu, terdapat parasit *P. knowlesi*, sejenis malaria yang secara alami menginfeksi kera di Asia Tenggara. Penyakit ini bisa menjangkit manusia dan menyebabkan penularan penyakit dari hewan ke manusia (malaria zoonosis). Akan tetapi, penyakit malaria tidak menular dari manusia ke manusia lainnya. Penyakit ini menular melalui jarum suntik, transfuse darah dan juga infeksi turunan dari ibu hamil ke bayi yang belum lahir (Fadli, R. 2023). Diantara semua parasit, *Plasmodium falciparum* dan *P. vivax* merupakan ancaman terbesar. Hal ini dikarenakan *P. falciparum* menyebabkan infeksi parah yang dimana jika gejala tersebut dibiarkan dapat menyebabkan masalah pada otak dan sistem saraf. Gangguan juga dapat menyebabkan kelumpuhan dan kejang-kejang parah jika tak ditangani dengan langkah yang tepat. Parasit ini terdapat di Afrika, Asia Tenggara, dan Amerika Selatan. Sedangkan *P. vivax* adalah jenis malaria yang umum dan paling banyak tersebar di seluruh dunia. Meskipun tidak berakibat fatal seperti *P. falciparum* tetapi bisa sangat melemahkan kekebalan tubuh (WHO, 2023).

Posisi geografis Indonesia yang merupakan negara tropis dan berada di kawasan Asia Tenggara menjadikan Indonesia salah satu negara yang rentan akan adanya penyakit malaria. Pada tahun 2022, WHO melaporkan adanya 249 juta kasus malaria secara global dengan 1,2 juta kasus diantaranya terjadi di Indonesia. Berdasarkan tabel jumlah kasus malaria pada negara-negara di Kawasan Asia Tenggara, ditampilkan bahwa Indonesia merupakan pemberi kontribusi terbesar kedua setelah India pada tahun 2023 dengan kasus malaria positif yang tercatat sebesar 811.636 kasus. (WHO, 2023) Adapun beberapa pencegahan terjadinya kasus malaria yang dapat dilakukan yakni kontrol vektor dengan insektisida atau penyemprotan residu, vaksin, kemoterapi pencegahan, dan lain-lain (WHO, 2023).

## 2.2. Jenis Malaria

Dengan jenis parasit yang berbeda, maka gejala dan penanganannya juga beragam. Terdapat beberapa jenis parasit malaria yang dapat menginfeksi manusia, yaitu

### 1. *Plasmodium Falciparum*

*Plasmodium Falciparum* merupakan parasit penyebab malaria tropika yang dapat ditemukan di Afrika, Asia Tenggara, dan Amerika Selatan (Fadli, 2023). Malaria dengan jenis ini merupakan malaria yang berat dan berbahaya karena berkembang biak dengan cepat didalam darah manusia dan jika tidak ditangani segera akan menyebabkan masalah pada otak dan sistem saraf. Gangguan juga dapat menyebabkan kelumpuhan dan kejang-kejang parah jika tidak ditangani dengan langkah yang tepat. Adapun gejala-gejala yang ditimbulkan yakni perasaan mual, kelelahan, nyeri badan, pembesaran limpa, nyeri pada perut, otot dan persendian, demam, sakit kepala, anemia, kebingungan dan kejang-kejang (Fadli, 2023).

### 2. *Plasmodium Vivax*

*Plasmodium vivax* merupakan parasit penyebab malaria tertiana. Malaria jenis ini merupakan jenis yang paling umum dan paling banyak tersebar di seluruh dunia. Jenis ini telah menyebar hampir ke seluruh pulau di Indonesia dan merupakan malaria yang paling banyak di temukan di daerah-daerah dari negara Indonesia dan secara klinis baik dari segi gejala maupun dampak jauh lebih ringan. Jenis ini jarang mencetak angka kematian jika dibandingkan dengan *Plasmodium falciparum* (Avichena and Anggriyani, 2023). Malaria tertiana memiliki gejala meliputi diare, rasa lelah yang parah, demam dan menggigil (Fadli, 2023).

### 3. *Plasmodium Ovale*

*Plasmodium ovale* merupakan parasit penyebab malaria *ovale*. Parasit ini biasanya ditemukan di negara-negara barat seperti Afrika, Ghana, Nigeria, dan Liberia (Fadli, 2023). Malaria *ovale* merupakan jenis malaria yang relatif ringan dan dapat sembuh dengan sendirinya (Irwan, 2016). Meskipun begitu malaria jenis ini cenderung kambuh karena parasitnya bisa menetap di hati hingga empat tahun. Kekambuhan dapat terjadi kapan saja selama periode tersebut (Fadli, 2023).

#### 4. *Plasmodium Malariae*

*Plasmodium malariae* merupakan parasit penyebab malaria *quartana* yang menyebabkan serangan demam setiap empat hari sekali. Malaria dengan jenis ini dapat terjadi di dataran rendah maupun dataran tinggi di daerah tropis (Avichena and Anggriyani, 2023). Jenis ini jarang terjadi, adapun prevalensi kasusnya kurang dari satu persen dari total keseluruhan. Gejala pada kasus ini yakni demam dan menggil yang tak kunjung sembuh (Fadli, 2023).

#### 5. *Plasmodium Knowlesi*

*Plasmodium knowlesi* merupakan parasit penyebab malaria pada primata yang bisa menular ke manusia. Parasit ini umum ditemukan di alam pada ekor kera dan ekor babi. Gejala yang ditimbulkan jugalah mirip dengan gejala infeksi *Plasmodium malariae*. Akan tetapi, gejala pada kasus ini dapat berkembang menjadi penyakit serius dengan sangat cepat (Fadli, 2023).

### 2.3. Meteorologi di Kabupaten Batu Bara

Kabupaten Batu Bara merupakan bagian dari wilayah Provinsi Sumatera Utara yang dimana merupakan salah satu provinsi di pulau Sumatera yang terletak di bagian Utara pulau. Provinsi ini terletak pada 10 – 40 lintang utara dan 980 - 1000 bujur timur dengan luas wilayah provinsi mencapai 71.680,68 km<sup>2</sup> atau 3,72% dari luas wilayah Republik Indonesia (Badan Pengawasan Keuangan dan Pembangunan, 2024).

Wilayah Kabupaten Batu Bara sendiri memiliki luas sebesar 904,96 km<sup>2</sup> dengan total 12 kecamatan, 10 kelurahan dan 141 jumlah desa (Badan Pusat Statistik Kabupaten Batu Bara, 2024). Data dari Dinas Kependudukan dan Pencatatan Sipil per 30 Juni 2024 menunjukkan bahwa jumlah populasi di Kabupaten Batu Bara mencapai 465.286 dengan populasi laki-laki sebanyak 234.817 dan perempuan sebanyak 230.469 (Kementerian Dalam Negeri, 2024). Kabupaten Batu Bara terletak di bagian timur Provinsi Sumatera Utara. Kabupaten ini berbatasan dengan Selat Malaka di sebelah timur, dan secara geografis berada di pesisir timur Sumatera. Terletak pada bagian timur yang merupakan dataran rendah menandakan bahwa wilayah ini memiliki tingkat kelembapan dan curah hujan yang tinggi(Sianturi, 2023). Dari segi iklim, Kabupaten

Batu Bara termasuk tropis dikarenakan rata-rata indeks curah hujan bulanan berada di angka 100 mm/tahun (Badan Pusat Statistik Kabupaten Batu Bara, 2024).



Gambar 2.1 Peta Wilayah Kabupaten Batu Bara (Fitri, H. et al., 2023)

#### 2.4. Hubungan Faktor Meteorologi dengan Penyakit Malaria

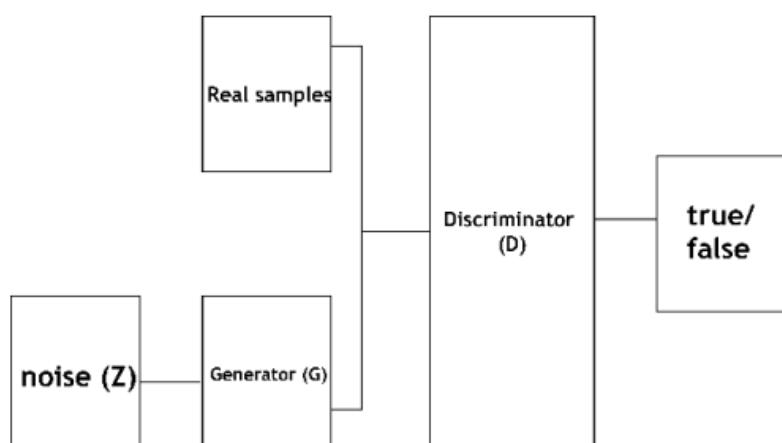
Iklim global merupakan salah satu faktor yang mempengaruhi terjadinya kasus malaria. Hal ini dikarenakan perubahan iklim dapat berpengaruh terhadap perkembangbiakan vektor penyakit seperti nyamuk *Anopheles*, *Aedes* dan lainnya, sehingga berpotensi meningkatkan kejadian berbagai penyakit. Salah satunya adalah penyakit seperti malaria dan demam berdarah *dengue* yang ditularkan melalui nyamuk. Perubahan iklim dari musim panas ke musim hujan, dianggap sebagai masa berbahaya disebabkan kondisi ini memungkinkan terjadinya penyebaran nyamuk ke wilayah-wilayah baru. Mirip dengan kejadian Badai El Nino pada tahun 1997, masa dimana nyamuk bermigrasi ke daerah dataran tinggi di Papua. Peningkatan suhu udara mengakibatkan perubahan pola vegetasi, menyebabkan serangga seperti nyamuk beradaptasi dengan perubahan tersebut dan dapat bertahan hidup di wilayah yang sebelumnya terlalu dingin untuk pembiakan (Apriliana, 2017). Faktor selain iklim seperti faktor geografis juga memiliki pengaruh dalam perkembangan kasus malaria.

Adapun kondisi dimana banyak terdapat rawa-rawa, hutan sagu, perbukitan, dan hutan merupakan lingkungan yang sangat ideal untuk habitat parasit *Plasmodium* (Lie, 2016).

Salah satu faktor pendorong vegetasi habitat nyamuk adalah curah hujan. Curah hujan dapat terbagi menjadi intensitas kecil, sedang dan tinggi. Curah hujan dengan intensitas sedang dapat meningkatkan kepadatan populasi nyamuk *Anopheles*. Hal ini mengakibatkan meningginya intensitas nyamuk *Anopheles* menghisap darah manusia yang dimana berpeluang terjadinya peningkatan kasus malaria di masyarakat (Suwito et al., 2010). Berkebalikan dengan intensitas sedang maupun tinggi, curah hujan yang rendah atau dengan jarang adanya hujan dapat merusak habitat nyamuk disebabkan oleh adanya paparan radiasi sinar matahari. Hal ini menyebabkan penurunan angka kasus malaria (Kazwaini et al., 2014). Adapun angka curah hujan minimum yang diperlukan untuk reproduksi nyamuk *Anopheles* adalah sebesar 1,5 mm per hari. Peningkatan kepadatan populasi nyamuk *Anopheles* terjadi jika angka curah hujan mencapai 150 mm per bulan (Kusuma & Widjyanto, 2016).

## 2.5. Generative Adversarial Networks (GAN)

*Generative Adversarial Networks* (GAN) merupakan model generative yang dapat menghasilkan sampel data yang sesuai dengan karakteristik data asli menggunakan basis *Neural Network*. Terdapat 2 *Neural Network* yang digunakan pada GAN yakni *Generator* dan *Discriminator*. *Generator* bertugas untuk menghasilkan data sampel yang susah dibedakan atau sesuai ketentuan dari data asli dengan mengambil *noise* acak  $z \in \mathbb{R}^r$  dan *Discriminator* bertugas untuk membedakan sampel data tersebut dari data asli (Goodfellow et al., 2014).



Gambar 2.2 Struktur *Generative Adversarial Networks* (Secada Purba, 2022)

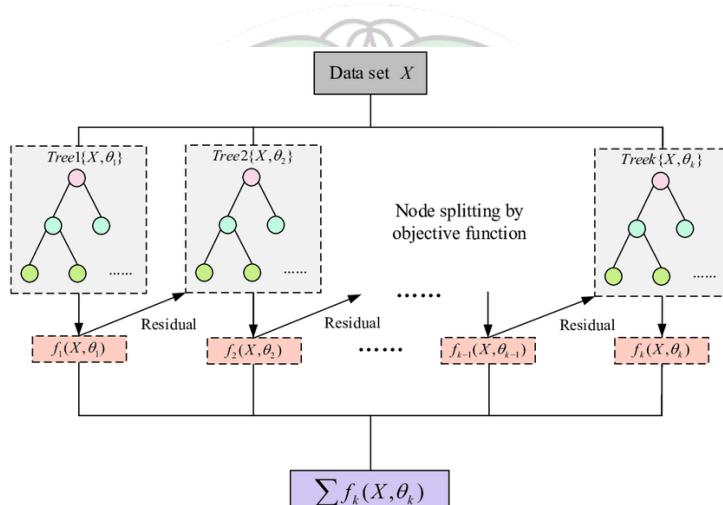
Struktur umum dari *Generative Adversarial Networks* terinspirasi dari konsep zero-sum game pada permainan dua orang. Konsep zero-sum game adalah kondisi dimana jumlah keuntungan dan kerugian dua orang adalah nol atau singkatnya keuntungan satu pihak merupakan kerugian dari pihak lainnya. Oleh karena itulah, GAN menggunakan dua *neural network* yang berbeda. Cara kerja *Generative Adversarial Networks* (GANs) yakni dengan mempertemukan *Generator* dan *Discriminator*, kemudian melawan satu sama lain hingga menghasilkan hasil yang sesuai. Dalam Langkah pertama, *Generator* menciptakan sampel palsu dari *noise* acak, sementara *Discriminator* mengecek perbedaan antara data nyata dan palsu. Melalui pelatihan yang berulang, *Generator* belajar untuk menghasilkan sampel yang lebih realistik dan juga sekaligus meningkatkan kemampuan *Discriminator* dalam mengevaluasi data. Hal tersebut menghasilkan sampel berkualitas tinggi yang sangat menyerupai sifat data nyata.

Pemodelan data *time series* merupakan tantangan bagi GAN yang umumnya digunakan untuk menghasilkan data non-sekuensial karena perlu memperhatikan dan mempertahankan dinamika temporal yang unik pada deret waktu. Oleh karena keperluan penelitian yang memperhatikan kepentingan deret waktu pada data yang akan diteliti, maka *Time-series Generative Adversarial Network* (TimeGAN) akan diutilisasi. TimeGAN merupakan GAN yang dirancang secara khusus untuk memodelkan data *time series*. Model ini terdiri dari empat komponen utama yakni fungsi *embedding*, fungsi *recovery*, *sequence generator*, dan *sequence discriminator*. Keunggulan utama dari TimeGAN adalah pendekatannya yang menggabungkan komponen *autoencoding* (*embedding* dan *recovery*) dengan komponen *adversarial* (*generator* dan *discriminator*) dalam satu proses pelatihan terpadu. Jaringan *embedding* berperan membentuk ruang laten yang merepresentasikan data asli. Komponen *adversarial* beroperasi di ruang laten ini untuk menghasilkan data sintetik yang realistik. Kemudian TimeGAN menyelaraskan dinamika laten antara data asli dan data sintetik melalui *supervised loss*, sehingga model tidak hanya belajar menghasilkan representasi data yang serupa secara statistik tetapi juga mampu mereplikasi hubungan temporal yang ada dalam data.

## 2.6. Extreme Gradient Boosting (XGBoost)

*Gradient Boosting* atau yang dikenal sebagai *eXtreme Gradient Boosting* diperkenalkan pertama kalinya oleh Dr. Tianqi Chen dari *University of Washington*

pada tahun 2014. *Gradient Boosting* merupakan model pembelajaran mesin yang menerapkan metode ensemble learning dalam pembangunan model prediksi dengan menggabungkan beberapa model lemah yang pada setiap prosesnya dirancang untuk menjadi lebih baik secara berurutan. *Extreme Gradient Boosting* merupakan algoritma implementasi dari *Gradient Boosting* yang sangat efisien dan dapat dikembangkan skala penggunaannya. Hal ini dikarenakan XGBoost memiliki fitur-fitur seperti penanganan data yang jarang (sparse data) dan penggunaan weighted quantile sketch untuk pembelajaran pohon yang lebih cepat dan efisien. Adapun beberapa keunggulan dari XGBoost yakni dalam menangani volume data yang besar, mampu melakukan komputasi out-of-core saat memori habis, dan mendukung parallelization untuk meningkatkan kecepatan dalam melatih model (Chen & Guestrin, 2016).



Gambar 2.3 Ilustrasi XGBoost (Sumber: Guo et al., 2020)

Gambar 2.3 menunjukkan proses awal algoritma XGBoost, dimulai dengan menghitung residu yang kemudian diikuti dengan pembelahan node berdasarkan fungsi tujuan. Setiap set data kemudian dijalankan melalui serangkaian pohon dalam ensemble, di mana masing-masing pohon berperan dalam membentuk nilai prediksi akhir. Algoritma secara berulang melewati pohon-pohon ini untuk menyempurnakan prediksi dan meningkatkan akurasi keseluruhan model. XGBoost terkenal akan kemampuannya menghadapi data yang kompleks dan menghasilkan prediksi yang akurat dengan memanfaatkan ensemble dari pohon keputusan (Guo et al., 2020). XGBoost dapat dijelaskan melalui persamaan berikut:

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad \dots(1)$$

Keterangan:

$\text{Obj}$  = fungsi objektif yang hendak dioptimalkan,

$l(y_i, \hat{y}_i)$  = fungsi kerugian yang mengukur kesalahan prediksi,

$\Omega(f_k)$  = fungsi regularisasi yang mencegah *overfitting* pada model.

Adapun *Pseudocode* proses XGboost (Demir, S., & Sahin, E. K., 2023) sebagai berikut :

For setiap boosting iteration( $t$ ) hingga  $M$  do:

    Compute first-order gradient:  $g_i = \partial(\text{Loss})/\partial(\hat{y}^{(t-1)})$

    Compute second-order gradient:  $h_i = \partial^2(\text{Loss})/\partial(\hat{y}^{(t-1)})^2$

    Construct a new tree  $f_m(x)$ :

        Solve for the optimal tree using:

$$\text{obj}^{(m)} = -1/2 * \sum (G_j^2 / (H_j + \lambda)) + \gamma M$$

        Where  $G_j = \sum g_i$  and  $H_j = \sum h_i$  for node  $j$

        Add the best tree  $f_m(x)$  to the model:  $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \eta * f_m(x)$

End For

Output regression tree kuat yang merupakan kombinasi *weak learners*.

