ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*

SKRIPSI

DISUSUN OLEH:

AYU SEKAR SARI
2009010106

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2024

ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Sistem Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

AYU SEKAR SARI

NPM. 2009010106

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2024

LEMBAR PENGESAHAN

Judul Skripsi	:	ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*
Nama Mahasiswa	:	AYU SEKAR SARI
NPM	:	2009010106
Program Studi	:	SISTEM INFORMASI

Menyetujui

Komisi Pembimbing

(Halim Maulana,ST., M.Kom)

NIDN. 0121119102

Ketua Program Studi	**Dekan**
(Martiano S.Pd, S.Kom., M.Kom)	**(Dr. Al-Khowarizmi, S.Kom., M.Kom.)**
NIDN. 0128029302	**NIDN. 0127099201**

PERNYATAAN ORISINALITAS

ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya

Medan, 25 Juni 2024

Yang Membuat Pernyataan

Ayu Sekar Sari

Npm.2009010106

PERNYATAAN PERSETUJUAN PUBLIKASI

KARYA ILMIAH UNTUK KEPENTINGAN

AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama : Ayu Sekar Sari
NPM : 2009010106
Program Studi : Sistem Informasi
Karya Ilmiah : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 25 Juni 2024

Yang Membuat Pernyatan

Ayu Sekar Sari
Npm.2009010106

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap : Ayu Sekar Sari

Tempat dan Tanggal Lahir : Medan 26 Maret 2003

Alamat Rumah : Medan Marelan

Telepon/Faks/HP : 081375409365

E-mail : Sekarsariayu641@gmail.com

Instansi Tempat Kerja : -

Alamat Kantor : -

DATA PENDIDIKAN

SD	:	SDN	TAMAT	:	2014
SMP	:	SMPN	TAMAT	:	2017
SMA	:	SMAN	TAMAT	:	2020

KATA PENGANTAR

Puji syukur atas kehadiran Allah SWT, berkat limpahan rahmat, hidayah dan karunianya, penulis bisa menyelesaikan skripsi dengan judul" **ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST***". Skripsi ini adalah salah satu dari beberapa persyaratan untuk menyelesaikan pendidikan dan memperoleh gelar sarjana pada program studi S1 Sistem Informasi di Universitas Muhammadiyah Sumatera Utara.

Penyusunan skripsi ini tidak lepas dari bimbingan, bantuan, arahan dan dukungan dari berbagai pihak terkait. Oleh karena itu pada kesempatan ini penulis menyampaikan terimakasih kepada:

1. Bapak Prof. Dr. Agussani, M.AP., Rektor Universitas Muhammadiyah Sumatera Utara (UMSU)
2. Bapak Dr. Al-khowarizmi, S.Kom.M.Kom selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi
3. Bapak Martiano, Martiano, S.Pd., S.Kom., M.Kom selaku Ketua Program Studi Sistem Informasi
4. Ibu Yoshida Sary, S.E., S.Kom., M.Kom selaku Sekretaris Program Studi Sistem Informasi
5. Bapak Halim Maulana,ST., S.Kom., M.Kom Selaku dosen pembimbing
6. Superhero dan panutanku,ayahanda Muhammad Aminin terima kasih telah berjuang untuk kehidupan penulis, terima kasih telah percaya atas semua keputusan yang telah penulis ambil untuk melanjutkan mimpinya, serta

cinta, doa dan motivasi yang selalu membuat penulis percaya bahwa penulis mampu menyelesaikan skripsi ini hingga akhir.

7. Pintu Surgaku,Ibunda Henny Sari Dewi yang tidak henti – hentinya memberikan kasih sayang dengan penuh cinta dan selalu memberikan dukungan,motivasi sert do'a yang dipanjatkan selama ini sehingga penulis mampu menyelesaikan studinya sampai sarjana
8. Kepada penyemangatku, adik tercintaku muhammad rahmad setiawan yang selama ini memberikan dukungan dan semangat kepada penulis, sehingga bisa menyelesaikan proposal penelitian sampai pada tahap penyusunan skripsi ini telah selesai.
9. Kepada Seha sebagai partner spesial saya, terimakasih telah menjadi sosok pendamping yang setia dalam segala hal, yang sudah meluangkan waktunya,menemani dan mendukung bahkan menghibur dalam kesedihan.tak hentinya memberikan semangat untuk terus maju tanpa kenal kata menyerah dalam meraih apa yang sudah menjadi impian saya.