## 2.7. Penelitian Terdahulu

Prediksi kejadian malaria dengan menggunakan data klinis kejadian malaria dan data lingkungan sudah banyak dilakukan sebelumnya. (Mbunge et al., 2022) mendemonstrasikan potensi *machine learning* dalam memprediksi penyebaran wabah malaria dengan menggunakan data kejadian malaria sebelumnya dan data lingkungan. Data lingkungan yang digunakan meliputi faktor-faktor yang berhubungan dengan resiko terjadinya malaria. Data kejadian yang terjadi sebelumnya berasal distrik Buhera, daerah yang dikenal akan kejadian malaria yang tinggi selama puncak musim hujan. Data diambil dari *Durban University of Technology* (DUT) dan MRCZ. Data dikumpulkan dari tahun 2015 hingga 2020. Data lingkungan dikumpulkan dari *database MODIS* dan *CHIRPS* dengan variabel curah hujan, *Day Land Surface Temperature* (LST), *Night Land Surface Temperature* (LST) dan indeks vegetasi

(NDVI). Hasil dari penelitian ini adalah performa model logistic regression lebih unggul dibandingkan model lainnya dengan tingkat akurasi 83%, presisi 82%, dan F1 score sebesar 90%.

Prediksi malaria sebelum dilakukan uji klinis menggunakan model berbasis *Machine Learning* dan *Deep Learning* dilakukan oleh (Islam *et al.*, 2023). Metode yang digunakan meliputi penggunaan siklus hidup mesin yang mencakup penyortiran fitur, pengumpulan data, *data preprocessing*, konstruksi model, validasi silang, *hyperparameter setting* dan evaluasi model. Model pembelajaran mesin yang digunakan *Gaussian NB*, *Bagging*, *Random Forest Classifier*, dan *XGBoost*. Total data sebanyak 1079 dengan 23 atribut, dikumpulkan dari *google form*, media sosial, mahasiswa *Daffodil University*, dan beberapa sumber lainnya. Hasil penelitian ini menunjukkan bahwa model *Gaussian NB* memberikan hasil dengan akurasi tertinggi yakni 97.66% diikuti oleh model *XGB classifier* dengan tingkat akurasi 97.31%.

Prediksi infeksi malaria menggunakan model pembelajaran mesin yang transparan dan dapat dimengerti pendekatannya. Penelitian ini dilakukan oleh (Rajab *et al.*, 2023). Penelitian ini menggunakan berbagai model pembelajaran mesin, yakni *Xtreme Gradient Boosting* (*XGBoost*), *AdaBoost*, *Random Forest*, *K-Nearest Neighbor*, *K-means*, *Support Vector Machine* (*SVM*), *Decision Tree*, *Logistic Regression* (*LR*), *Naive Bayes*, dan *Explainable Boosting Machines* (*EBMs*). Penerapan teknik *Explainable Artificial Intelligence* (*XAI*) seperti *Shapley Additive Explanation* (*SHAP*) dan *Local Interpretable Model-agnostic Explanation* (*LIME*) pada penelitian ini berfungsi untuk meningkatkan interpretabilitas model dan memberikan wawasan yang bermakna terkait prediksi malaria yang tergolong parah. Dataset yang digunakan adalah dataset klinis terbuka yang berisi informasi tentang malaria golongan parah, infeksi non-malaria, dan malaria golongan ringan dari total partisipan sebesar 2207. Data klinis malaria tersebut berasal dari berbagai fasilitas kesehatan di Uganda, yang menyediakan atribut yang relevan untuk analisis dan prediksi pembelajaran mesin. Hasil penelitian menunjukkan bahwa *Random Forest* dan *Explainable Boosting Machines* (*EBMs*) mencapai akurasi tertinggi sebesar 84% dalam memprediksi kejadian malaria parah. *XGBoost* juga menunjukkan akurasi sebesar 83% dalam tugas klasifikasi. Penerapan teknik *Explainable AI* seperti *SHAP* dan *LIME* meningkatkan interpretabilitas model, memberikan wawasan tentang fitur yang memungkinkan adanya prediksi yang jelas terkait kasus malaria golongan parah.

Penelitian prediksi malaria menggunakan data lingkungan sebelumnya juga telah dilakukan oleh (Nkiruka et al, 2021). Penelitian ini berfokus pada negara-negara di sub-saharan Africa yakni pada 6 negara meliputi Burkina Faso, Mali, Niger, Nigeria, Cameroon, dan *Democratic Republic of Congo*. Penelitian ini melakukan rekayasa fitur untuk mengetahui faktor iklim yang berpengaruh dalam terjadinya malaria, *K-means clustering* sebagai pendekripsi *outlier*, dan XGBoost untuk klasifikasi kejadian malaria. Data yang digunakan meliputi variable-variabel iklim seperti curah hujan, suhu, radiasi permukaan, kelembaban relatif, dan tekanan atmosfer yang telah dianalisis selama 28 tahun (1990-2017) yang diambil dari *National Centre for Atmospheric Research* (NCAR) untuk memahami dampaknya terhadap kejadian malaria. Dataset klinis sendiri diambil dari data repositori WHO. Hasil penelitian menunjukkan bahwa suhu, curah hujan, dan radiasi permukaan secara signifikan mempengaruhi terjadinya wabah malaria. Model XGBoost memberikan hasil dengan tingkat akurasi tertinggi disbanding model *Logistic Regression*, *Naïve Bayes*, *Support Vector Machine* (SVM) dan *Decision Tree* dengan rata-rata nilai akurasi 95.83% di enam negara tersebut.

Penelitian prediksi demam berdarah menggunakan XGBoost juga telah dilakukan oleh (Methiyothin & Ahn, 2022) di Prancis dan Thailand. Data yang digunakan meliputi data meteorologi, data *Google Trend*, dan data survei dari berbagai negara, yang dikumpulkan dari sumber seperti *Google Trends*, *National Oceanic and Atmospheric Administration* (NOAA), dan laporan survei dari Thailand dengan rentang waktu dari tahun 2014 hingga 2020. Penelitian ini menerapkan pemilihan fitur-fitur input berdasarkan analisis *cross-correlation* untuk mengidentifikasi keterkaitan antara data demam berdarah dengan fitur input lainnya. Penelitian ini bertujuan untuk menganalisis tren jangka panjang dan jangka pendek terkait wabah Demam Berdarah di Prancis dan Thailand. Hasil penelitian menunjukkan bahwa model XGBoost dapat secara akurat meramalkan wabah demam berdarah baik dalam jangka waktu yang panjang maupun pendek. Diketahui prediksi jangka pendek lebih akurat dibandingkan dengan prediksi jangka panjang dan pentingnya pemilihan fitur input yang relevan dalam meningkatkan akurasi model.

Penelitian penggunaan XGBoost dalam memprediksi *relative density of SLMed Ti-6Al-4V parts* menggunakan dataset berukuran kecil (Zou, M et al., 2022). Penggunaan model XGBoost yang dioptimalisasikan dengan GridsearchCV diusulkan dalam penelitian dan menunjukkan hasil bahwa akurasi prediksi model ini lebih baik

dibandingkan dengan model prediksi *Artifical Neural Network* (ANN) dan *Support Vector Regression* (SVR). Jumlah terkecil pada dataset training yang dicoba adalah 486 dengan data testing berjumlah 122 menghasilkan nilai MAE 1.5577, RMSE 5.1405, R<sup>2</sup> 0.7632. Nilai-nilai evaluasi ini menunjukkan hasil prediksi model bagus dengan dataset yang sangat kecil.

## 2.8. Perbedaan Penelitian

Penelitian prediksi kejadian malaria telah banyak dilakukan penelitian dengan menggunakan beragam model, batasan dan data. Penelitian yang dilakukan oleh (Mbunge et al., 2022) menggunakan data kejadian sebelumnya dari daerah distrik Buhera. Metode-metode yang digunakan pada penelitian ini yakni *Support Vector Machine* (SVM), *Logistic Regression*, *Decision Tree Classifier* dan *Random Forest Classifier*. Penelitian ini tidak menggunakan metode XGBosot dan data yang digunakan berfokus pada area tersebut.

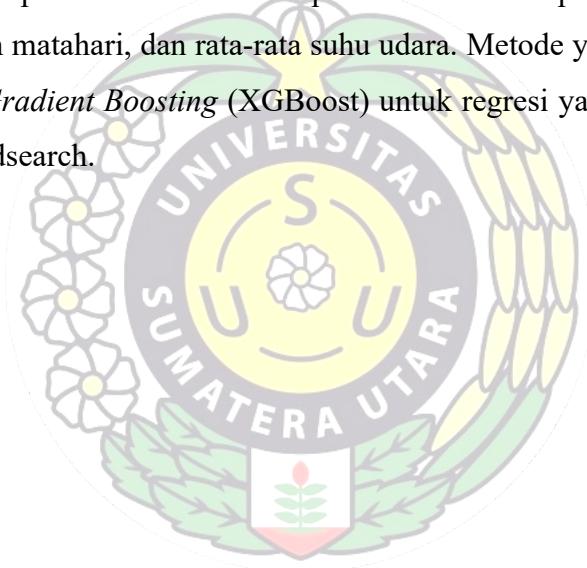
Perbedaan penelitian yang dilakukan oleh (Islam et al., 2023) yakni pada data yang digunakan. Penelitian ini menggunakan data dari google form, media sosial, mahasiswa *Daffodil University*, dan beberapa sumber lainnya. Data hanya terbatas pada data klinis. Metode-metode yang digunakan meliputi *Gaussian NB*, *Bagging*, *Random Forest Classifier*, dan *XGBoost*. Adapun fitur-fitur yang digunakan untuk melatih model penelitian ini yakni usia, pembengkakan pada area mata, jenis kelamin, gejala berkeringat, gejala sakit kepala, pendarahan hidung, nyeri *retro-ocular*, dan adanya kesulitan bernapas.

Perbedaan penelitian yang dilakukan oleh (Rajab et al., 2023) terletak pada tujuan penelitian yang berfokus pada peningkatan interpretabilitas model. Data yang digunakan terbatas pada data klinis malaria yang berasal dari berbagai fasilitas kesehatan di Uganda. Hasil penelitian menunjukkan bahwa XGBoost akurasi yang tergolong akurat yakni sebesar 83% dalam tugas klasifikasi.

Perbedaan penelitian yang dilakukan oleh (Methiyothin & Ahn, 2022) terletak pada metode yang digunakan dan wilayah penelitian. Penelitian yang dilakukan menggunakan metode *Xtreme Gradient Boosting* (XGBoost) pada wilayah Prancis dan Thailand. Penelitian ini bertujuan untuk memprediksi malaria untuk jangka panjang dan jangka pendek. Dataset yang digunakan memiliki rentang waktu dari tahun 2014 hingga 2020 dengan variabel data survei yang berasal dari beberapa negara yang berisi jumlah

kasus, lokasi dan rentang waktu. Data meteorologi juga digunakan seperti suhu, cuaca, curah hujan dan kelembapan. Selain itu, dikumpulkan juga data dari Google Trends untuk mengetahui wawasan public tentang wabah malaria.

Berdasarkan perbandingan penelitian yang telah diuraikan di atas, penelitian ini memiliki perbedaan dengan penelitian yang sudah dilakukan sebelumnya dari segi wilayah penelitian, data variabel input yang akan digunakan dalam model prediksi, dan metode penelitian. Wilayah penelitian ini yakni di Kabupaten Batu Bara, Sumatera Utara dan menggunakan data klinis yang diperoleh dari penelitian sebelumnya (Fahmi, F et al., 2022) yakni data tahun 2019 hingga 2022. Data ini kemudian akan disintesis menggunakan *Generative Adversarial Network* (GAN) dari tahun 2015 hingga tahun 2018 dengan tujuan meningkatkan akurasi model yang didapat. Variabel input yang akan digunakan pada penelitian ini mencakup rata-rata kelembapan udara, curah hujan, lamanya penyinaran matahari, dan rata-rata suhu udara. Metode yang digunakan yakni algoritma *Xtreme Gradient Boosting* (XGBoost) untuk regresi yang dioptimalisasikan dengan metode Gridsearch.



## BAB 3

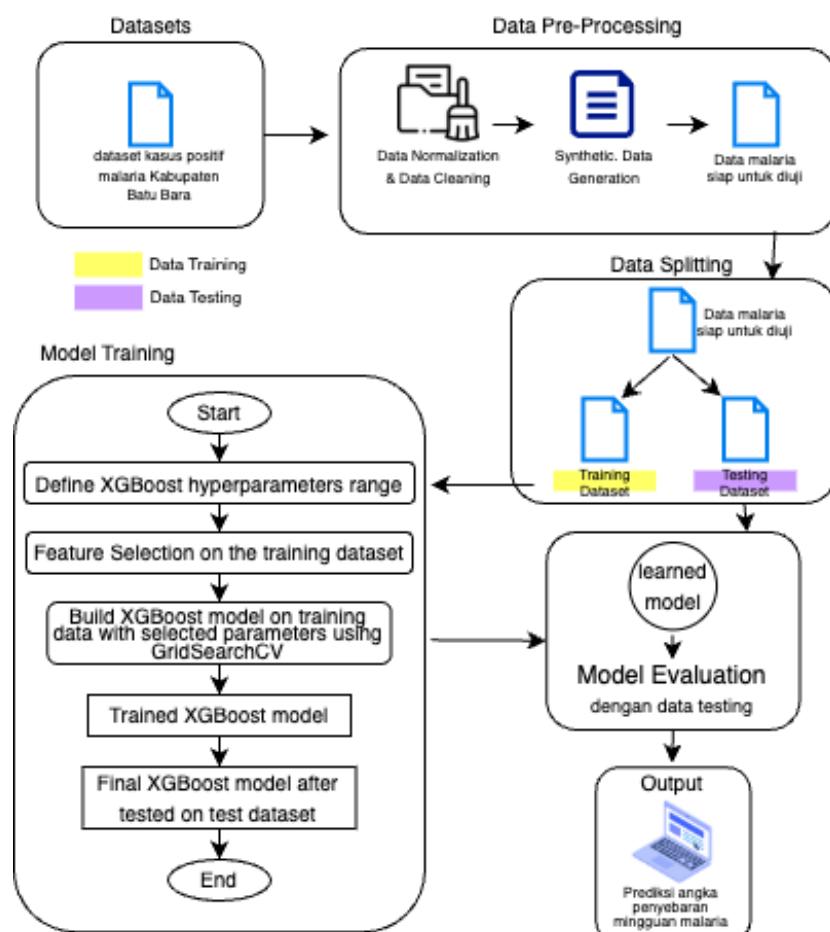
### ANALISIS DAN PERANCANGAN SISTEM

#### 3.1.Data

Penulis memperoleh data kasus positif malaria yang terjadi pada Kabupaten Batu Bara, Sumatera Utara. Data kasus ini berdasarkan variabel-variabel iklim seperti temperature maksimum, curah hujan rata-rata harian, curah hujan kumulatif, indeks kelembapan, indeks tingkat genangan air ataupun banjir, indeks kepadatan penduduk, dan total kasus positif yang diperoleh secara mingguan. Adapun sumber dari data kasus ini yakni dari penelitian terdahulu yang dilakukan oleh Fahmi et.al., pada tahun 2023 dengan judul “Pemodelan Penyebaran Penyakit Malaria Berbasis Geographic Information System dengan Mobile Application dan Analisis Big Data sebagai Alat Bantu Pendukung Program Eliminasi 2030”.

#### 3.2.Arsitektur Umum

Rancangan arsitektur umum pada penelitian ini beserta setiap tahapannya yakni:



Gambar 3.1 Arsitektur Umum Penelitian

### 3.2.1. Datasets

Data yang digunakan pada penelitian ini adalah data kasus positif malaria di Kabupaten Batu Bara pada tahun 2019 hingga tahun 2022 yang berjumlah sebanyak 210 data. Penulis memperoleh data ini dari penelitian Fahmi et al. (2023). Penulis nantinya akan men-generate data sintesis dengan tujuan untuk memperbanyak data training yang didapat dengan model GAN. Model ini nantinya akan mempelajari pola data klinis yang telah ada yakni dari 210 data awal.

### 3.2.2. Data Preprocessing

Pada tahapan ini, dilakukan pemrosesan terhadap data yang sudah dikumpulkan agar siap diproses yakni dengan dibersihkan, disintesis, dan digabungkan, serta dilakukan pemilihan variabel yang mempengaruhi hasil dari model prediksi. Pada tahap *preprocessing* meliputi beberapa tahapan, yaitu:

#### 3.2.2.1. Data Checking

Data-data yang telah dikumpulkan memiliki ketidaksempurnaan atau disebut “data kotor”, karena mungkin masih terdapat *missing value*, dan data yang berulang atau duplikat. Oleh sebab itu, diperlukan tahap pembersihan atau ‘data *cleaning*’. Tahap ini berisi pembersihan data dengan menghapus informasi yang tidak relevan (*outlier*) dan data duplikat, serta mengisi *missing value* dengan nilai rata-rata atau median dari jumlah data. Adapun *Pseudocode* proses *data cleaning* sebagai berikut :

*Function data\_checking (data):*

```

Inisialisasi checked_data sebagai salinan data asli
#menangani missing values
For setiap kolom dalam checked_data:
    If kolom memiliki missing values Then:
        Ganti missing values dengan rata-rata dari kolom tersebut
    End If
End For
#menangani outliers
For setiap kolom dalam checkned_data:
    Hitung Q1 (kuartil pertama) dan Q3 (kuartil ketiga)
    Hitung IQR (interquartile range) sebagai Q3 - Q1
    Hitung lower_bound sebagai Q1 - 1.5 * IQR

```

*Hitung upper\_bound sebagai  $Q3 + 1.5 * IQR$*

*For setiap nilai dalam kolom:*

*If nilai di luar lower\_bound atau upper\_bound Then:*

*Ganti outlier dengan nilai median dari kolom*

*End If*

*End For*

*End For*

*Return checked\_data*

### 3.2.2.2. Synthetic Data Generation

Dengan tujuan memperbanyak data yang akan digunakan dalam penelitian, maka diperlukan pembuatan data sintesis. Data sintesis yang dibutuhkan adalah dataset angka penyebaran kasus positif di Kabupaten Batu Bara pada tahun 2015 hingga tahun 2018. Metode yang digunakan dalam tahap ini yakni model *Generative Adversial Network* (GAN). Model ini merupakan model *Time-series* yang dimana dipilih dikarenakan kemampuannya meniru distribusi data asli secara efektif, mampu mengatasi masalah keterbatasan data yang dihadapi dan kemampuan mempertahankan dinamika temporal deret waktu sehubungan dengan data dalam penelitian ini berhubungan dengan waktu (Yoon et al., 2019). Implementasi timeGAN dilakukan dengan penggunaan library *ydata\_synthetic*. Library ini secara otomatis menormalisasikan data dengan *MinMaxScaler* sehingga harus dilakukan *inverse transform* pada data sintesis yang dihasilkan untuk mengembalikan data ke rentang skala aslinya. Adapun *pseudocode* untuk proses ini yakni:

*Function generate\_synthetic\_data (data):*

*Import TimeGAN from ydata\_synthetic*

*Initialize the TimeGAN model*

*Fit the model on the preprocessed data*

*Generate synthetic\_data samples*

*Inverse transform synthetic\_data*

*Tambahkan kolom weeks pada synthetic\_data*

*Return synthetic\_data*

### 3.2.2.3. Data Unification

Setelah melewati tahapan pembersihan data dan pembuatan data sintesis, maka data-data yang telah dikumpulkan akan disatukan menjadi 1 dataset yang utuh. Data asli dari tahun 2019-2022 dan data sintesis tahun 2015-2018 digabung dan kemudian diurutkan berdasarkan kolom *weeks*. Tahapan penggabungan ini menggunakan bahasa pemrograman Python beserta *library-library* seperti numpy dan pandas. Dataset kemudian siap untuk digunakan baik untuk *training* maupun *testing*.

### 3.2.3. Data Splitting

Setelah melewati tahap preprocessing, selanjutnya data akan dibagi menjadi dua bagian yaitu data *training* dan data *testing* dengan rasio 80:20. Sebanyak 80% data akan digunakan sebagai data *training* dalam pembuatan model dan sisa 20% lainnya akan menjadi data *testing* yang berfungsi untuk menguji model yang telah dibuat. Pemilihan rasio *data splitting* yang digunakan adalah 80:20, karena merupakan rasio yang umum digunakan praktisi mengikuti *Pareto principle* (Joseph, 2022). Penggunaan rasio yang berbeda dapat mempengaruhi nilai akurasi performa model (Nazarkar et al., 2023).

### 3.2.4. Feature Selection on The Training Dataset

Tahapan ini hanya terjadi pada data *training* dengan tujuan mempercepat proses, meningkatkan kinerja model, dan mengurangi kemungkinan terjadinya *overfitting*. Dalam tahap ini, dilakukan pemilihan fitur-fitur yang berpengaruh secara signifikan pada hasil prediksi target. Metode yang penulis gunakan adalah *Pearson Correlation*. Metode ini digunakan untuk mengukur angka hubungan linear antar 2 variabel dengan rentang nilai dari -1 hingga 1. Nilai 1 menandakan korelasi positif sempurna, 0 menandakan tidak adanya korelasi dan -1 menandakan korelasi negative sempurna. Dapat disimpulkan bahwa apabila nilai korelasi semakin mendekati angka 1 dan -1, maka semakin terdapat korelasi pada hubungan antar variabel tersebut.

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad \dots(2)$$

Dimana  $r$  adalah nilai korelasi antara variabel X dan Y,  $n$  adalah jumlah pasangan data yang diobservasi,  $X_i$  dan  $Y_i$  adalah nilai-nilai dari variabel X dan Y dari data ke-i,  $\bar{X}$  dan  $\bar{Y}$  adalah nilai rata-rata dari variabel X dan Y.

Penerapan perhitungan *Pearson Correlation* terhadap sample data untuk variabel input indeks curah hujan kumulatif dan total kasus pada tabel 3.2.1 adalah sebagai berikut:

Tabel 3.2.1 Tabel variabel indeks curah hujan kumulatif dan total kasus

Indeks Curah Hujan Kumulatif	Total kasus
1,51	7
2,01	10
2,01	6

$$\bar{y} = \frac{7 + 10 + 6}{3} \approx 7.67$$

$$\bar{x} = \frac{1.51 + 2.01 + 2.01}{3} \approx 1.84$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = ((1.51 - 1.84)(7 - 7.67)) + ((2.01 - 1.84)(10 - 7.67)) + ((2.01 - 1.84)(6 - 7.67)) \approx 0.33$$

$$\sum_{i=1}^n (X_i - \bar{x})^2 = (1.51 - 1.84)^2 + (2.01 - 1.84)^2 + (2.01 - 1.84)^2 \approx 1.67$$

$$\sum_{i=1}^n (Y_i - \bar{y})^2 = (7 - 7.67)^2 + (10 - 7.67)^2 + (6 - 7.67)^2 \approx 8.67$$

$$r = \frac{0.33}{\sqrt{1.67 \cdot 8.67}} \approx 0.28$$

### 3.2.5. Model Training

Tahap ini adalah tahap pembuatan model menggunakan algoritma XGBoost dan data *training* yang telah melewati *feature selection*. Sebelum pembuatan model, dilakukan optimasi model dengan pencarian rasio pembagian data *training* dan data *testing* yang kemudian dilanjutkan dengan pencarian parameter-parameter model yang memberikan hasil paling optimal menggunakan GridSearchCV. Penggunaan GridSearchCV dikarenakan metode ini mencoba semua gabungan value parameter yang diberikan dan bisa digabungkan dengan *k-fold cross validation*. Dalam kasus ini, penulis akan menggunakan *4-fold cross validation*. Dengan menggunakan parameter yang optimal, dilakukan inisialisasi model XGBoost yang kemudian dilatih dengan data *training* untuk mempelajari pola dan hubungan antara fitur-fitur input dan variabel target regresi. Ini dilakukan dengan mengoptimalkan fungsi tujuan, yang biasanya

berupa fungsi kerugian seperti *mean squared error* (MSE). Selanjutnya, model dievaluasi secara iteratif pada setiap langkah dalam pelatihan menggunakan teknik pemberian umpan balik seperti *gradient boosting*. Proses ini berlanjut hingga model mencapai performa yang optimal atau kriteria penghentian tertentu. Setelah pelatihan selesai, model XGBoost siap untuk diuji. Saat diuji dengan data testing, model menggunakan pengetahuan yang diperoleh dari data training untuk membuat prediksi dan mengevaluasi tingkat ketidakpastian.

### 3.2.6. Model Evaluation and Comparison

Model yang sudah dilatih akan dievaluasi untuk menghitung tingkat akurasi dari hasil prediksi model dengan nilai asli. Metode evaluasi pada penelitian ini, yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), *Mean Square Error* (MSE), serta *Coefficient of Determination* ( $R^2$ ).

#### 3.2.6.1. Root Mean Square Error (RMSE)

RMSE merupakan metrik evaluasi yang umum digunakan untuk mengukur kinerja model regresi dalam memprediksi nilai sebenarnya. RMSE menghitung akar dari rata-rata selisih kuadrat nilai prediksi ( $\hat{y}_i$ ) dengan nilai sebenarnya ( $y_i$ ). RMSE memberikan bobot lebih besar pada kesalahan besar, sehingga lebih sensitif terhadap *outlier* dibandingkan dengan metrik evaluasi lainnya. Hal ini jugalah yang menunjukkan bahwa RMSE rentan terhadap *outlier* dan juga perubahan skala pada data. Dengan begitu, dapat disimpulkan bahwa semakin kecil nilai RMSE, maka hasil prediksi akan semakin akurat. Secara sistematis, RMSE didefinisikan dengan rumus sebagai berikut:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad \dots(3)$$

Penerapan perhitungan matriks RMSE terhadap sample perbandingan nilai aktual dan prediksi pada tabel 3.2.2 adalah sebagai berikut:

Tabel 3.2.2 Tabel perbandingan nilai aktual dan prediksi RMSE

Weeks	Actual	Predicted
2019-01-28 00.00.00	7	5
2019-02-04 00.00.00	10	7
2019-02-11 00.00.00	10	7
2019-02-18 00.00.00	6	6
2019-02-25 00.00.00	4	6

$$RMSE = \sqrt{\frac{(7-5)^2 + (10-7)^2 + (10-7)^2 + (6-6)^2 + (4-6)^2}{5}} \approx 2.28$$

### 3.2.6.2. Mean Absolute Error (MAE)

MAE merupakan metrik evaluasi yang mengukur kinerja model regresi dengan menghitung rata-rata dari selisih absolut antara nilai prediksi ( $\hat{y}_i$ ) dan nilai aktual ( $y_i$ ). MAE kurang rentan terpengaruh oleh *outlier*, karena selisih absolut yang digunakan tidak memberikan dampak yang lebih besar pada kesalahan besar. MAE memberikan hasil dengan satuan yang sama dengan nilai aktual. Sama dengan RMSE, semakin kecil nilai MAE, semakin baik kualitas model yang dibuat. Secara sistematis, MAE didefinisikan dengan rumus sebagai berikut:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \dots (4)$$

Penerapan perhitungan matriks MAE terhadap sample perbandingan nilai aktual dan prediksi pada tabel 3.2.3 adalah sebagai berikut:

Tabel 3.2.3 Tabel perbandingan nilai aktual dan prediksi MAE

Weeks	Actual	Predicted
2021-05-17 00.00.00	15	17
2021-05-24 00.00.00	12	13
2021-05-31 00.00.00	15	16

$$MAE = \frac{|15 - 17| + |12 - 13| + |15 - 16|}{3} \approx 1,33$$

### 3.2.6.3. Mean Square Error (MSE)

MSE adalah metrik evaluasi yang mengukur rata-rata selisih kuadrat nilai prediksi ( $\hat{y}_i$ ) dengan nilai sebenarnya ( $y_i$ ). Berbeda dengan RMSE, MSE memberikan memberikan bobot yang lebih besar pada perbedaan yang signifikan. Tetapi MSE juga rentan terhadap *outlier*. Rumus MSE sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \dots (5)$$

Penerapan perhitungan matriks MSE terhadap sample perbandingan nilai aktual dan prediksi pada tabel 3.2.4 adalah sebagai berikut:

Tabel 3.2.4 Tabel perbandingan nilai aktual dan prediksi MSE

Weeks	Actual	Predicted
2021-05-17 00.00.00	15	17
2021-05-24 00.00.00	12	13
2021-05-31 00.00.00	15	16

$$MSE = \frac{(15 - 17)^2 + (12 - 13)^2 + (15 - 16)^2}{3} = 2$$

#### 3.2.6.4. Coefficient of Determination ( $R^2$ )

$R^2$  atau yang dikenal juga sebagai *R-squared* merupakan suatu nilai yang memperlihatkan seberapa besar variabel independen (eksogen) mempengaruhi variabel dependen (endogen) atau seberapa sesuai model dengan data. Nilai  $R^2$  berada di rentang angka 0 hingga 1, dengan 0 yang mengindikasikan tidak adanya kesesuaian terhadap nilai variabel dependen dan 1 yang mengindikasikan adanya kesesuaian dengan data. Sehingga berbeda dengan metrik-metrik lainnya, semakin besar nilai  $R^2$ , semakin baik. Nilai  $R^2$  dikalkulasikan dengan membandingkan *Regression Sum of Squares* (SSR) dengan *Total Sum of Squares* (SST), dimana SSR adalah jumlah kuadrat dari perbedaan antara nilai prediksi ( $\hat{y}_i$ ) dan nilai aktual ( $y_i$ ) dan SST adalah jumlah jumlah kuadrat perbedaan antara nilai aktual ( $y_i$ ) dan rata-rata nilai aktual ( $\bar{y}$ ).  $R^2$  dapat dirumuskan sebagai berikut:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum i (y_i - \hat{y}_i)^2}{\sum i (y_i - \bar{y})^2} \quad \dots(6)$$

Penerapan perhitungan matriks  $R^2$  terhadap sample perbandingan nilai aktual dan prediksi pada tabel 3.2.5 adalah sebagai berikut:

Tabel 3.2.5 Tabel perbandingan nilai aktual dan prediksi  $R^2$ 

Weeks	Actual	Predicted
2020-08-10 00.00.00	8	6
2020-08-17 00.00.00	4	5
2020-08-24 00.00.00	5	5
2020-08-31 00.00.00	9	10
2020-09-07 00.00.00	13	15

$$\bar{y} = \frac{8 + 4 + 5 + 9 + 13}{5} = 7.8$$

$$SSR = (8 - 6)^2 + (4 - 5)^2 + (5 - 5)^2 + (9 - 10)^2 + (13 - 15)^2 = 10$$

$$SST = (8 - 7.8)^2 + (4 - 7.8)^2 + (5 - 7.8)^2 + (9 - 7.8)^2 + (13 - 7.8)^2 = 50.8$$

$$R^2 = 1 - \frac{10}{50.8} \approx 0.803$$

### 3.2.7. Output

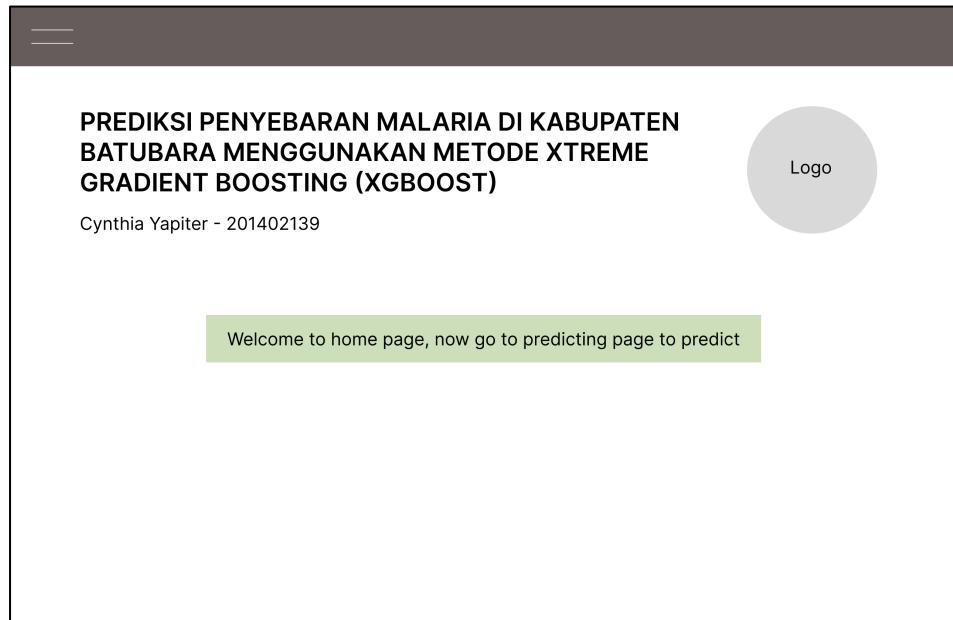
Pada tahap ini, telah dihasilkan model prediksi menggunakan algoritma XGBoost dengan hyperparameter yang optimal. Model ini mampu memprediksi angka penyebaran malaria mingguan di Kabupaten Batu Bara, Sumatera Utara. Selanjutnya, model tersebut akan diimplementasikan ke dalam sistem berupa *website* untuk digunakan pada masa yang akan datang.

## 3.3.Rancangan Antarmuka Sistem

Perancangan antarmuka sistem dilakukan terlebih dahulu untuk mengetahui tampilan website yang hendak dibuat nantinya untuk pengaplikasian model. Pada bagian *sidebar* website, terdapat 2 menu pilihan utama. Pilihan-pilihan utama itu yakni *testing* dan *predicting*. Hasil dari semua proses akan ditampilkan di halaman *result*.

### 1. Rancangan tampilan *home page*

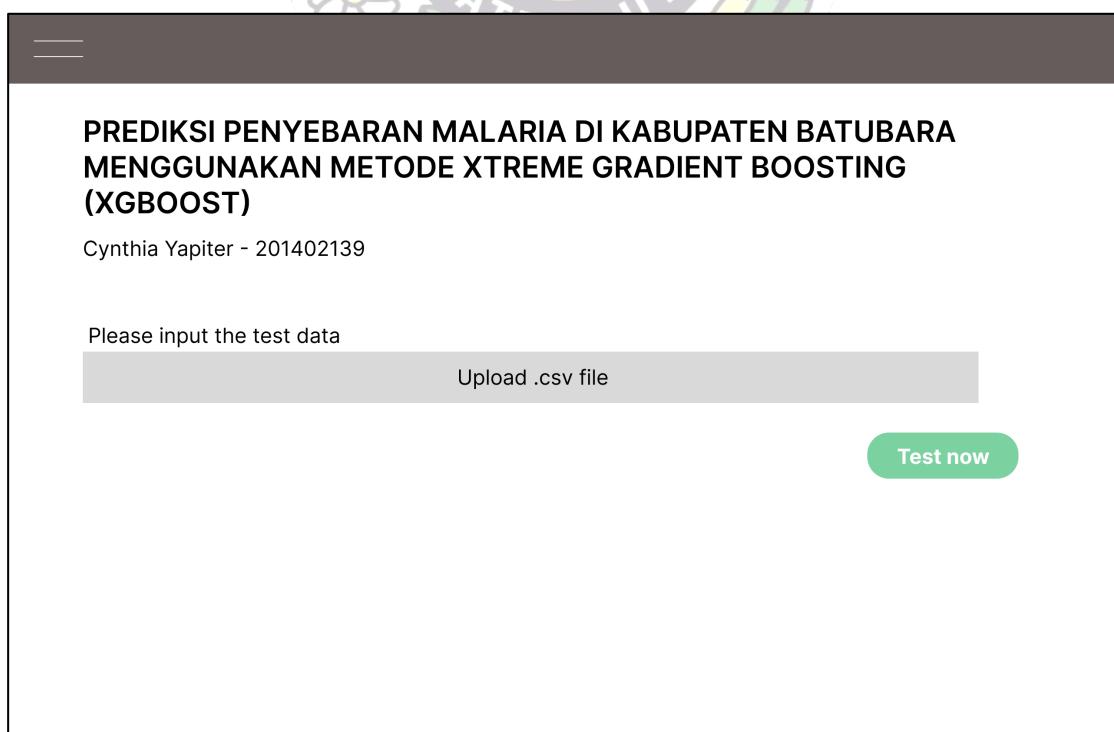
Halaman ini merupakan tampilan awal ketika *user* mengakses *website* atau yang biasa disebut sebagai *landing page*. Isi dari halaman ini yakni berupa keterangan-keterangan penelitian seperti judul, logo universitas, nama peneliti, nomor induk mahasiswa peneliti, nama program studi dan fakultas peneliti. *Sidebar* terletak di bagian pojok kiri atas halaman yang memiliki fungsi mengarahkan *user* ke halaman *testing* dan *predicting*. Adapun rancangan tampilan *home page* dapat dilihat pada gambar 3.2.



Gambar 3.2 Rancangan tampilan *home page*

## 2. Rancangan tampilan *testing page*

Tampilan yang muncul ketika *user* memilih *menu testing* pada *sidebar*. Halaman ini memiliki keterangan-keterangan yang serupa dengan *home page* yakni judul penelitian, nama dan nomor induk mahasiswa peneliti. Terdapat *form* untuk memasukkan *data* yang hendak diuji dengan tombol *submit* ‘*Test now*’. Adapun rancangan tampilan *testing page* dapat dilihat pada gambar 3.3.



Gambar 3.3 Rancangan tampilan *testing page*

Hasil proses *testing* setelah pengumpulan data melalui tombol submit akan ditampilkan pada halaman *result page*. Terdapat tabel perbandingan data asli dengan data yang diprediksi model. Tabel ini ditampilkan sebagian dengan pembagian *pagination* 10 data pada setiap *page*. Adapun tampilan tabel perbandingan yang dirancang dapat dilihat pada gambar 3.4.

PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA  
MENGGUNAKAN METODE XTREME GRADIENT BOOSTING  
(XGBOOST)

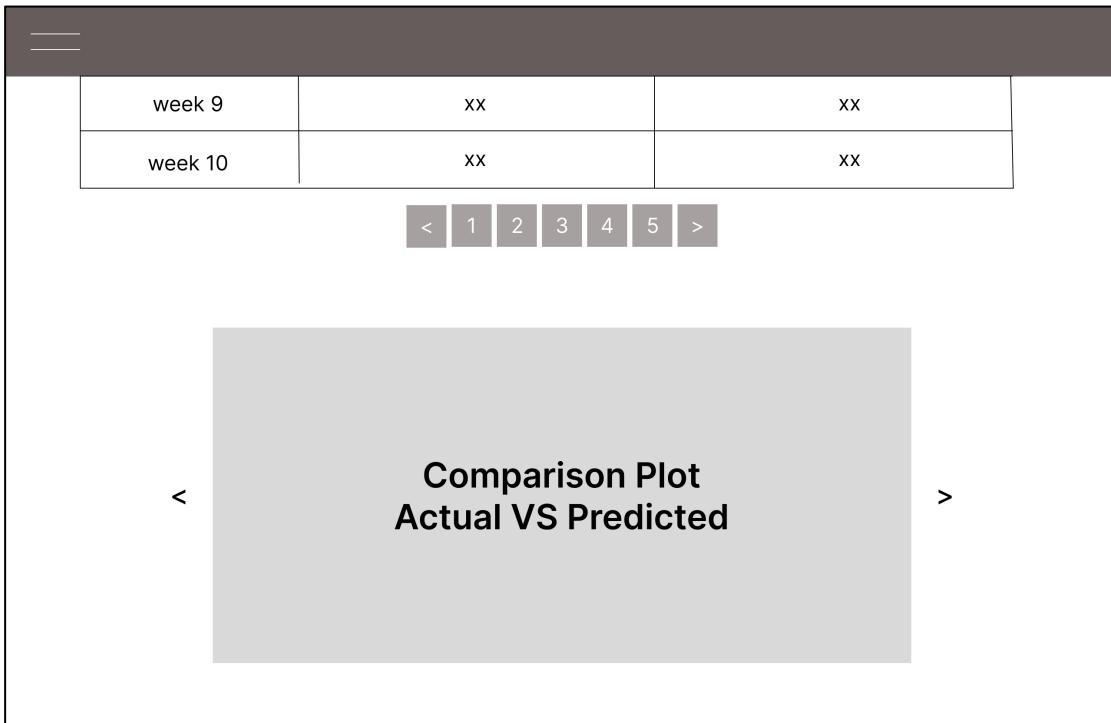
Cynthia Yapiter - 201402139

Testing Result  
Comparison Table

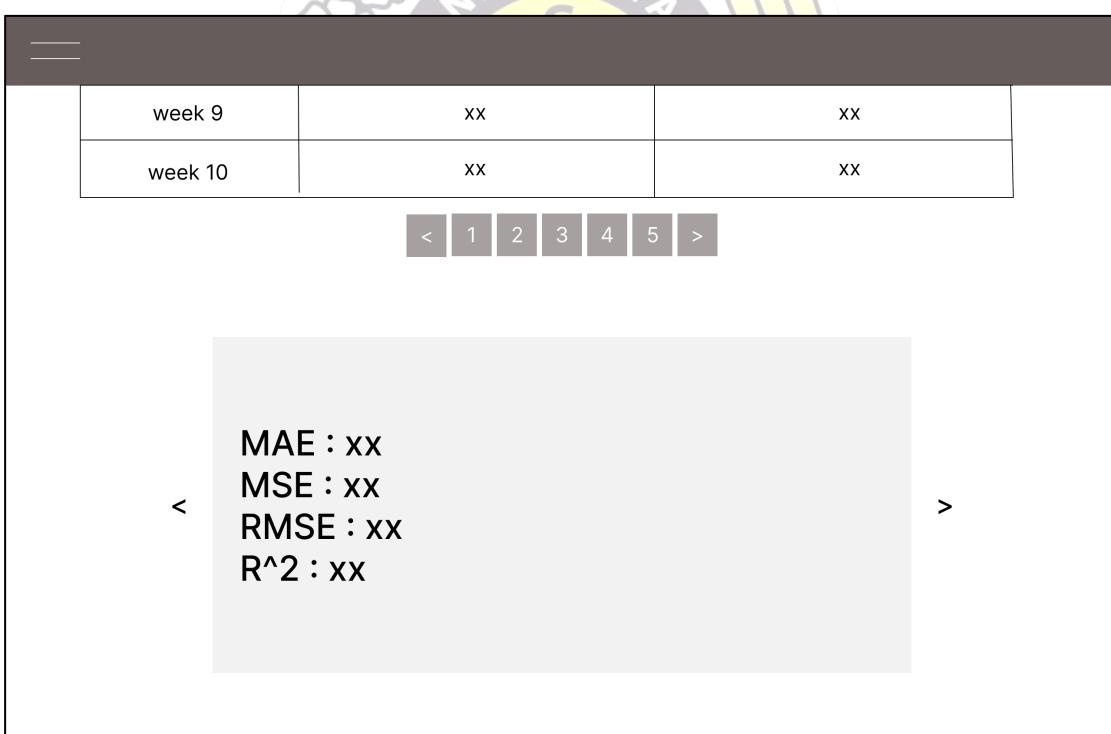
Weeks	Actual	Predicted
week 1	xx	xx
week 2	xx	xx
week 3	xx	xx
week 4	xx	xx
week 5	xx	xx
week 6	xx	xx

Gambar 3.4 Rancangan tampilan *result page* dengan tabel perbandingan

Pada *result page*, dibawah tabel perbandingan terdapat gambar plot perbandingan data asli dengan data hasil prediksi model dan juga tabel nilai evaluasi model. Tabel ini berisi nilai MAE, MSE, RMSE dan  $R^2$ . Gambar plot perbandingan dan tabel nilai evaluasi ditampilkan dalam bentuk slideshow. Tampilan tabel perbandingan yang dirancang dapat dilihat pada gambar 3.5 dan gambar 3.6.



Gambar 3.5 Rancangan tampilan *result page* dengan plot perbandingan



Gambar 3.6 Rancangan tampilan *result page* dengan tabel evaluasi

### 3. Rancangan tampilan *predicting page*

Halaman yang muncul ketika *menu testing* pada *sidebar* ditekan. Halaman ini memiliki keterangan-keterangan yang serupa dengan *home page* yakni judul penelitian,

nama dan nomor induk mahasiswa peneliti. Terdapat *form* untuk memasukkan *data* yang hendak diprediksi dengan tombol *submit* ‘*Predict*’. Adapun rancangan tampilan *predicting page* dapat dilihat pada gambar 3.7.

The screenshot shows a web-based prediction interface. At the top, there is a dark header bar. Below it, the main title is displayed in bold capital letters: "PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE XTREME GRADIENT BOOSTING (XGBOOST)". Underneath the title, the author's name "Cynthia Yapiter - 201402139" is shown. The main content area has a light gray background. It contains a text input field with the placeholder "Please input the dataset that needs to be predicted" and a sub-placeholder "Upload .csv file". To the right of this input field is a green rounded rectangular button with the word "Predict" in white. The overall design is clean and functional.

Gambar 3.7 Rancangan tampilan *predicting page*

Hasil proses *predicting* setelah pengumpulan data pada *form* akan ditampilkan pada halaman *predict result page*. Terdapat tabel yang menampilkan hasil prediksi model secara mingguan. Sama seperti halaman *result*, tabel ini juga ditampilkan sebagian dengan pembagian *pagination* 10 data pada setiap *page*. Adapun tampilan tabel hasil yang dirancang dapat dilihat pada gambar 3.8.

Weeks	Predicted
week 1	xx
week 2	xx
week 3	xx
week 4	xx
week 5	xx
week 6	xx

Gambar 3.8 Rancangan tampilan *predict result page* dengan tabel hasil

Pada *predict result page*, dibawah tabel hasil terdapat gambar plot hasil prediksi model. Rancangan tampilan tabel hasil dapat dilihat pada gambar 3.9.

week 9	xx
week 10	xx

< 1 2 3 4 5 >

**Predict Result Plot**

Gambar 3.9 Rancangan tampilan *result page* dengan plot prediksi

## **BAB 4**

### **IMPLEMENTASI DAN PENGUJIAN SISTEM**

#### **4.1. Implementasi Sistem**

Untuk implementasi sistem prediksi kasus positif malaria pada Kabupaten Batu Bara menggunakan metode XGBoost di penelitian ini, penulis harus mempertimbangkan hal-hal seperti *hardware* dan *software* yang digunakan.

##### **4.1.1. Spesifikasi perangkat**

*Hardware* yang digunakan dalam perancangan sistem pada penelitian ini memiliki spesifikasi sebagai berikut:

1. Laptop Macbook Pro dengan Apple M1 chip
2. Kapasitas Memori sebesar 8 GB

Adapun *software* yang digunakan dalam perancangan sistem pada penelitian ini memiliki spesifikasi sebagai berikut:

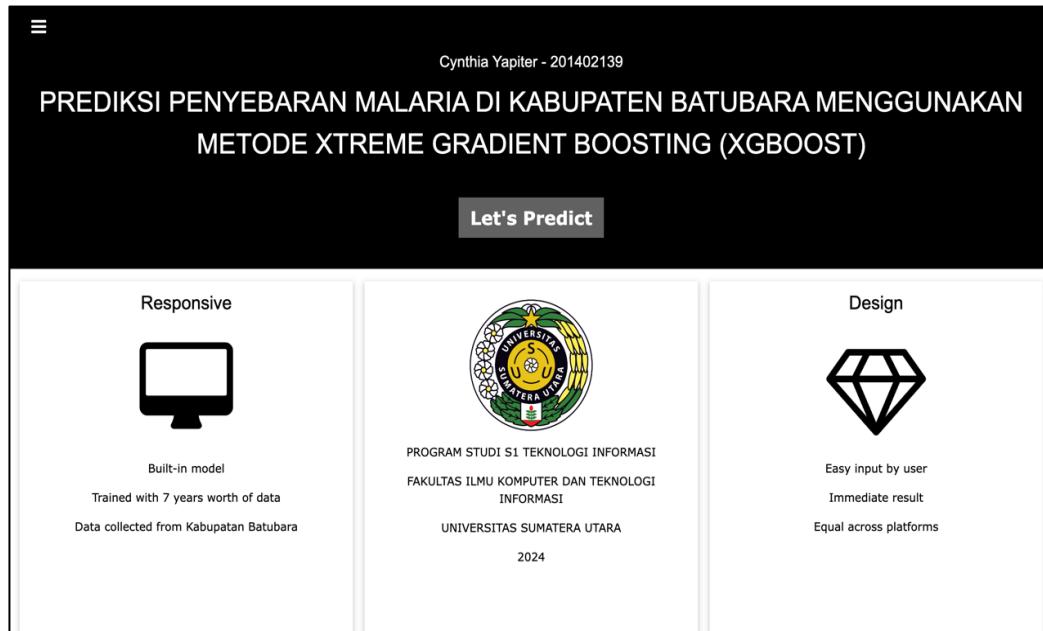
1. *Operating System* : macOS Sonoma
2. Bahasa pemrograman *Python* versi 3.11.3 dan *library-library* yang digunakan yakni *xgboost* versi 2.1.1, *pandas* versi 2.2.2, *numpy* versi 2.1.1, *matplotlib* versi 3.9.2, *flask* versi 3.0.3, dan *scipy* versi 1.14.1.

##### **4.1.2. Implementasi perancangan antarmuka**

Rancangan *interface* yang telah dibuat sebelumnya pada Bab 3 diimplementasikan pada sebuah website. Berikut tampilan hasil implementasi rancangan tersebut secara detail :

###### **1. Tampilan *Home Page***

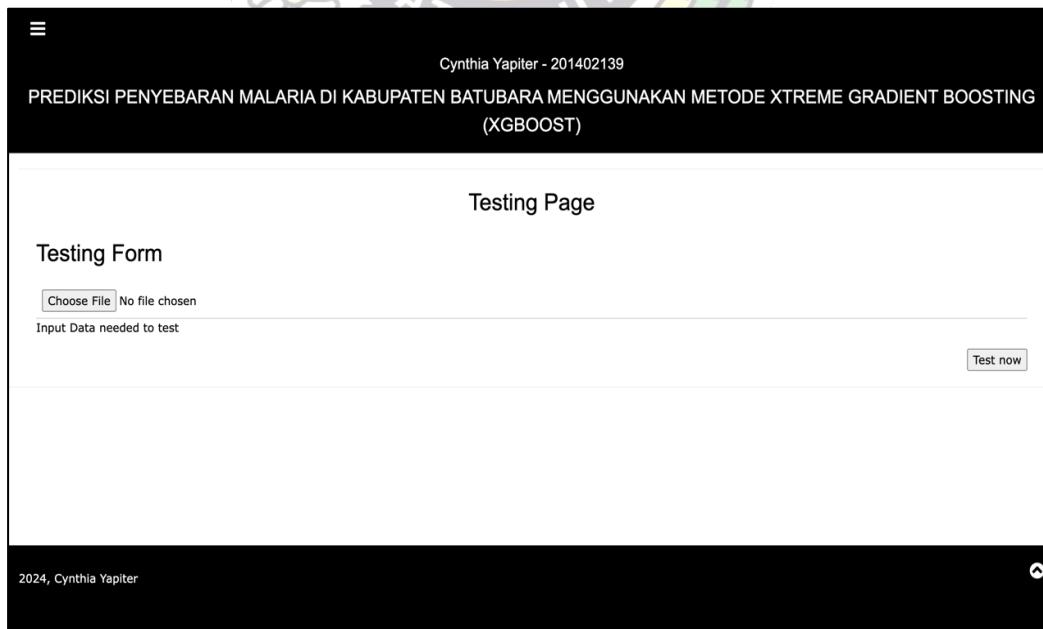
Tampilan awal yang muncul ketika *user* mengakses *website*. Halaman ini berisi informasi *website* dan tombol *sidebar* terletak di bagian pojok kiri atas halaman. Terdapat juga tombol *Let's Predict* yang mengarahkan *user* langsung ke halaman *predicting*. Adapun tampilan Home Page dapat dilihat pada gambar 4.1.1.



Gambar 4.1.1 Implementasi tampilan *home page*

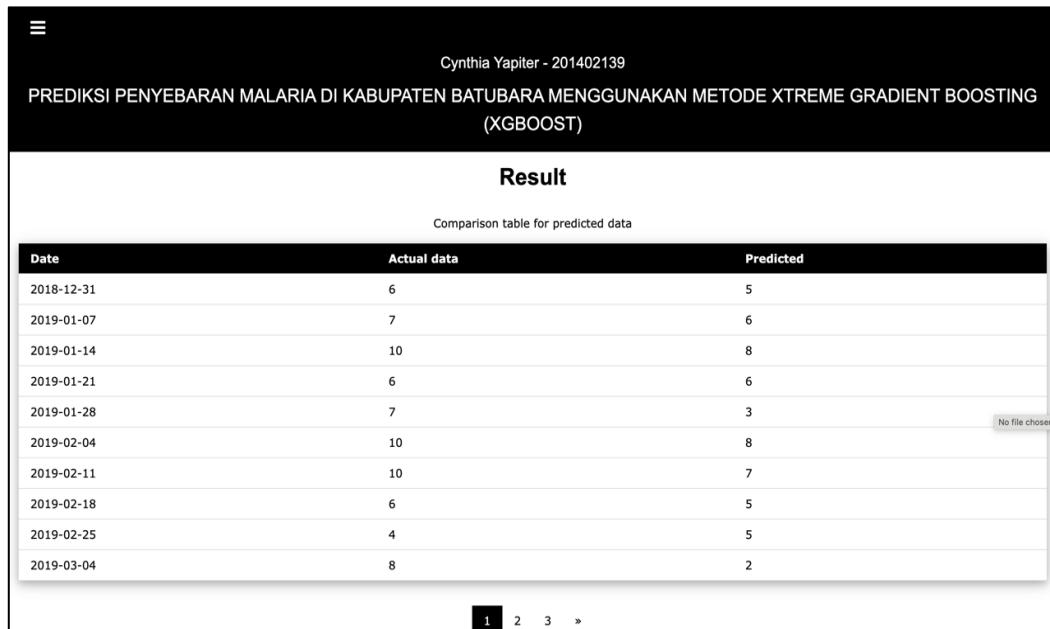
## 2. Tampilan *Testing Page*

Tampilan yang muncul ketika memilih *menu testing* dari *sidebar*. Pada halaman ini terdapat informasi singkat mengenai penelitian dan sebuah *form* untuk memasukkan *data* yang hendak diuji dengan tombol *submit* ‘*Test now*’. Adapun implementasi tampilan *testing page* dapat dilihat pada gambar 4.1.2.



4.1.2 Implementasi tampilan *testing page*

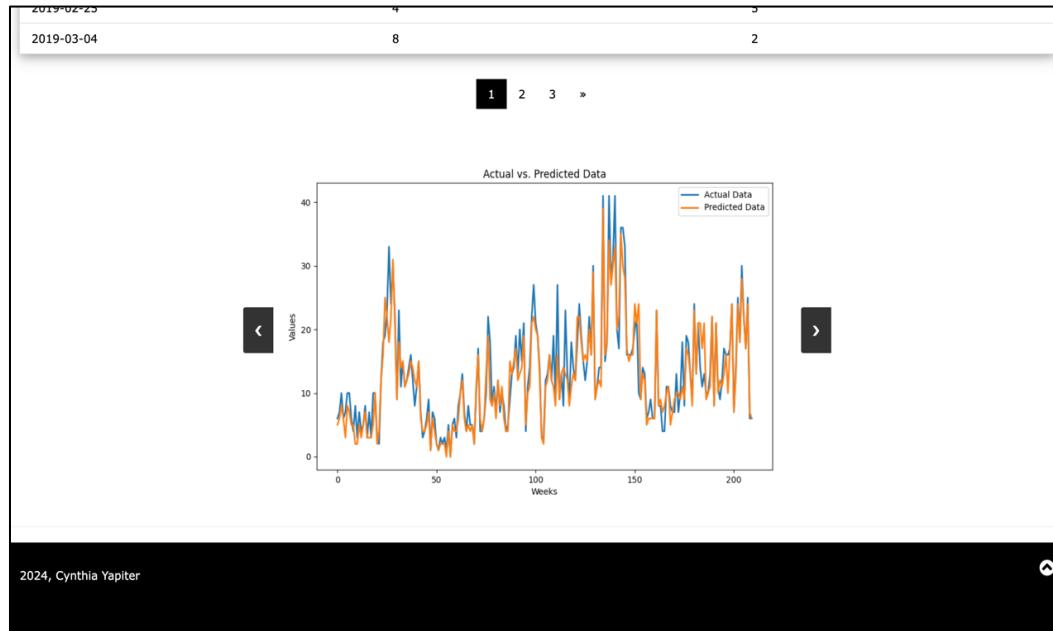
Dengan menekan tombol submit, halaman *result page* akan muncul. Terdapat tabel perbandingan data asli dengan data yang diprediksi model. Tabel ini ditampilkan sebagian dengan pembagian *pagination* 10 data pada setiap *page*. Adapun tampilan hasil tabel perbandingan dapat dilihat pada gambar 4.1.3.