ABSTRAK

ANALISIS PREDIKSI PENYEBARAN PENYAKIT DEMAM BERDARAH DENGAN PENDEKATAN *ENSEMBLE LEARNING* DENGAN *XGBOOST* DAN *RANDOM FOREST*

Demam berdarah *Dengue* (DBD) adalah penyakit menular yang signifikan di negara tropis, dengan dampak kesehatan masyarakat yang besar. Penelitian ini bertujuan mengembangkan model prediktif untuk memperkirakan jumlah kasus DBD di dua kota, *San Juan* dan *Iquitos*, menggunakan algoritma *Random Forest* dan *XGBoost*. Dataset yang digunakan adalah DengAI: Predicting Disease Spread, yang mencakup berbagai fitur lingkungan dan cuaca seperti suhu, curah hujan, kelembaban, dan *indeks vegetasi*, serta jumlah kasus DBD yang dilaporkan. Proses penelitian dimulai dengan pra-pemrosesan data untuk memastikan kualitas dan kesesuaian data. Setelah itu, model prediktif dibangun menggunakan *Random Forest* dan *XGBoost*. Evaluasi kinerja model dilakukan menggunakan *Mean Absolute Error* (MAE). Hasil penelitian menunjukkan bahwa model *XGBoost* memiliki kinerja yang lebih baik dalam memprediksi jumlah kasus DBD dibandingkan model *Random Forest*, dengan MAE yang lebih rendah untuk kedua kota. Model prediktif yang dihasilkan dapat membantu otoritas kesehatan dalam perencanaan dan pelaksanaan tindakan pencegahan yang lebih efektif. Penelitian ini menegaskan potensi penggunaan teknik pembelajaran mesin dalam epidemiologi penyakit menular dan memberikan wawasan penting tentang faktor-faktor lingkungan yang mempengaruhi penyebaran DBD.

Kata kunci: Demam Berdarah *Dengue*, Prediksi Kasus, *Random Forest*, *XGBoost*, Pembelajaran Mesin, Epidemiologi**.**

ABSTRACT

ANALYSIS OF DENGUE FEVER SPREAD PREDICTION USING ENSEMBLE LEARNING APPROACHES WITH XGBOOST AND RANDOM FOREST

Dengue fever is a significant infectious disease in tropical countries, with considerable public health impact. This study aims to develop a predictive model to estimate the number of Dengue cases in two cities, San Juan and Iquitos, using Random Forest and XGBoost algorithms. The dataset used is DengAI: Predicting Disease Spread, which includes various environmental and weather features such as temperature, rainfall, humidity, and vegetation indices, as well as reported Dengue case numbers.The research process began with data pre-Processing to ensure data quality and suitability. Subsequently, predictive models were built using Random Forest and XGBoost. Model performance was evaluated using Mean Absolute Error (MAE). The results indicated that the XGBoost model outperformed the Random Forest model in predicting the number of Dengue cases, with a lower MAE for both cities. The predictive models developed can assist health authorities in planning and implementing more effective preventive measures. This study underscores the potential of Machine Learning techniques in the epidemiology of infectious diseases and provides valuable insights into the environmental factors influencing the spread of Dengue.

Keywords: *Dengue Fever, Case Prediction, Random Forest, XGBoost, Machine Learning, Epidemiology***.**

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Demam berdarah *Dengue* (DBD) merupakan gejala virus yang disebarkan pada individu proses gigitan nyamuk spesies *Aedes* yang terinfeksi termasuk *Aedes aegyptidan Aedes albopictus* (Sabir dkk, 2021). Lebih dari 100 juta orang terkena demam berdarah setiap tahunnya, virus *Dengue* merupakan virus yang penyebaran dan penularannya paling cepat secara global dan ditularkan melalui gigitan nyamuk (Schaefer, 2023). Prevalensi demam berdarah meningkat pesat dari waktu ke waktu, mengakibatkan antara 100 dan 400 juta kasus setiap tahunnya (Sabir dkk, 2021). Spesies *Aedes* semampai didapatkan pada wilayah tropis dan subtropis sehingga sekitar 4 miliar orang tinggal di daerah yang berisiko terkena demam berdarah.

Penyebaran penyakit demam berdarah dipengaruhi oleh sejumlah faktor, termasuk kondisi lingkungan, iklim, kepadatan populasi, sanitasi, dan upaya pencegahan dan pengendalian yang dilakukan. Oleh karena itu, pemahaman yang mendalam tentang pola dan yang berdampak pada perluasan demam berdarah pada Kota Medan menjadi krusial dalam upaya pencegahan dan pengendaliannya. Pada tahun 205, Kamran Shaukat dkk. Melakukan studi tentang Prediksi DBD di Pakistan menggunakan lima algoritma untuk mencari akurasi terbaik dari dua algoritma, Naive Bayes dan J48, memiliki akurasi berurut 92% dan 88%, masing-masing. Dalam penelitian Jackins et al., metode Random Forest digunakan untuk memprediksi penyakit gula, penyakit jantung koroner, dan kanker payudara.

Semua dataset yang digunakan adalah NIDDK untuk diabetes, Studi Jantung Framingham untuk penyakit jantung koroner, dan Wisconsin Breast Cancer untuk kanker payudara. Peneliti memanfaatkan alat Anaconda. Hasil penelitian menunjukkan bahwa metode Random Forest adalah algoritma klasifikasi yang lebih baik daripada metode Bayesian. Dengan nilai akurasi sebesar 74.3% untuk diabetes, 83.85% untuk penyakit jantung koroner, dan 92.40% untuk kanker payudara, metode Random Forest mengalahkan metode Bayesian (Jackins et al., 2021). Percobaan ini bertujuan untuk mengembangkan model prediktif menggunakan algoritma *Machine Learning*, yaitu *Random Forest* dan *XGBoost*, untuk memprediksi jumlah kasus demam berdarah di masa mendatang. Dengan menggunakan data historis yang mencakup berbagai fitur lingkungan dan meteorologi, diharapkan model ini dapat memberikan prediksi yang akurat sehingga otoritas kesehatan dapat mengambil tindakan preventif yang lebih tepat waktu. Penelitian ini menggunakan data historis dari dua kota, *San Juan* dan *Iquitos*, yang mencakup berbagai fitur seperti temperatur udara, kelembaban relatif, curah hujan, dan Normalized Difference Vegetation Index (NDVI). Data tersebut diproses dan dibagi menjadi set pelatihan, validasi, dan pengujian untuk memastikan keakuratan model. Algoritma *Random Forest* dan *XGBoost* spesifik karena daya mereka analitis menangani data pada kompleks dengan beragam juga memberikan hasil yang dapat diinterpretasikan dengan baik.

1.2 Ringkasan Permasalahan

1. Bagaimana cara menggunakan data historis lingkungan dengan meteorologi untuk memprediksi jumlah kasus demam berdarah?

2. Algoritma *Machine Learning* apa yang terbaik dalam memprediksi jumlah kasus demam berdarah?
3. Bagaimana performa model *Ensemble Learning XGBoost* dan *Random Forest* dalam mengklasifikasikan daerah dengan risiko penyebaran penyakit demam ?

1.3 Batasan Kajian

Makna penulisan proposal ini adalah batasan masalah yang ada pada rumusan masalah yang disebutkan sebelumnya, yaitu:

1) Penelitian ini tidak akan membahas aspek biologi dan sosial dari demam berdarah, seperti perilaku nyamuk dan intervensi kesehatan masyarakat.
2) Menentukan algoritma Machine Learning ini semuanya efisien ketika memprediksi jumlah kasus demam berdarah di wilayah tertentu
3) Penelitian ini berfokus pada wilayah yang spesifik. Hasilnya mungkin tidak dapat digeneralisasi ke wilayah lain dengan karakteristik yang berbeda.