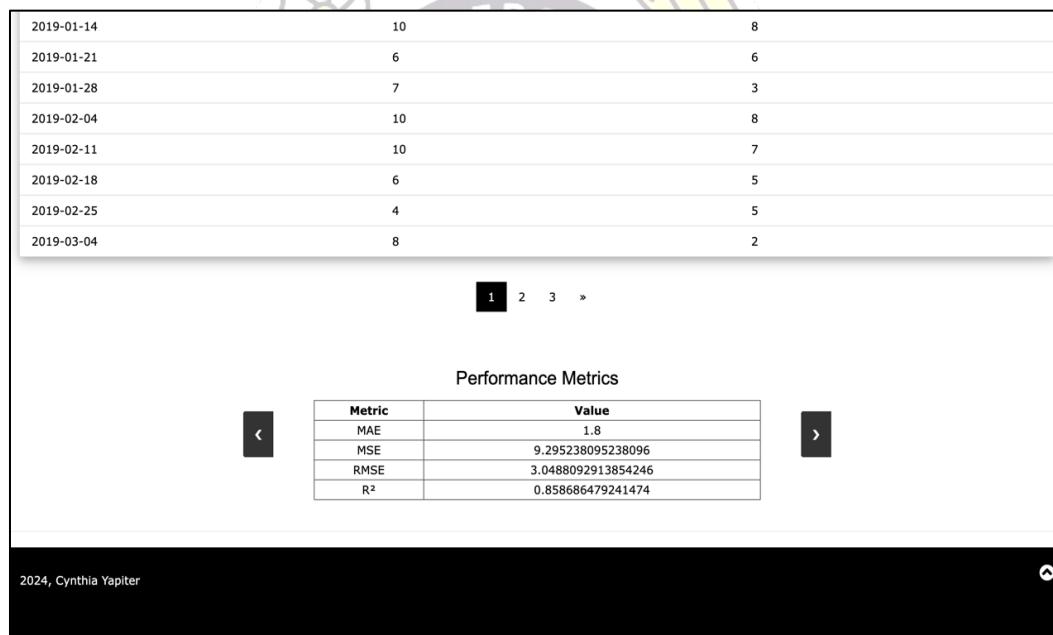
Date	Actual data	Predicted
2018-12-31	6	5
2019-01-07	7	6
2019-01-14	10	8
2019-01-21	6	6
2019-01-28	7	3
2019-02-04	10	8
2019-02-11	10	7
2019-02-18	6	5
2019-02-25	4	5
2019-03-04	8	2

Gambar 4.1.3 Implementasi tampilan *result page comparison table*

Selain itu juga terdapat *plot* yang menunjukkan perbandingan nilai asli dengan nilai hasil prediksi dan tabel hasil evaluasi yakni nilai-nilai MAE, MSE, RMSE dan  $R^2$ . Plot dan tabel ditampilkan secara *slideshow*. Tampilan implementasi ini ditunjukkan pada gambar 4.1.4 dan gambar 4.1.5.



Gambar 4.1.4 Implementasi tampilan *result page comparison plot*



Gambar 4.1.5 Implementasi tampilan *result page evaluation metrics*

### 3. Tampilan *Predicting Page*

Tampilan yang muncul ketika memilih *menu predicting* dari *sidebar*. Pada halaman ini juga terdapat informasi singkat mengenai penelitian dan sebuah *form* untuk memasukkan *data* yang hendak diprediksi total kasus kejadian positif malaria dengan

tombol *submit* ‘*Predict*’. Adapun implementasi tampilan *predicting page* dapat dilihat pada gambar 4.1.6.

Cynthia Yapiter - 201402139  
PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE XTREME GRADIENT BOOSTING (XGBOOST)

Predicting Page

Predicting Form

Choose File No file chosen

Input Data needed to predict

Predict

2024, Cynthia Yapiter

Gambar 4.1.6 Implementasi tampilan *predicting page*

Dengan menekan tombol *Predict*, maka akan muncul halaman *result predict page*. Terdapat tabel yang menunjukkan data asli dengan data hasil prediksi model menggunakan metode XGBoost. Tabel ini ditampilkan dengan pembagian *pagination* 10 data pada setiap halamannya. Adapun tampilan hasil tabel prediksi dapat dilihat pada gambar 4.1.7.

Cynthia Yapiter - 201402139

PREDIKSI PENYEBARAN MALARIA DI KABUPATEN BATUBARA MENGGUNAKAN METODE XTREME GRADIENT BOOSTING (XGBOOST)

### Result

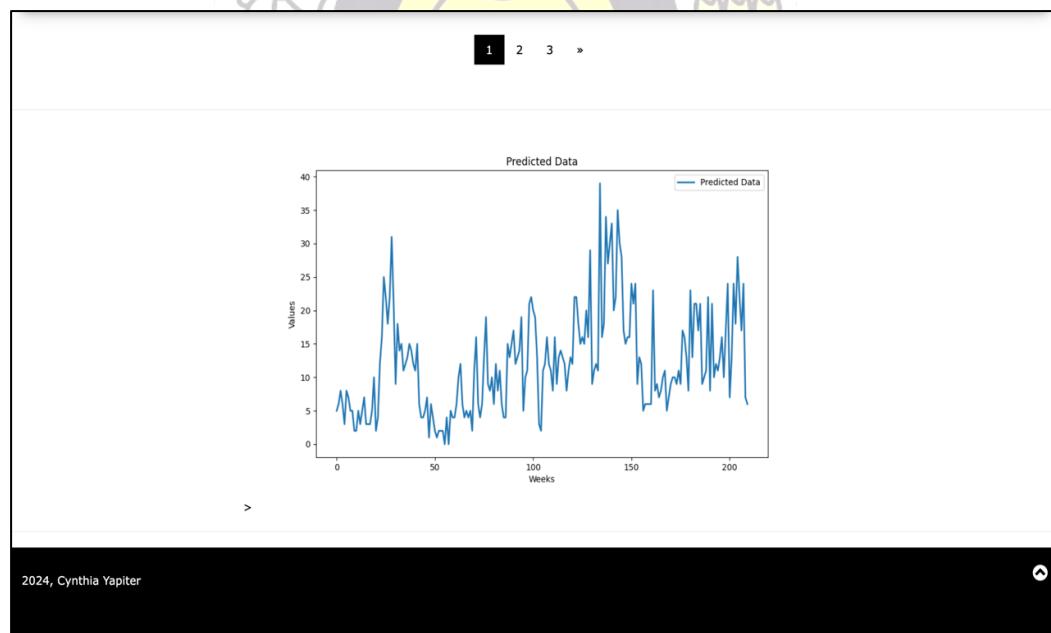
Table for predicted data

Date	Predicted
2018-12-31	5
2019-01-07	6
2019-01-14	8
2019-01-21	6
2019-01-28	3
2019-02-04	8
2019-02-11	7
2019-02-18	5
2019-02-25	5
2019-03-04	2

1 2 3 >

Gambar 4.1.7 Implementasi tampilan *result predict page table*

Selain tabel hasil, halaman ini juga menunjukkan plot hasil prediksi. Tampilan plot hasil prediksi ditunjukkan melalui gambar 4.1.8.

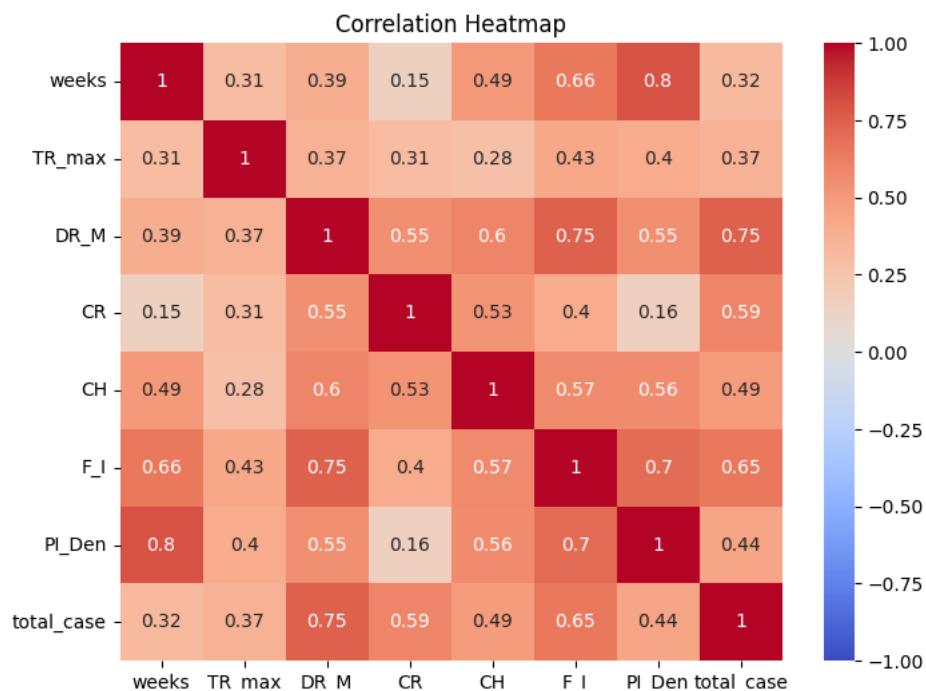


Gambar 4.1.8 Implementasi tampilan *result predict page plot*

## 4.2. Implementasi dan Pengujian Model Data Awal Tahun 2019-2022

### 4.2.1. Datasets

Pada bagian ini, data yang akan digunakan adalah data awal yang didapat dari penelitian sebelumnya yakni data kasus positif malaria di Kabupaten Batu Bara dengan rentang waktu dari tahun 2019 hingga tahun 2022. Pengecekan hubungan antar variabel dalam dataset dilakukan terlebih dahulu sebelum memproses model. Hal ini dilakukan dengan tujuan memastikan bahwa semua variabel yang digunakan efektif dalam menentukan target data. Hasil *correlation heatmap* pada data awal tertera pada gambar 4.2.1.



Gambar 4.2.1 *Correlation Heatmap* pada data awal

Berdasarkan gambar 4.2.1, ditemukan bahwa semua variabel berkorelasi positif. Hal ini menandakan bahwa semua variabel pada *dataset* sangatlah berpengaruh satu sama lain sehingga tidak ada variabel yang akan dieliminasi dari pelatihan model dalam penentuan data target prediksi.

Sebelum pembuatan model, eksperimen-eksperimen untuk mengoptimalkan model perlu dilakukan. Hal pertama yang dilakukan yakni pencarian rasio pembagian data *training* dan data *testing* terbaik. Hasil eksperimen ini ditunjukkan pada tabel 4.2.1.

Tabel 4.2.1 Eksperimen optimalisasi rasio pembagian data tahun 2019-2022

Ratio (train : test)	MAE	MSE	Training Score (R <sup>2</sup> )	Testing Score (R <sup>2</sup> )
6:4	4.453	40.506	0.896	0.471
7:3	5.354	50.087	0.913	-0.332
8:2	5.514	44.961	0.896	-0.196
9:1	3.485	22.704	0.875	0.505

Berdasarkan hasil eksperimen tabel 4.2.1, diketahui bahwa rasio pembagian data 9:1 memperoleh nilai MAE dan MSE terendah, yakni 3.485 dan 22.704 secara berurut. Perbandingan nilai uji MSE dari rasio 9:1 dengan rasio lainnya berbanding sangatlah jauh terutama pada rasio 7:3 yang menunjukkan 50.087. Selain itu pada rasio 9:1, nilai *training score* bukanlah yang tertinggi tetapi memiliki nilai *testing score* yang tertinggi. Nilai *training score* dan *testing score* menggunakan metriks perhitungan R<sup>2</sup>. Hal ini menunjukkan bahwa tidak terjadinya *overfitting* pada data dan membuat rasio ini sebagai rasio terbaik untuk model.

Selanjutnya dengan menggunakan rasio 9:1 sebagai rasio pembagian data *training* dan data *testing* terbaik, dilakukan optimalisasi parameter-parameter model. Parameter-parameter yang dimaksud yakni rasio pembagian data *training* dan data *testing*, *subsample*, *colsample\_bytree*, *learning\_rate*, *max\_depth*, dan *n\_estimators*. *Subsample* merupakan parameter yang mengatur persentase pemilihan acak pada data yang akan digunakan dalam pembangunan setiap pohon pada model XGBoost. Pemilihan acak ini memaksa setiap pohon untuk tidak bergantung pada seluruh data yang berfungsi untuk mencegah *overfitting*. *Colsample\_bytree* adalah parameter yang mengatur persentasi kolom atau fitur pada setiap pohon keputusan model. Parameter ini membatasi jumlah fitur yang digunakan dalam setiap *tree* dengan tujuan meningkatkan generalisasi dan juga mengurangi *overfitting*. *Learning\_rate* atau yang dikenal sebagai eta dalam XGBoost merupakan parameter yang menentukan besaran kontribusi dari setiap pohon yang ditambahkan ke model secara keseluruhan. *Max\_depth* merupakan parameter kedalaman maksimum setiap pohon. *N\_estimators* merupakan jumlah total pohon yang akan dibangun oleh model. Optimalisasi parameter-parameter ini dilakukan dengan metode *Grid Search*. Dengan metode ini, kita dapat mengetes semua kombinasi nilai parameter dan menemukan pilihan yang terbaik. Penggunaan metode *Grid Search* untuk optimalisasi data awal tertera pada gambar 4.2.2.

```
[ ] # hyperparameter tuning with grid search
from sklearn.model_selection import GridSearchCV

param_grid = {
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'max_depth': [2, 3, 4, 5],
    'subsample': [0.6, 0.7, 0.8, 0.9],
    'colsample_bytree': [0.6, 0.7, 0.8, 0.9],
    'n_estimators': [100, 500, 700, 1000]
}

grid_search = GridSearchCV(estimator=XGBRegressor(),
                           param_grid=param_grid,
                           cv=4, # 4-fold cross-validation
                           scoring=scoring,
                           refit='rmse',
                           verbose=1,
                           n_jobs=-1)
grid_search.fit(X_train, y_train)
```

→ Fitting 4 folds for each of 1024 candidates, totalling 4096 fits

- ▶ GridSearchCV
- ▶ estimator: XGBRegressor
  - ▶ XGBRegressor

Gambar 4.2.2 *Grid Search* pada data awal

Adapun pseudocode untuk metode *grid search* yakni:

*Function grid\_search\_hyperparameter():*

    Import GridSearchCV from sklearn.model\_selection

    Import XGBRegressor from xgboost

    Create instance for XGBRegressor

    Create hyperparameter grid

    For each hyperparameter in hyperparameter grid:

        Create all possible combinations

    End For

    For each hyperparameter combination:

        Set model parameters with current value

        Perform k-fold cross-validation

        compute average performance score for this combination

    End For

    Return the hyperparameter combination with the best performance

Hasil dari grid search kemudian dimasukkan ke dalam suatu tabel perbandingan yang diurut berdasarkan nilai evaluasi *Root Mean Square Error* (RMSE). Terdapat total sebanyak 1024 kombinasi nilai parameter yang diuji. Sebanyak 30 hasil kombinasi teratas ditampilkan pada tabel 4.2.2.

Tabel 4.2.2 Eksperimen optimalisasi parameter model pada data tahun 2019-2022

colsample_bytree	learning_rate	max_depth	n_estimators	subsample	mean_test_rmse	mean_test_mae	mean_test_mse	mean_test_r <sup>2</sup>
0,6	0,2	3	100	0,8	4.686	3.611	23.205	0.583
0,6	0,1	3	100	0,8	4.734	3.602	23.868	0.571
0,8	0,1	2	100	0,9	4.754	3.593	24.485	0.562
0,7	0,1	2	100	0,9	4.754	3.593	24.485	0.562
0,6	0,01	3	700	0,8	4.755	3.593	24.788	0.570
0,9	0,01	2	1000	0,9	4.764	3.578	24.835	0.559
0,6	0,2	3	100	0,7	4.768	3.616	23.858	0.572
0,9	0,01	2	700	0,9	4.769	3.581	24.794	0.559
0,6	0,01	3	500	0,8	4.773	3.612	24.914	0.568
0,7	0,2	5	100	0,6	4.775	3.728	24.450	0.568
0,8	0,2	5	100	0,6	4.775	3.728	24.450	0.568
0,6	0,2	3	100	0,9	4.776	3.679	24.544	0.565
0,6	0,01	4	500	0,8	4.780	3.619	24.990	0.567
0,6	0,01	3	700	0,9	4.782	3.594	24.971	0.566
0,8	0,2	5	700	0,6	4.782	3.742	24.487	0.566
0,7	0,2	5	700	0,6	4.782	3.742	24.487	0.566
0,7	0,2	5	1000	0,6	4.782	3.742	24.487	0.566
0,8	0,2	5	1000	0,6	4.782	3.742	24.487	0.566

colsample_bytree	learning_rate	max_depth	n_estimators	subsample	mean_test_rmse	mean_test_mae	mean_test_mse	mean_test_r^2
0,7	0,2	5	500	0,6	4.782	3.742	24.487	0.566
0,8	0,2	5	500	0,6	4.782	3.742	24.487	0.566
0,7	0,01	2	1000	0,9	4.783	3.587	25.051	0.558
0,8	0,01	2	1000	0,9	4.783	3.587	25.051	0.558
0,6	0,01	3	1000	0,8	4.784	3.616	25.104	0.564
0,6	0,01	4	700	0,8	4.789	3.616	25.084	0.565
0,7	0,01	2	700	0,9	4.791	3.596	25.074	0.558
0,8	0,01	2	700	0,9	4.791	3.596	25.074	0.558
0,6	0,2	3	500	0,8	4.797	3.709	24.196	0.561
0,6	0,2	3	700	0,8	4.798	3.709	24.200	0.560
0,6	0,2	3	1000	0,8	4.798	3.710	24.203	0.560
0,6	0,01	3	500	0,9	4.799	3.604	25.198	0.563

Berdasarkan hasil eksperimen tabel 4.2.2, kombinasi nilai parameter terbaik yakni *colsample\_bytree* sebesar 0.6, *learning\_rate* sebesar 0.2, *max\_depth* sebesar 3, *n\_estimators* sebesar 100, dan *subsample* dengan nilai 0.8. Kombinasi ini memiliki nilai evaluasi error terendah yakni RMSE sebesar 4.686, dan MSE sebesar 23.204. Nilai evaluasi lainnya yakni MAE sebesar 3.611 dan R<sup>2</sup> sebesar 0.583.

#### 4.2.2. Pengujian Model

Model dengan parameter-parameter yang telah dioptimalisasi akan diuji dengan data mingguan kasus positif malaria dengan rentang waktu dari tanggal 15 Agustus 2022 hingga 2 Januari 2023. Beberapa parameter dari hasil GridSearchCV secara acak akan dibandingkan dengan parameter terbaik dengan menggunakan data *testing*.