1.4 Visi Kajian

Observasi ini bermaksud mengembangkan model prediktif akan memperkirakan jumlah kasus demam berdarah di *San Juan* dan *Iquitos* menggunakan algoritma *XGBoost* dan *Random Forest*. Dengan memanfaatkan dataset DengAI, penelitian ini berusaha mengidentifikasi faktor lingkungan yang mempengaruhi penyebaran penyakit. Tujuan utama adalah meningkatkan akurasi prediksi untuk membantu otoritas kesehatan dalam merencanakan tindakan pencegahan yang lebih efektif.

1.5 Keunggulan Kajian

Salah satu keuntungan dari penelitian adalah sebagai berikut:

1.5.1 Keunggulan Akademis

1) Diharapkan penelitian ini akan memberikan perkembangan ilmu komputer, khususnya sistem informasi di Program studi Fakultas Ilmu Komputer Dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
2) Sebagai salah satu syarat untuk memperoleh gelar Sarjana (S1), penelitian ini dapat meningkatkan pengetahuan dan pengalaman dalam penerapan disiplin ilmu yang telah diterima selama perkuliahan.di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
3) Penelitian ini bisa menambah pengetahuan terutama tentang memprediksi jumlah kasus DBD serta bisa memahami sumber literatur tambahan pada pengkaji tersebut.

1.5.2 Keunggulan Praktis

1) Model prediktif yang dikembangkan dapat membantu otoritas kesehatan meningkatkan akurasi prediksi jumlah kasus demam berdarah, sehingga memungkinkan perencanaan yang lebih tepat waktu dan efektif dalam mengatasi wabah.
2) Dengan prediksi yang lebih akurat, otoritas kesehatan dapat merencanakan dan mengimplementasikan tindakan pencegahan yang lebih efektif, seperti kampanye pemberantasan nyamuk dan edukasi masyarakat.

3) Anggapan penelitian ini dapat membantu dalam pengambilan keputusan bagi pembuat kebijakan kesehatan masyarakat untuk mengalokasikan sumber daya dan merancang strategi intervensi terbaik.

BAB II

LANDASAN TEORI

2.1 Perkiraan

Perkiraan atau peramalan penjualan, juga dikenal sebagai prediksi, adalah proses mengevaluasi kondisi masa lalu untuk memprediksi kondisi masa depan. (Alfani W.P.R. et al., 2021). Prediksi menunjukkan apa yang akan terjadi pada situasi tertentu dan merupakan komponen dalam proses perencanaan dan pengambilan keputusan; namun, mereka tidak harus memberikan jawaban pasti untuk kasus yang hendak ada, kecuali bergerak selama menemukan jawaban sedekat mungkin.

Prediksi bisa ilmiah atau subjektif. Ambil contoh, prediksi cuaca selalu didasarkan pada data dan informasi terbaru, termasuk pengamatan satelit. Pada awalnya, walaupun pengkajian mendalam tentang alternatif masa depan adalah suatu disiplin baru, orang mungkin sangat tertarik pada apa yang akan terjadi kemudian karena individu tiba mengetahui sesuatu perkiraan mampu berdasarkan metode ilmiah ataupun subjektif belaka. Ambil model, perkiraan cuaca selalu didasarkan pada data dan informasi terbaru, termasuk pengamatan satelit. Pada awalnya, walaupun pengkajian mendalam tentang alternatif masa depan adalah disiplin baru, orang mungkin sangat memperhatikan apa yang akan terjadi kemudian karena manusia mulai mengetahui sesuatu. (Kafil, 2019).

2.2 Demam Berdarah

Demam Berdarah *Dengue* (DBD) adalah salah satu penyakit melandai yang ditimbulkan maka dari itu virus Dengue ditularkan lewat gigitan nyamuk Aedes aegypti atau Aedes albopictus. Nyamuk Aedes aegypti biasanya berkembang biak di bak mandi, ember, ban bekas, dan tempat minum burung, sementara Aedes albopictus lebih banyak ditemukan di tempat penampungan air alami di luar rumah, seperti lubang pohon dan potongan bambu, terutama di daerah pinggi. (Sukendra, D., M., 2021).

Penyakit dengue fever (DBD) juga disebabkan oleh kondisi kesehatan lingkungan yang buruk, seperti perilaku masyarakat yang buruk atau sering membuang sampah di tempat yang tidak perlu, di mana sampah menjadi tempat nyamuk berkembang biak, dan tiap-tiap nyamuk membawa virus Dengue fever ke tempat yang lebih tinggi. Tingkat kelembaban udara juga memengaruhi kejadian demam berdarah, seperti yang ditunjukkan oleh penelitian Agus tentang faktor-faktor yang memengaruhi kejadian demam berdarah, yang menyebabkan populasi nyamuk perantara virus Dengue meningkat.(Timah, 2021)

2.3 Data *Science*

Data Science adalah bidang ilmu yang khusus mempelajari data, terutama data kuantitatif (angka), baik yang terstruktur maupun tidak terstruktur. Bidang ilmu ini mencakup semua proses yang berkaitan dengan data, seperti pengumpulan, analisis, pengolahan, manajemen, kearsipan, pengelompokan, penyajian, distribusi, dan cara mengubah data menjadi kumpulan informasi yang mudah dipahami. Data *scientist* biasanya disebut statistikawan. Oleh karena itu, tidak mengherankan bahwa data *scientist* lebih sering membuat algoritma program komputer sehingga

komputer dapat langsung mengolah data yang masuk. Big data, machine learning, dan internet of things adalah beberapa bagian teknologi yang sangat membantu ilmu data.(Aditya et al., 2020).

2.4 *Machine Learning*

Machine learning adalah cabang kecerdasan buatan yang membantu sistem mengadaptasi kemampuan manusia untuk belajar. pada aplikasi machine learning, algoritma atau urutan proses statistik untuk membuat prediksi dalam pengembangan data melalui penggunaan statistik. Data digunakan untuk membuat keputusan dan prediksi. Semakin baik algoritma, semakin akurat prediksi dan keputusan sistem (Mahendra et al., 2022). Ada beberapa manfaat pengajaran mesin, antara lain:

• *Classification (Klasifikasi)* untuk memprediksi nilai atau kelas seseorang dalam sebuah populasi, klasifikasi adalah teknik pembelajaran mesin.

•Pencocokan kemiripan, juga dikenal sebagai teknik pembelajaran mesin yang digunakan untuk menemukan kemiripan antara orang-orang berdasarkan data yang ada.

•*Clustering* (Pengklasteran) juga dikenal sebagai metode pengajaran mesin yang digunakan untuk mengelompokkan orang dalam grup yang sama bersumber kesamaan yang mereka miliki.

Dalam penelitian ini, kegunaan metode pengajaran mesin adalah pada poin pertama, yaitu untuk memprediksi nilai atau kelas seseorang.

Machine learning adalah bidang ilmu komputer lain yang mencakup pengembangan algoritma yang memungkinkan komputer untuk belajar melalui data, yang sering disebut sebagai belajar dari data. Dengan kata lain, machine

learning adalah pemrograman komputer yang menggunakan data masa lampau untuk pembelajaran model, yang memungkinkan komputer untuk mendapatkan hasil yang optimal dari kumpulan data yang digali. Inti pembelajaran mesin merupakan model yang menggambarkan pola data. Namun, menurut Tom M. Mitchell, machine learning merupakan semacam program komputer sehingga menggunakan kinerja yang terukur untuk belajar dari pengalaman dari tugas yang dibebankan. Secara umum, pembelajaran mesin dibagi menjadi tiga kategori: pembelajaran yang diawasi, pembelajaran yang tidak diawasi, dan pembelajaran pendukung. (Sidik & Ansawarman, 2022).