Tabel 4.2.3 Parameter hasil optimalisasi GridSearch dengan data *training* awal

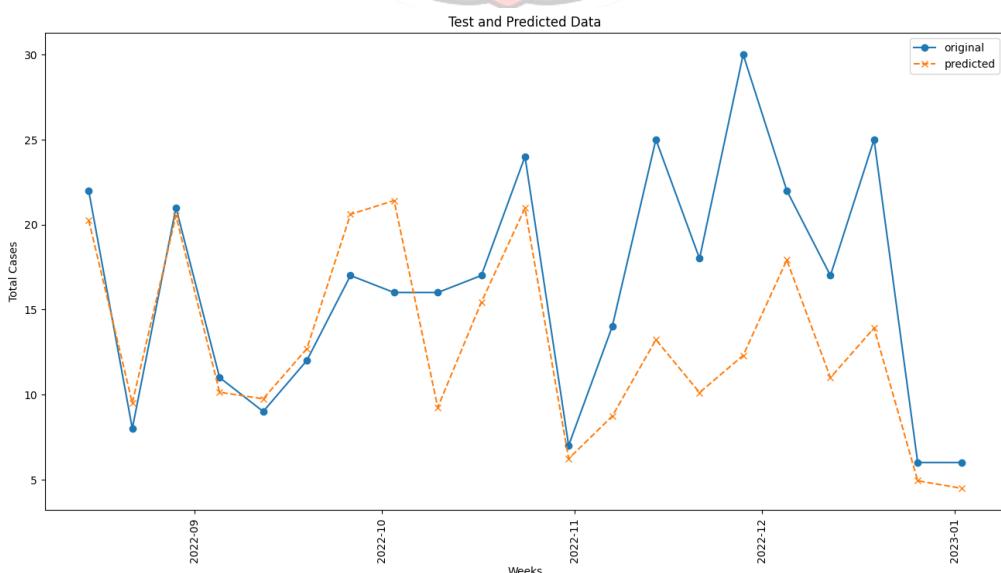
colsample_bytree	learning_rate	max_depth	n_estimators	subsample	mean_test_rmse	mean_test_mae	mean_test_mse	mean_test_r2
0,6	0,2	3	100	0,8	4.686	3.611	23.205	0.583
0,9	0,1	2	500	0,9	5.238	3.923	30.047	0.457
0,9	0,05	2	1000	0,9	5.219	3.963	29.788	0.461
0,8	0,1	2	500	0,7	5.430	4.177	31.663	0.420
0,8	0,2	2	1000	0,8	5.604	4.196	34.606	0.377

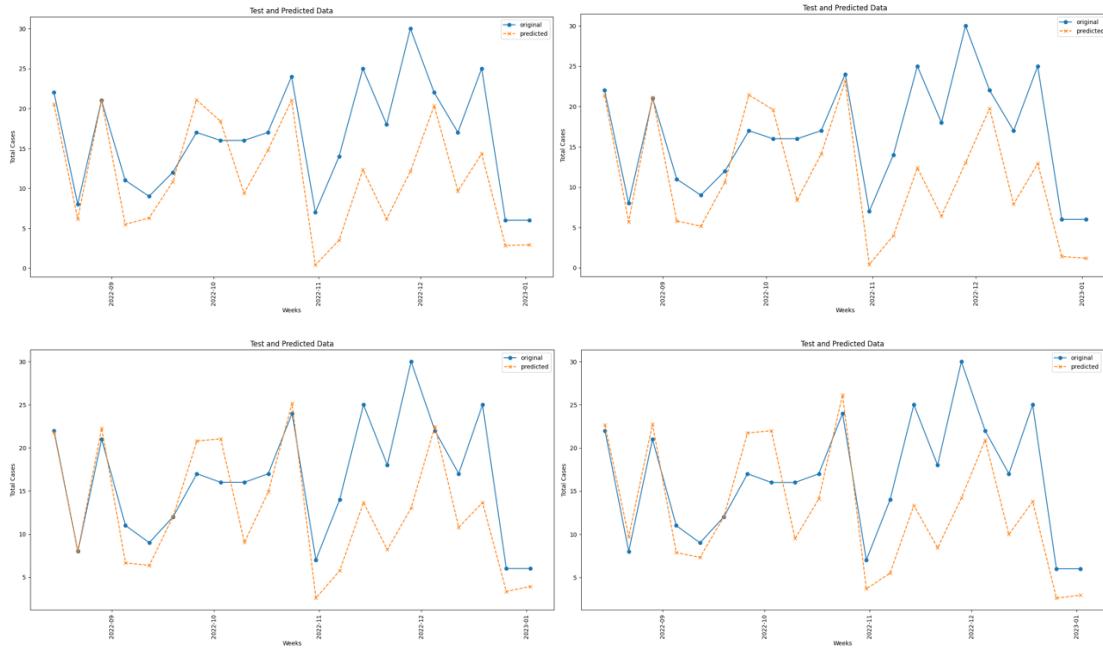
Tabel 4.2.4 Parameter hasil optimalisasi GridSearch dengan data *testing* awal

colsample_bytree	learning_rate	max_depth	n_estimators	subsample	rmse	mae	mse	R <sup>2</sup>
0,6	0,2	3	100	0,8	6.282	4.456	39.461	0.139
0,7	0,05	4	500	0,7	7.199	5.547	51.829	-0.131
0,9	0,05	2	1000	0,9	7.406	5.877	54.849	-0.196
0,8	0,1	2	500	0,7	6.551	4.817	42.921	0.064
0,8	0,2	2	1000	0,8	6.513	5.042	42.425	0.074

Perbedaan nilai hasil uji pada tabel 4.2.3 dan tabel 4.2.4 yakni dikarenakan pada tabel 4.2.3 menggunakan data *training* sedangkan pada tabel 4.2.4 menggunakan data *testing*. Oleh karena itu, nilai hasil uji yang didapatkan menjadi sedikit lebih rendah adalah hal yang normal.

Perbandingan hasil prediksi model dengan nilai asli dapat dilihat dalam bentuk plot pada gambar 4.2.3 dan gambar 4.2.4.

Gambar 4.2.3 Plot perbandingan nilai asli dengan prediksi *test* data 2019-2022



Gambar 4.2.4 Plot perbandingan nilai asli dengan prediksi *test data* 2019-2022 (2)

Berdasarkan hasil perbandingan plot hasil prediksi model, diketahui bahwa plot model dengan parameter terbaik terdapat pada gambar 4.2.3. Hasil prediksi model menggunakan parameter tersebut kemudian ditampilkan dengan nilai asli kejadian positif malaria sesuai rentang waktu di dalam sebuah tabel perbandingan. Tabel perbandingan nilai asli dengan nilai hasil prediksi model ditunjukkan pada tabel 4.2.5.

Tabel 4.2.5 Perbandingan nilai asli dan prediksi model pada *test data* 2019-2022

Weeks	Actual	Predicted
2022-08-15 00.00.00	22	20
2022-08-22 00.00.00	8	10
2022-08-29 00.00.00	21	21
2022-09-05 00.00.00	11	10
2022-09-12 00.00.00	9	10
2022-09-19 00.00.00	12	13
2022-09-26 00.00.00	17	21
2022-10-03 00.00.00	16	21
2022-10-10 00.00.00	16	9
2022-10-17 00.00.00	17	15
2022-10-24 00.00.00	24	21
2022-10-31 00.00.00	7	6
2022-11-07 00.00.00	14	9
2022-11-14 00.00.00	25	13

Weeks	Actual	Predicted
2022-11-21 00.00.00	18	10
2022-11-28 00.00.00	30	12
2022-12-05 00.00.00	22	18
2022-12-12 00.00.00	17	11
2022-12-19 00.00.00	25	14
2022-12-26 00.00.00	6	5
2023-01-02 00.00.00	6	4

#### 4.2.3. Evaluasi Model

Setelah melakukan evaluasi hasil prediksi, peneliti menemukan model yang paling optimal dengan rasio split data 9:1 dan parameter-parameter model XGBoost yakni `colsample_bytree` sebesar 0.6, `learning_rate` sebesar 0.2, `max_depth` sebesar 3, `n_estimators` sebesar 100, dan `subsample` sebesar 0.8. Peneliti menggunakan 4 perhitungan evaluasi model yaitu *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Mean Square Error* (MSE), serta *Coefficient of Determination* ( $R^2$ ). Adapun nilai hasil evaluasi yakni 4.456, 6.282, 39.461 dan 0.139 secara berurut. Data yang dievaluasi adalah data yang diuji yakni dari rentang waktu 15 Agustus 2022 hingga 2 Januari 2023. Penjabaran rumus keempat evaluasi *metrics* yang digunakan dapat ditinjau pada bagian landasan teori.

### 4.3. Implementasi dan Pengujian Model Data Gabungan Tahun 2015-2022

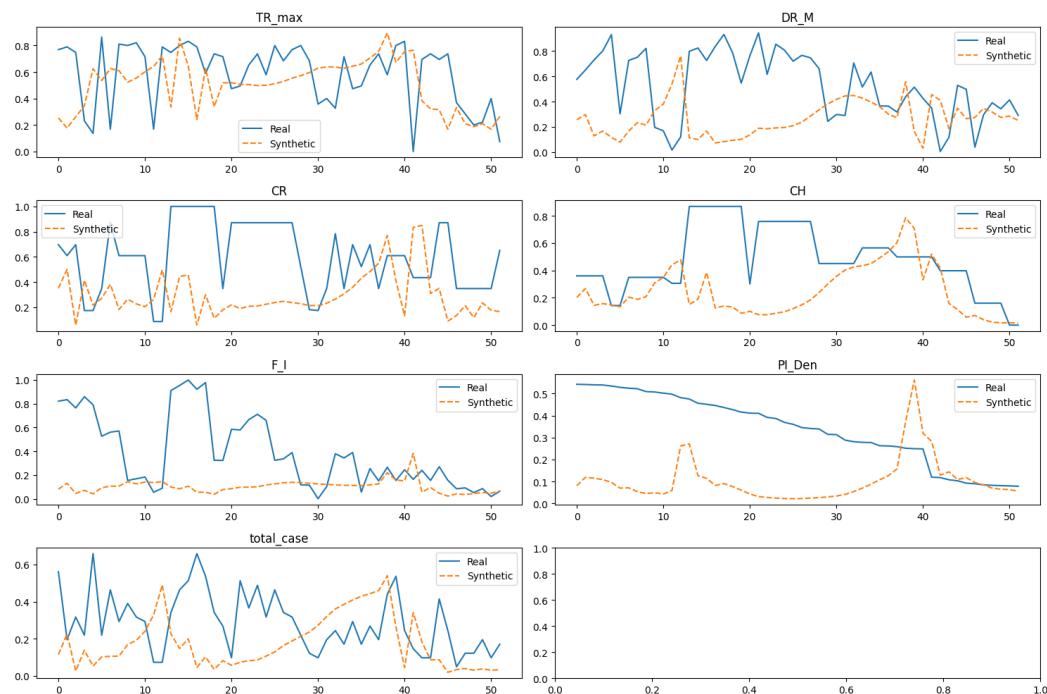
#### 4.3.1. Datasets

Data yang digunakan pada bagian ini adalah gabungan data awal yang diperoleh dari penelitian sebelumnya yakni data kasus positif malaria di Kabupaten Batu Bara dengan rentang waktu dari tahun 2019 hingga tahun 2022 dengan data sintesis tahun 2015-2018 yang di-generate dengan metode TimeGAN. Hal ini menjadikan adanya peningkatan pada jumlah data yang digunakan dengan total rentang waktu selama 8 tahun. Tentunya dalam generasi data sintesis dengan TimeGAN, peneliti melakukan optimalisasi parameter yakni pada `seq_len` dan `batch_size`.

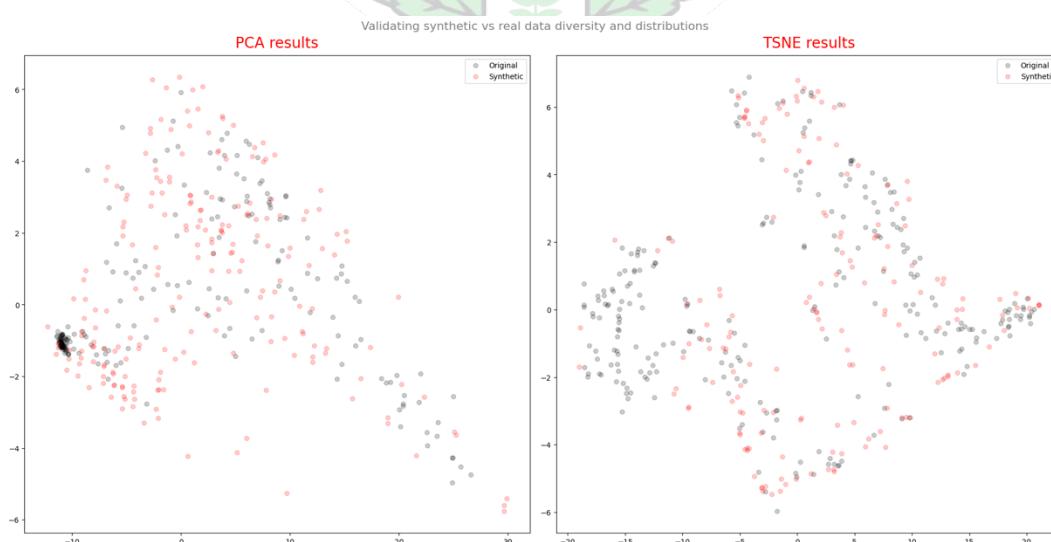
*Sequence Length (seq\_len)* adalah panjang dari setiap *sequence* yang dipelajari model. Contohnya jika `seq_len` bernilai 24, maka model memproses 24 langkah waktu secara bersamaan yang dimana mewakili data 24 jam jikalau data yang digunakan merupakan data per jam. *Batch\_size* adalah jumlah *sequence* yang diproses secara

bersamaan dalam satu iterasi pelatihan. Kedua parameter ini menentukan bagaimana TimeGAN mempelajari ketergantungan temporal dan generalisasi pola pada data. Peneliti melakukan eksperimen dengan menentukan nilai parameter yang hendak diuji, yakni 24 dan 52 untuk *seq\_len* dan 128 dan 64 untuk *batch\_size*.

Plot perbandingan nilai asli dengan hasil sintesis, hasil PCA dan TSNE dan hasil evaluasi kombinasi *seq\_len* 52 dan *batch\_size* 64 dapat dilihat pada gambar 4.3.1, gambar 4.3.2 dan tabel 4.3.1.



Gambar 4.3.1 Plot perbandingan nilai sintesis *seq\_len* 52 dan *batch\_size* 64

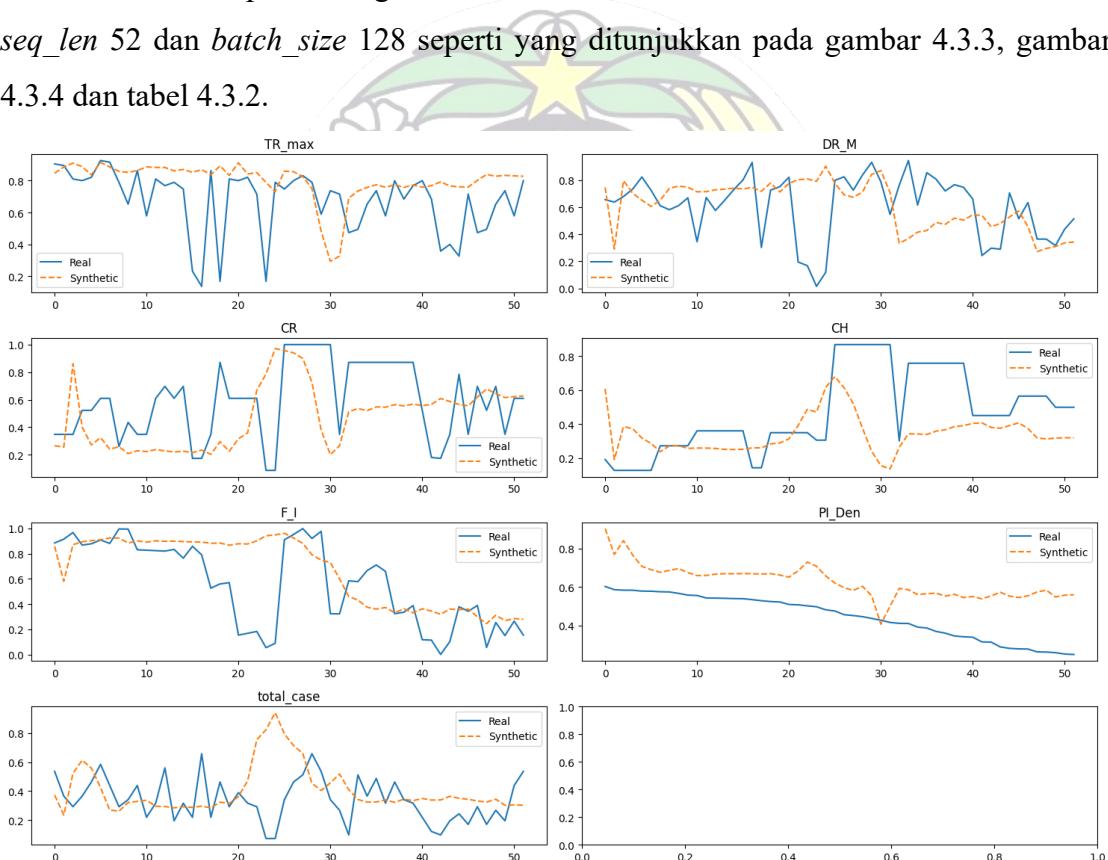


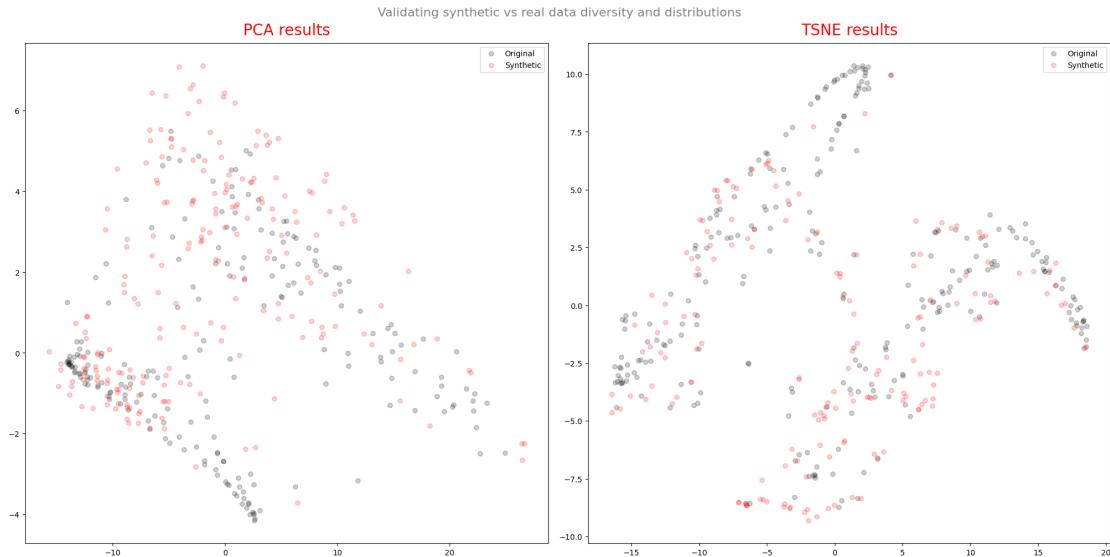
Gambar 4.3.2 hasil PCA dan TSNE *seq\_len* 52 dan *batch\_size* 64

Tabel 4.3.1 Hasil evaluasi *seq\_len* 52 dan *batch\_size* 64

	R <sup>2</sup>	MAE	MRLE
real	0.331546	0.142330	0.01744
synthetic	0.350317	0.141436	0.01769

Ditemukan dari hasil eksperimen bahwa nilai kombinasi optimal dari parameter *seq\_len* dan *batch\_size* adalah 52 dan 128 secara berurut. Alasan terpilihnya kombinasi nilai tersebut tergambar pada plot perbandingan nilai asli dan sintesis dan juga hasil PCA dan TSNE. Selain itu, perbedaan nilai evaluasi pada kombinasi ini merupakan yang terkecil yang dimana menunjukkan bahwa kemungkinan terjadinya *overfitting* lebih minim. Plot perbandingan, hasil PCA dan TSNE dan hasil evaluasi kombinasi *seq\_len* 52 dan *batch\_size* 128 seperti yang ditunjukkan pada gambar 4.3.3, gambar 4.3.4 dan tabel 4.3.2.