a) Algoritma pembelajaran mesin yang diawasi memiliki proses pembelajaran yang diawasi. Klasifikasi dan regression adalah contoh pembelajaran diawasi. (Dalam penyelidikan ini, poin pertama digunakan)
b) Pembelajaran tanpa pengawasan adalah algoritma pembelajaran mesin yang melakukan pembelajaran tanpa pengawasan. Yang termasuk dalam pembelajaran tanpa pengawasan adalah pengurangan dimensi dan kelompokan.
c) Pembelajaran meningkatkan adalah algoritma pembelajaran mesin yang memiliki kemampuan untuk membuat agent software mesin bekerja secara otomatis untuk menentukan perilaku yang ideal untuk mengembangkan kinerja algoritma. Yang termasuk dalam pembelajaran meningkatkan

2.5 *Ensemble Learning*

Menurut (Nguyen et al., 2021) Ensemble Machine Learning didefinisikan sebagai metode yang menggabungkan berbagai model dasar pembelajaran mesin, baik heterogen maupun homogen, untuk meningkatkan prediksi serta memudahkan noise atau kelalaian pusat data yang diamati dan diprediksi. Sebagian besar orang membagi metode ensemble ke dalam tiga kategori: bootstrap aggregating (bagging), boosting, dan squared. Ketiga kategori ini melakukan penyesuaian prediksi bersumber perhitungan segala model, bias, maupun keduanya dengan bersamaan. Salah satu selisih utamanya merupakan bahwa bagging dan boosting umumnya berfungsi dengan model yang homogen, sedangkan squared lebih baik dalam menggabungkan model yang heterogen.

2.6 *Extreme Gradient Boosting* (*XGBoost*)

Extreme Gradient boosting, merupakan implementasi dari kerangka *gradient boosting* yang efisien dan terukur (Chen & He, 2021). *XGBoost* adalah salah satu implementasi paling populer dan efisien dari algoritma *Gradient Boosted Trees*, sebuah metode *supervised learning* yang didasarkan pada perkiraan fungsi dengan mengoptimalkan fungsi kerugian spesifik dan menerapkan beberapa teknik regularisasi.

XGBoost adalah metode pembelajaran *ensemble*. Metode pembelajaran *ensemble* merupakan cara pembelajaran mesin yang menghasilkan model prediksi yang optimal dengan mempersatukan selama model dasar. Pembelajaran *ensemble* menawarkan solusi sistematis untuk menggabungkan kekuatan prediktif banyak *learner*. Hasilnya adalah model tunggal yang memberikan *output* agregat dari beberapa model. *Boosting* mengacu pada algoritma yang mampu mengubah *weak*

learner menjadi *strong learner* dengan prinsip utama menyesuaikan urutan model *weak learner*, yang hanya sedikit lebih baik daripada menebak secara acak seperti pohon keputusan sederhana, untuk versi data yang berbobot. Lebih banyak bobot diberikan pada contoh yang salah diklasifikasi oleh putaran sebelumnya. Prediksinya kemudian digabungkan melalui suara terbobot mayoritas atau *weighted majority vote* (klasifikasi) ataupun jumlah tertimbang atau *weighted sum* (regresi) untuk menghasilkan prediksi akhir.

Adapun langkah-langkah dalam menerapkan algoritma *XGBoost* ini di antaranya sebagai berikut:

a) Langkah pertama yaitu menentukan nilai *initial prediction*. Nilai *initial prediction* memiliki nilai *default* 0,5 baik untuk regresi atau klasifikasi. Nilai ini bisa mencakup hal-hal seperti probabilitas dari pengamatan pada data latih yang diobservasi. Angka tersebut menandakan bahwa ada 50% kemungkinan yang sedang diobservasi bernilai positif (*true*).

b) Selanjutnya setelah nilai *initial prediction* ditentukan, akan diperoleh nilai *residual*. Nilai *residual* merupakan perbedaan yang ada atau *gap* antara sampel yang sedang diobservasi dengan yang diprediksi yang dapat dilihat pada Persamaan (1).