Gambar 4.3.3 *Plot* perbandingan nilai sintesis *seq\_len* 52 dan *batch\_size* 128

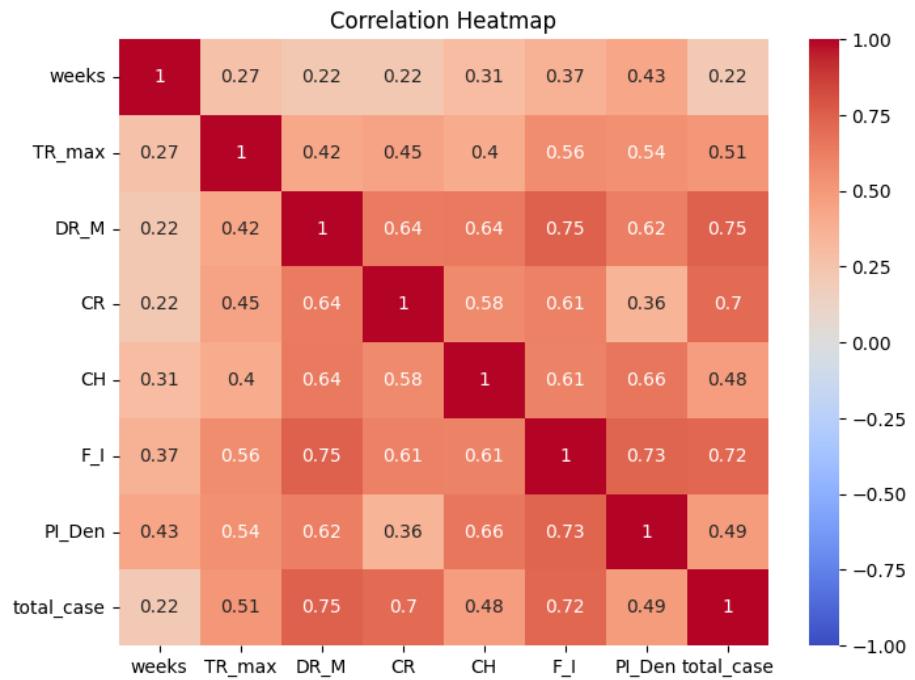


Gambar 4.3.4 hasil PCA dan TSNE *seq\_len* 52 dan *batch\_size* 128

Tabel 4.3.2 Hasil evaluasi *seq\_len* 52 dan *batch\_size* 128

	$R^2$	MAE	MRLE
real	0.339447	0.154485	0.021455
synthetic	0.324727	0.161474	0.021296

Dengan menggunakan timeGAN dengan parameter yang telah dioptimalkan, peneliti memperoleh data sintesis dari tahun 2015-2018 yang kemudian digabung dengan data awal. Adapun *correlation heatmap* dari data gabungan tersebut terdapat pada gambar 4.3.5.



Gambar 4.3.5 Correlation Heatmap data 2015-2022

Berdasarkan gambar 4.3.5, diketahui bahwa semua variabel berkorelasi positif. Hal ini menandakan bahwa semua variabel pada *dataset* sangatlah berpengaruh satu sama lain. Oleh karena itu, penulis memutuskan untuk menggunakan semua variabel untuk pengembangan model.

Selanjutnya, dilakukan optimalisasi parameter-parameter model seperti rasio pembagian data *training* dan data *testing*, *subsample*, *colsample\_bytree*, *learning\_rate*, *max\_depth*, dan *n\_estimators*. Optimalisasi pertama yakni eksperimen pencarian rasio pembagian data *training* dan data *testing* terbaik. Hasil eksperimen ini ditunjukkan pada tabel 4.3.3.

Tabel 4.3.3 Eksperimen optimalisasi rasio pembagian data tahun 2015-2022

Ratio (train : test)	MAE	MSE	Training Score ( $R^2$ )	Testing Score ( $R^2$ )
6:4	4.034	30.347	0.999	0.554
7:3	4.39	32.579	0.999	0.495
8:2	4.361	32.033	0.999	0.582
9:1	4.72	36.432	0.999	0.031

Berdasarkan hasil eksperimen tabel 4.3.3, diketahui bahwa rasio pembagian data 6:4 memperoleh nilai MAE dan MSE terendah, yakni 4.034 dan 30.347 secara berurut. Perbandingan nilai uji MSE dari rasio 6:4 dengan nilai rasio lainnya berbanding sangatlah jauh terutama pada rasio 9:1 yang menunjukkan 36.432. Nilai *training score* dan *testing score* menggunakan metriks perhitungan  $R^2$  yang dimana berarti semakin tinggi dan mendekati 1 nilainya maka semakin bagus. Pada rasio 6:4, nilai *training score* termasuk salah satu yang tertinggi dan memiliki nilai *testing score* yang stabil tingginya juga. Hal ini menunjukkan bahwa tidak terjadinya *overfitting* pada data dan membuat rasio ini sebagai rasio terbaik untuk pembuatan model.

Optimalisasi parameter-parameter XGBoost selanjutnya dilakukan dengan metode *Grid Search*. Penggunaan metode ini dikarenakan kemampuan mencoba semua kombinasi nilai parameter dan menemukan nilai kombinasi parameter terbaik. Hasil dari grid search kemudian dimasukkan ke dalam suatu tabel perbandingan yang diurut berdasarkan nilai evaluasi *Mean Absolute Error* (MAE). Terdapat total sebanyak 1024 kombinasi nilai parameter yang diuji. Total 30 hasil kombinasi terbaik ditampilkan pada tabel 4.3.4.

Tabel 4.3.4 Eksperimen optimalisasi parameter model pada data tahun 2015-2022

colsample_bytree	learning_rate	max_depth	n_estimators	subsample	mean_test_rmse	mean_test_mae	mean_test_mse	mean_test_r2
0,9	0,2	5	100	0,6	4,068	3,154	17,298	0,772
0,9	0,2	5	500	0,6	4,075	3,155	17,370	0,771
0,9	0,2	5	700	0,6	4,075	3,155	17,370	0,771
0,9	0,2	5	1000	0,6	4,075	3,155	17,370	0,771
0,7	0,1	3	100	0,6	4,152	3,103	18,435	0,760
0,8	0,1	3	100	0,6	4,152	3,103	18,435	0,760
0,9	0,1	4	100	0,6	4,183	3,187	18,582	0,761
0,9	0,1	3	100	0,6	4,229	3,161	19,180	0,752
0,7	0,1	3	100	0,7	4,260	3,239	19,156	0,752
0,8	0,1	3	100	0,7	4,260	3,239	19,156	0,752
0,6	0,2	5	100	0,8	4,273	3,189	19,350	0,755
0,9	0,1	4	500	0,6	4,274	3,297	19,320	0,751
0,9	0,1	4	700	0,6	4,277	3,299	19,347	0,751
0,9	0,1	4	1000	0,6	4,277	3,300	19,352	0,751
0,6	0,1	5	100	0,6	4,277	3,243	18,832	0,751
0,6	0,2	5	500	0,8	4,284	3,199	19,445	0,754

<i>colsample_bytree</i>	<i>learning_rate</i>	<i>max_depth</i>	<i>n_estimators</i>	<i>subsample</i>	<i>mean_test_rmse</i>	<i>mean_test_mae</i>	<i>mean_test_mse</i>	<i>mean_test_r2</i>
0,6	0,2	5	700	0,8	4,284	3,199	19,445	0,754
0,6	0,2	5	1000	0,8	4,284	3,199	19,445	0,754
0,9	0,01	3	1000	0,6	4,288	3,193	19,567	0,745
0,9	0,05	3	500	0,6	4,291	3,318	19,428	0,746
0,7	0,1	3	500	0,6	4,293	3,314	19,503	0,745
0,8	0,1	3	500	0,6	4,293	3,314	19,503	0,745
0,8	0,1	2	500	0,6	4,294	3,272	19,684	0,743
0,7	0,1	2	500	0,6	4,294	3,272	19,684	0,743
0,6	0,1	2	100	0,7	4,297	3,130	19,600	0,744
0,9	0,01	3	700	0,6	4,299	3,173	19,656	0,744
0,6	0,2	5	100	0,7	4,300	3,200	18,998	0,744
0,7	0,1	3	700	0,6	4,307	3,326	19,627	0,743
0,8	0,1	3	700	0,6	4,307	3,326	19,627	0,743
0,6	0,2	5	500	0,7	4,314	3,216	19,129	0,742

Berdasarkan hasil eksperimen tabel 4.3.4, kombinasi nilai parameter terbaik yakni *colsample\_bytree* sebesar 0,9, *learning\_rate* sebesar 0,2, *max\_depth* sebesar 5, *n\_estimators* sebesar 100, dan *subsample* dengan nilai 0,6. Kombinasi ini memiliki nilai evaluasi error terendah yakni RMSE sebesar 4.068, dan MSE sebesar 17.298. Nilai evaluasi lainnya yakni MAE sebesar 3.154 dan R<sup>2</sup> sebesar 0,772.

#### 4.3.2. Pengujian Model

Model beserta parameter-parameter yang telah dioptimalisasi selanjutnya diuji dengan data kasus positif malaria mingguan dengan rentang waktu dari tanggal 21 Oktober 2019 hingga 2 Januari 2023. Beberapa parameter dari hasil GridSearchCV secara acak akan dibandingkan dengan parameter terbaik dengan menggunakan data *testing*.

Tabel 4.3.5 Parameter optimalisasi GridSearch dengan data *training* gabungan

<i>colsample_bytree</i>	<i>learning_rate</i>	<i>max_depth</i>	<i>n_estimators</i>	<i>subsample</i>	<i>mean_test_rmse</i>	<i>mean_test_mae</i>	<i>mean_test_mse</i>	<i>mean_test_r2</i>
0,9	0,2	5	100	0,6	4,068	3,154	17,298	0,772
0,7	0,2	2	500	0,6	4.767	3.694	24.829	0.6814
0,9	0,01	4	100	0,8	5.429	4.533	30.696	0.599
0,9	0,01	2	100	0,7	6.016	4.926	37.298	0.500

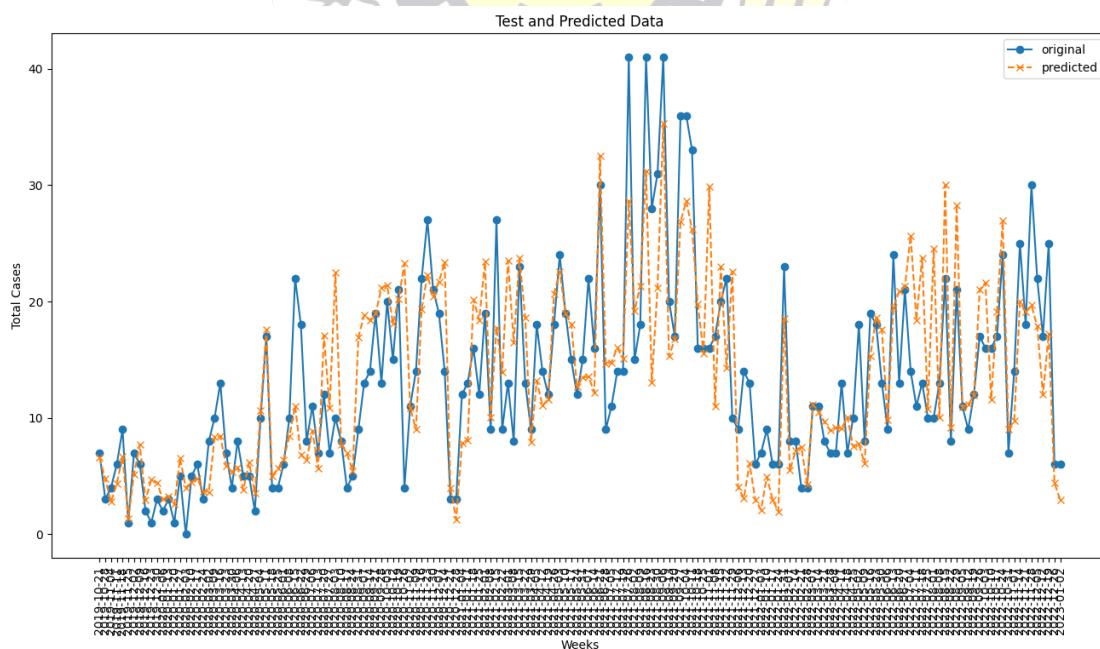
colsam ple_byt ree	learnin g_rate	max_ depth	n_estim ators	subsa mple	mean _test_ rmse	mean _test_ mae	mean_t est_mse	mean_t est_r2
0,9	0,01	2	100	0,9	6.073	4.968	38.063	0.490

Tabel 4.3.6 Parameter optimalisasi GridSearch dengan data *testing* gabungan

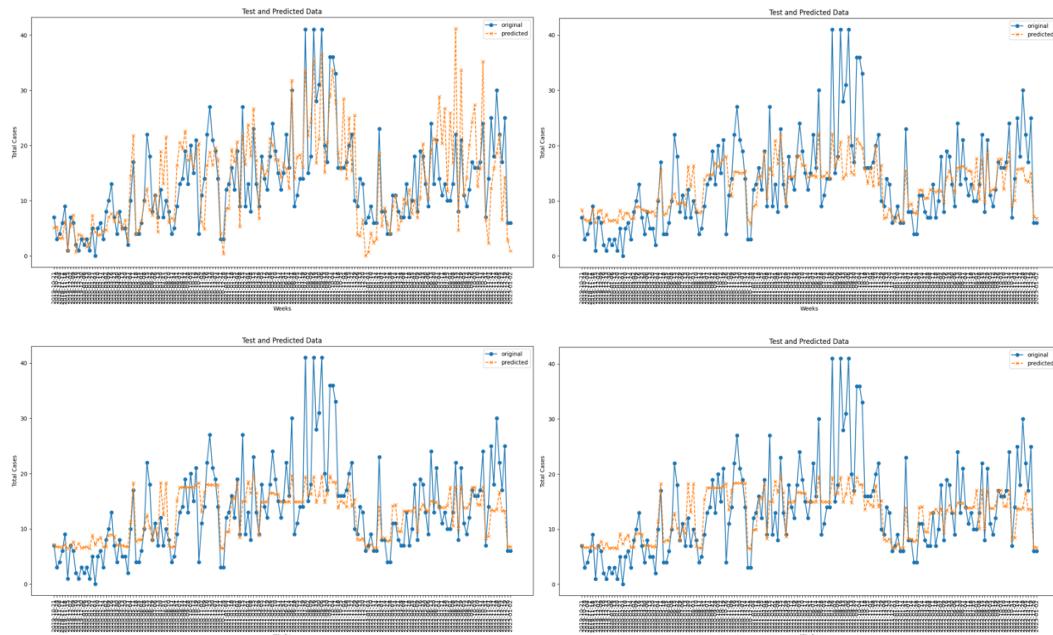
colsam ple_byt ree	learnin g_rate	max_ depth	n_estim ators	subsa mple	rmse	mae	mse	R <sup>2</sup>
0,9	0,2	5	100	0,6	5,363	3,958	28,762	0,578
0,7	0,2	2	500	0,6	5,977	4,558	35,726	0,475
0,9	0,01	4	100	0,8	5,639	4,039	31,793	0,533
0,9	0,01	2	100	0,7	5,920	4,187	35,042	0,486
0,9	0,01	2	100	0,9	5,930	4,190	35,164	0,484

Perbedaan nilai hasil uji pada tabel 4.3.5 dan tabel 4.3.6. yakni dikarenakan pada tabel 4.3.5 menggunakan data training sedangkan pada tabel 4.3.6 menggunakan data testing. Oleh karena itu, nilai hasil uji yang didapatkan menjadi sedikit lebih rendah adalah hal yang normal.

Perbandingan hasil prediksi model dengan nilai asli dapat dilihat dalam bentuk plot pada gambar 4.3.6 dan gambar 4.3.7.



Gambar 4.3.6 Plot perbandingan nilai asli dengan prediksi *test* data 2015-2022 (1)



Gambar 4.3.7 Plot perbandingan nilai asli dengan prediksi *test* data 2015-2022 (2)

Berdasarkan hasil perbandingan plot hasil prediksi model, diketahui bahwa plot model dengan parameter terbaik terdapat pada gambar 4.3.6. Hasil prediksi model menggunakan parameter terbaik kemudian ditampilkan dengan nilai asli kejadian positif malaria sesuai rentang waktu di dalam sebuah tabel perbandingan. Tabel perbandingan nilai asli dengan nilai hasil prediksi model tersebut ditunjukkan pada tabel 4.3.7.

Tabel 4.3.7 Perbandingan nilai asli dan prediksi model pada *test* data 2015-2022

Weeks	Actual	Predicted
2019-10-21	7	7
2019-10-28	3	5
2019-11-04	4	3
2019-11-11	6	4
2019-11-18	9	7
2019-11-25	1	1
2019-12-02	7	5
2019-12-09	6	8
2019-12-16	2	3
2019-12-23	1	5
2019-12-30	3	4
2020-01-06	2	3
2020-01-13	3	3

Weeks	Actual	Predicted
2020-01-20	1	3
2020-01-27	5	7
2020-02-03	0	4
2020-02-10	5	4
2020-02-17	6	5
2020-02-24	3	4
2020-03-02	8	4
2020-03-09	10	8
2020-03-16	13	8
2020-03-23	7	6
2020-03-30	4	5
2020-04-06	8	6
2020-04-13	5	4
2020-04-20	5	6
2020-04-27	2	4
2020-05-04	10	11
2020-05-11	17	18
2020-05-18	4	5
2020-05-25	4	6
2020-06-01	6	6
2020-06-08	10	8
2020-06-15	22	11
2020-06-22	18	7
2020-06-29	8	6
2020-07-06	11	9
2020-07-13	7	6
2020-07-20	12	17
2020-07-27	7	11
2020-08-03	10	22
2020-08-10	8	8
2020-08-17	4	7
2020-08-24	5	5
2020-08-31	9	17
2020-09-07	13	19
2020-09-14	14	18
2020-09-21	19	19
2020-09-28	13	21
2020-10-05	20	21
2020-10-12	15	18
2020-10-19	21	20

Weeks	Actual	Predicted
2020-10-26	4	23
2020-11-02	11	11
2020-11-09	14	9
2020-11-16	22	19
2020-11-23	27	22
2020-11-30	21	20
2020-12-07	19	22
2020-12-14	14	23
2020-12-21	3	4
2020-12-28	3	1
2021-01-04	12	8
2021-01-11	13	8
2021-01-18	16	20
2021-01-25	12	18
2021-02-01	19	23
2021-02-08	9	10
2021-02-15	27	18
2021-02-22	9	14
2021-03-01	13	24
2021-03-08	8	17
2021-03-15	23	24
2021-03-22	13	19
2021-03-29	9	8
2021-04-05	18	13
2021-04-12	14	11
2021-04-19	12	12
2021-04-26	18	21
2021-05-03	24	23
2021-05-10	19	19
2021-05-17	15	18
2021-05-24	12	13
2021-05-31	15	13
2021-06-07	22	14
2021-06-14	16	12
2021-06-21	30	33
2021-06-28	9	15
2021-07-05	11	15
2021-07-12	14	16
2021-07-19	14	15
2021-07-26	41	29

Weeks	Actual	Predicted
2021-08-02	15	19
2021-08-09	18	21
2021-08-16	41	31
2021-08-23	28	13
2021-08-30	31	21
2021-09-06	41	35
2021-09-13	20	15
2021-09-20	17	17
2021-09-27	36	27
2021-10-04	36	29
2021-10-11	33	26
2021-10-18	16	20
2021-10-25	16	16
2021-11-01	16	30
2021-11-08	17	11
2021-11-15	20	23
2021-11-22	22	14
2021-11-29	10	23
2021-12-06	9	4
2021-12-13	14	3
2021-12-20	13	6
2021-12-27	6	3
2022-01-03	7	2
2022-01-10	9	5
2022-01-17	6	3
2022-01-24	6	2
2022-01-31	23	19
2022-02-07	8	5
2022-02-14	8	7
2022-02-21	4	7
2022-02-28	4	4
2022-03-07	11	11
2022-03-14	11	11
2022-03-21	8	10
2022-03-28	7	9
2022-04-04	7	9
2022-04-11	13	9
2022-04-18	7	10
2022-04-25	10	8
2022-05-02	18	8

Weeks	Actual	Predicted
2022-05-09	8	6
2022-05-16	19	15
2022-05-23	18	19
2022-05-30	13	18
2022-06-06	9	10
2022-06-13	24	20
2022-06-20	13	21
2022-06-27	21	21
2022-07-04	14	26
2022-07-11	11	18
2022-07-18	13	24
2022-07-25	10	11
2022-08-01	10	25
2022-08-08	13	10
2022-08-15	22	30
2022-08-22	8	9
2022-08-29	21	28
2022-09-05	11	11
2022-09-12	9	11
2022-09-19	12	12
2022-09-26	17	21
2022-10-03	16	22
2022-10-10	16	12
2022-10-17	17	19
2022-10-24	24	27
2022-10-31	7	9
2022-11-07	14	10
2022-11-14	25	20
2022-11-21	18	19
2022-11-28	30	20
2022-12-05	22	18
2022-12-12	17	12
2022-12-19	25	17
2022-12-26	6	4
2023-01-02	6	3

#### 4.3.3. Evaluasi Model

Setelah melakukan evaluasi hasil prediksi, peneliti menemukan model yang paling optimal dengan rasio split data 6:4 dan parameter-parameter model XGBoost yakni `colsample_bytree` sebesar 0.9, `learning_rate` sebesar 0.2, `max_depth` sebesar 5, `n_estimators` sebesar 100, dan `subsample` sebesar 0.6. Evaluasi model dilakukan dengan menggunakan 4 perhitungan seperti pada data awal yaitu *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), *Mean Square Error* (MSE), serta *Coefficient of Determination* ( $R^2$ ). Adapun nilai hasil evaluasi yakni 3.958, 5.363, 28.762 dan 0.578 secara berurut. Data yang dievaluasi adalah data yang diuji yakni dari rentang waktu 5 Januari 2015 hingga 2 Januari 2023. Adapun penjabaran rumus keempat evaluasi yang digunakan dapat ditinjau pada bagian landasan teori.

#### 4.4. Diskusi

Penelitian ini mengimplementasi metode *Extreme Gradient Boosting* (XGBoost) dalam memprediksi kasus positif malaria di Kabupaten Batu Bara dengan memanfaatkan faktor meteorologi. Dengan tujuan memperoleh hasil akhir model paling optimal, penulis melakukan beberapa uji coba pada hyperparameter XGBoost yakni `subsample`, `colsample_bytree`, `learning_rate`, `max_depth`, dan `n_estimators`. Peneliti juga melakukan percobaan dalam menentukan rasio pembagian data *training* dan *testing* yang paling sesuai. Hasil prediksi yang didapat yaitu model dengan hasil evaluasi nilai train  $R^2$  sebesar 0.99 dan nilai test  $R^2$ , MAE, RMSE dan MSE sebesar 0.578, 3.958, 5.363 dan 28.762 secara berturut-turut. Hasil plot hasil prediksi dan data aktual juga menunjukkan bahwa model dapat mengikuti tren kenaikan dan penurunan pola kejadian kasus positif malaria dengan cukup baik.

Perbedaan nilai yang terjadi dikarenakan jumlah data training yang tergolong sedikit untuk pelatihan model XGBoost tetapi hal ini yang menjadi salah satu *challenge* yang dicoba pada penelitian ini. Berhubung data awal yang hendak digunakan hanya terdiri dari 4 tahun, penambahan data sintesis diperlukan untuk menghindari baik terjadinya *underfitting* maupun *overfitting*. Penambahan data sintesis untuk 4 tahun lainnya yakni dari tahun 2015 hingga 2018 dilakukan dengan *Time-series Generative Adversarial Network* (TimeGAN). Hasil penggunaan data gabungan ini terbukti berhasil meningkatkan kinerja model seperti yang dapat dilihat dari perbandingan hasil evaluasi kedua percobaan tersebut.

Terdeteksi adanya tren kejadian kasus positif malaria pada beberapa periode waktu tertentu dari data *test* yang digunakan. Hal ini dikarenakan pengaruh kuat jumlah kasus positif terhadap faktor iklim dan lingkungan. Tren peningkatan ini terjadi pada akhir 2020, pertengahan hingga akhir tahun 2021, dan akhir tahun 2022.

Pada penelitian selanjutnya, peneliti merekomendasikan penggunaan data yang lebih banyak untuk meningkatkan akurasi model lebih jauh lagi. Eksplorasi metode yang digunakan dalam *hyperparameter tuning* juga sangat disarankan. Hal ini tidak terbatas pada *hyperparameter* yang diuji. *Colsample\_bylevel*, *early\_stopping\_rounds*, dan *min\_child\_weight* merupakan contoh *hyperparameter-hyperparameter* lain yang bisa diuji untuk membantu mencegah *overfitting* secara lebih baik lagi. Penelitian ini secara umum memberikan kontribusi penting dalam upaya memprediksi kasus positif malaria berdasarkan faktor iklim. Walaupun demikian, terdapat peluang untuk penyempurnaan lebih lanjut, terutama terkait dengan peningkatan jumlah data dan optimalisasi proses *hyperparameter tuning* pada model.



## **BAB 5** **KESIMPULAN DAN SARAN**

### **5.1. Kesimpulan**

Kesimpulan yang dapat ditarik dari penelitian prediksi kejadian malaria menggunakan metode *Extreme Gradient Boosting* (XGBoost) yakni sebagai berikut:

1. Prediksi kejadian malaria menggunakan metode XGBoost dan data iklim mendapatkan nilai MAE, MSE, RMSE dan R<sup>2</sup> sebesar 4.456, 39.461, 6.282 dan 0.139 secara berturut-turut.
2. Penambahan data sintesis dengan menggunakan *Time-series Generative Adversarial Networks* (TimeGAN) melalui penggunaan salah satu *library python* yaitu *ydata-synthetic* terbukti sangatlah membantu dalam meningkatkan performa model prediksi. Hal ini dibuktikan dengan peningkatan performa nilai uji yakni MAE, MSE, RMSE dan R<sup>2</sup> sebesar 3.958, 28.762, 5.363 dan 0.578 secara berturut-turut.
3. Data iklim seperti curah hujan, suhu, kelembapan dan juga data lingkungan seperti kepadatan penduduk dan tingkat genangan air berperan sebagai penentu angka kejadian malaria di Kabupaten Batu Bara.
4. Algoritma XGBoost terbukti dapat digunakan dalam memprediksi data dengan jumlah data yang sedikit dan terbatas seperti data malaria tanpa terjadinya overfitting. Meskipun data dengan jumlah yang lebih banyak tentunya akan menghasilkan nilai yang lebih bagus.

### **5.2. Saran**

Sebagai bahan pertimbangan penelitian selanjutnya, berikut saran-saran yang dianjurkan peneliti:

1. Penggunaan data historikal kejadian malaria yang lebih banyak dalam pelatihan model agar dapat lebih jauh meningkatkan akurasi dan mengurangi tingkat *error*.
2. Eksplorasi lebih lanjut terhadap metode dan *hyperparameter* yang digunakan pada *hyperparameter tuning* untuk meningkatkan kinerja model, efisiensi waktu yang digunakan sebelum pembuatan model dan mencegah *overfitting* secara lebih baik lagi.

3. Pertimbangan dalam penggunaan data pendukung selain faktor iklim seperti tingkat wilayah, indeks perpindahan penduduk, faktor sosial-ekonomi, kebersihan lingkungan dan faktor-faktor lainnya yang berkaitan dengan kejadian malaria.
4. Penggunaan data dari wilayah lain agar penggunaan model dalam memprediksi kejadian malaria tidak terbatas pada suatu wilayah saja.



## DAFTAR PUSTAKA

- Avichena, A. and Anggriyani, R. (2023) ‘Analisis Penyakit Malaria Akibat Infeksi Plasmodium sp. terhadap Darah Manusia’, EKOTONIA: Jurnal Penelitian Biologi, Botani, Zoologi dan Mikrobiologi, 8(1), pp. 30–37.
- Ahmad, H. I., Prasad, R., Sharma, B. K., Madaki, A. Y., & Shuaibu, A. R. (2023). Malaria Disease Prediction Using Frequency-Based Machine Learning Algorithms and Ensembles Algorithms. 2023 2nd International Conference on Multidisciplinary Engineering and Applied Science, ICMEAS 2023. Available at: <https://doi.org/10.1109/ICMEAS58693.2023.10379206>
- Alkaff, M., Baskara, A., & Ainiyyah, A. (2023). Penerapan Metode XGBoost Untuk Memprediksi Jumlah Kejadian Kecelakaan Lalu Lintas di Kota Banjarmasin. Generation Journal, 7(1). Available at: <https://doi.org/10.29407/gj.v7i1.19807>
- Apriliana (2017). Pengaruh Iklim terhadap Insidens Malaria di Provinsi Lampung. Cermin Dunia Kedokteran, 44(7), pp.464-470.
- Apriliana (2021). ANALISIS FAKTOR RISIKO KEJADIAN MALARIA DI INDONESIA (Analisis Data Riskesdas 2018). Data Riskesdas 2018, pp. 10–25. Available at: <http://repository.uinsu.ac.id/id/eprint/15340>.
- Badan Pusat Statistik Kabupaten Batu Bara. (n.d.). Jumlah curah hujan dan hari hujan Kabupaten Batu Bara. Diakses dari <https://batubarabakab.bps.go.id/id/statistics-table/2/MTA2IzI=/jumlah-curah-hujan-dan-hari-hujan-kabupaten-batu-barabara.html>
- BPKP. (2024). Diakses pada 29 April 2024. <https://www.bpkp.go.id/sumut/konten/236/>
- Dinas Komunikasi dan Informatika Provinsi Sumatera Utara. (2023). Statistik Sektoral Provinsi Sumatera Utara.
- Chen, T., & Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
- Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, 35(4), 3173–3190. <https://doi.org/10.1007/s00521-022-07856-4>

- Fahmi, F., Pasaribu, A. P., Theodora, M., & Wangdi, K. (2022). Spatial analysis to evaluate risk of malaria in Northern Sumatera, Indonesia. *Malaria Journal*, 21(1). Available at: <https://doi.org/10.1186/s12936-022-04262-y>
- Fadli, R. (2023, November 17). Malaria. HaloDoc. Diakses pada 24 April 2024. <https://www.halodoc.com/kesehatan/malaria>
- Fitri, H., & Adlina, Z. (2023). Traces of Social History in Batu Bara Songket Traditional Crafts in Batubara Regency, North Sumatra Province. *Historical Studies Journal*, 7(1), 86–101. <http://jurnal.fkip.unmul.ac.id/index.php/yupa>
- Fischer, L., Güttekin, N., Kaelin, M. B., Fehr, J., & Schlagenhauf, P. (2020). Rising temperature and its impact on receptivity to malaria transmission in Europe: A systematic review. In *Travel Medicine and Infectious Disease* (Vol. 36). Elsevier USA. Available at: <https://doi.org/10.1016/j.tmaid.2020.101815>
- Guo R., Zhao Z., Wang T., Liu G., Zhao J., & Gao D. 2020. Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. *Applied Sciences* (Switzerland). Available at: <https://doi.org/10.3390/APP10186593>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).
- Garrido-Cárdenas, J. A., Cebrián-Carmona, J., González-Cerón, L., Manzano-Agugliaro, F., & Mesa-Valle, C. (2019). Analysis of global research on malaria and Plasmodium vivax. *International Journal of Environmental Research and Public Health*, 16(11). Available at: <https://doi.org/10.3390/ijerph16111928>
- Global Health, D. of P.D. and M. (2023) Symptoms of Malaria. Available at: [https://www.cdc.gov/malaria/about/symptoms\\_malaria.html](https://www.cdc.gov/malaria/about/symptoms_malaria.html).
- Islam, M. R., Jeba, T. N., Zulfiker, M. S., Rahman, M. S., & Rahman, M. (2023). Analysing the ML and DL-based Models for Predicting Malarial Fever Prior to Clinical Trial. 7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings, 1040–1045. Available at: <https://doi.org/10.1109/ICOEI56765.2023.10126059>
- Irwan (2016) Epidemiologi Penyakit Menular, Pengaruh Kualitas Pelayanan... *Jurnal EMBA*.
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15, 531–538. Available at: <https://doi.org/10.1002/sam.12000>

- <https://doi.org/10.1002/sam.11583>
- Kazwaini, M., & Willa, R. W. (2014). Korelasi Kepadatan Anopheles spp. dengan Curah Hujan serta Status Vektor Malaria pada Berbagai Tipe Geografi di Kabupaten Sumba Timur, Provinsi Nusa Tenggara Timur. *Buletin Penelitian Kesehatan*, 43(2), 77-88.
- Kementerian Dalam Negeri. (2024). Peta sebaran data Kependudukan Indonesia: Kabupaten Batu Bara. Diakses pada 18 September 2024, dari <https://gis.dukcapil.kemendagri.go.id/peta/>
- Kesehatan, K. (2022). Wilayah-wilayah Endemis Malaria Tinggi di Indonesia. Available at: <https://p2pm.kemkes.go.id/publikasi/artikel/wilayah-wilayah-endemis-malaria-tinggi-di-indonesia>.
- Kusuma, U., & Widjianto, A. (2016). Deskripsi bionomik nyamuk Anopheles Sp di wilayah Kecamatan Parigi Kabupaten Pangandaran Provinsi Jawa Barat Tahun 2016. *Keslingmas*, 35, 278-396
- Lie K. (2016). Profil Kesehatan Kabupaten Jayapura Tahun 2016. Dinas Kesehatan Kabupaten Jayapura. Jayapura.
- Latief, M.A., Bustamam, A., & Siswantining, T. (2020) "Performance Evaluation XGBoost in Handling Missing Value on Classification of Hepatocellular Carcinoma Gene Expression Data," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), pp. 1–6, 2020. doi: 10.1109/ICICoS51170.2020.9299012.
- Mbunge, E., et al. (2022). Application of machine learning models to predict malaria using malaria cases and environmental risk factors. 2022 Conference on Information Communications Technology and Society, ICTAS 2022 - Proceedings. Available at: <https://doi.org/10.1109/ICTAS53252.2022.9744657>
- Methiyothin, T., Ahn, I. (2022). Forecasting Dengue Fever in France and Thailand using XGBoost. *Journal of Epidemiology and Global Health*, Volume X, Issue X, 677-680.
- Nazarkar, A., Kuchulakanti, H., Paidimarry, C. S., & Kulkarni, S. (2023). Impact of Various Data Splitting Ratios on the Performance of Machine Learning Models in the Classification of Lung Cancer (pp. 96–104). Available at: [https://doi.org/10.2991/978-94-6463-252-1\\_12](https://doi.org/10.2991/978-94-6463-252-1_12)
- Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using

- climate variability and machine learning. *Informatics in Medicine Unlocked*, 22. Available at: <https://doi.org/10.1016/j.imu.2020.100508>
- Prasetyo, B., Irwandi, H., Pusparini, N., Besar Meteorologi Klimatologi dan Geofisika Wil -Medan, B. I., Ngumban Surbakti No, J., Medan, S. I., Klimatologi Deli Serdang, S., Meteorologi Klimatologi dan Geofisika, B., Meteorologi Raya No, J., & Medan, S. (2018). KARAKTERISTIK CURAH HUJAN BERDASARKAN RAGAM TOPOGRAFI DI SUMATERA UTARA Variable Topography-Based Rainfall Characteristic in North Sumatera. In *Jurnal Sains & Teknologi Modifikasi Cuaca* (Vol. 19, Issue 1).
- Rajab, S., Nakatumba-Nabende, J., & Marvin, G. (2023). Interpretable Machine Learning Models for Predicting Malaria. 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2023. Available at: <https://doi.org/10.1109/ICSTSN57873.2023.10151538>
- Rokom. (2022). Kejar Target Bebas Malaria 2030, Kemenkes Tetapkan 5 Regional Target Eliminasi. Sehat Negeriku. Available at: <https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220422/1439692/kejar-target-bebas-malaria-2030-kemenkes-tetapkan-5-regional-target-eliminasi/>
- Secada Purba, E. (2022) ‘Implementation of Generative Adversarial Networks for Creating Digital Artwork in the Form of Abstract Images’, *Jurnal Teknik Informatika (JUTIF)*, 3(3), pp. 707–715.
- Sandy, S., & Wike, I. (2019). Pengaruh iklim terhadap Annual Parasite Incidence malaria di Kabupaten Jayapura tahun 2011 – 2018. *Journal of Health Epidemiology and Communicable Diseases*, 5(1), 9–15. Available at: <https://doi.org/10.22435/jhecds.v5i1.1031>
- Sianturi, Aprilda Ariana. (2023). Letak Geografis, Batas Wilayah, serta Iklim di Sumatera Utara. *detikSumut*. Diakses pada 29 April 2024. Available at: <https://www.detik.com/sumut/berita/d-7076810/letak-geografis-batas-wilayah-serta-iklim-di-sumatera-utara>
- Suwito, Hadi UK, Sigit SH, Supratman S. (2010) . Hubungan Iklim , Kepadatan Nyamuk Anopheles dan Kejadian Penyakit Malaria. *Entomologi Indonesia*.7(1):42-53.
- Widi, S. (2022) Kasus Malaria Indonesia Melonjak 36,29% pada 2022. Available at: <https://dataindonesia.id/kesehatan/detail/kasus-malaria-indonesia-melonjak->

3629-pada-2022.

World Health Organization. (2023). Malaria. Diakses pada 8 April 2024.

<https://www.who.int/news-room/fact-sheets/detail/malaria>

WHO. (2023). World malaria report 2023, 1–356.

<https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023>

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series Generative Adversarial Networks.

Zou, M., Jiang, W. G., Qin, Q. H., Liu, Y. C., & Li, M. L. (2022). Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting. Materials, 15(15).  
<https://doi.org/10.3390/ma15155298>