$$Residual = Observe\ Predicted \quad \text{.....(1)}$$

c) Sama halnya dengan *XGBoost* regresi, untuk menyesuaikan bentuk *XGBoost tree* ke nilai-nilai *residual* digunakan rumus *Similarity Score* yang sedikit berbeda untuk regresi. Rumus *Similarity Score* untuk klasifikasi tertera seperti dalam Persamaan (2).

$$\text{Similarity score} = \frac{(\sum Residual_i)^2}{\sum[Previsiousprobability_i x(1-PreviousProbabilit y_i)]+\lambda} \quad \ldots\ldots(2)$$

Keseluruhan *residual* dimasukkan ke dalam satu *leaf* yang sama dan dihitung nilai *Similarity Score* dari *leaf* tersebut.

d) Untuk mengetahui apakah kinerja mengelompokkan *residual* yang serupa akan lebih baik apabila dilakukan pemisahan menjadi dua grup terpisah digunakan perbandingan nilai *Gain* dari tiap kemungkinan pemisahan yang ada. Nilai *Gain* menentukan seperti apa *branch* dari tree yang sedang dikerjakan. *Gain* dirumuskan seperti dalam Persamaan (3).

$$Gain = Leftsimilarity + Rightsimilarity - Rootsimilarity \ldots\ldots..(3)$$

Dari semua kemungkinan nilai *Gain* yang didapatkan setelah pemisahan tiap sampel yang diobservasi, nilai *Gain* tertinggi dipilih menjadi *branch* yang memisahkan *residual*.

e) Selanjutnya dilakukan pengecekan apakah masih ada *residual-residual* di dalam *leaf* yang masih dapat dipisahkan dan dibentuk menjadi *branch*. Proses *splitting* atau pemisahan dilakukan sesuai dengan batas kedalaman tree yang ditentukan. Kedalaman dari *Tree XGBoost* standarnya memungkinkan hingga 6 level kedalaman dan kedalaman levelnya bebas ditentukan.

f) Jumlah minimum dari *residual* dalam tiap-tiap *leaf* dan ditentukan dengan menghitung nilai *Cover*. Secara *default*, nilai minimum dari *Cover* adalah 1 dan apabila nilai *Cover* suatu *leaf* kurang dari 1 maka *XGBoost* tidak memperbolehkan *leaf* tersebut. Cover dirumuskan seperti dalam Persamaan (4)

$$Cover = \sum[\ Previous\ Probability_i \times (1 - Previous\ Probability_i\ )] \quad (4)$$

g) Untuk melakukan pemangkasan atau *pruning* dari *tree*, yang perlu dilakukan adalah menghitung selisih antara *Gain* dari cabang atau *branch* paling bawah dari *tree* dengan nilai γ (gamma) yang sudah ditetapkan. Apabila selisih yang didapatkan bernilai positif maka tidak akan dilakukan pemangkasan pada *branch* tersebut. Namun, jika bernilai negatif maka pemangkasan dilakukan dan berhenti hingga mencapai *branch* lain yang bernilai positif.

h) Dalam melakukan pemangkasan atau *pruning* ini dapat terjadi pemangkasan ekstrim yaitu kondisi dimana *tree* habis terpangkas dan menyisakan nilai *initial prediction*. Untuk mencegah terjadinya *extreme pruning*, *Regularization Parameter*, atau λ (lambda) pada mengurangi nilai *Similarity Scores* yang berdampak ke menurunnya nilai *Gain*. Nilai λ (lambda) lebih dari 0 mengurangi sensitivitas dari *tree* ke *individual observation* dengan *pruning* dan menggabungkan mereka dengan observasi lain.

i) Jika *tree* sudah terbentuk, langkah selanjutnya adalah menghitung nilai *Output*

Value dari tiap-tiap node pada *tree* yang dirumuskan seperti Persamaan 5

$$Output\ Value = \frac{\Sigma Residual_i)}{\Sigma(PrevisioudsProbability_i x(1 - PrevisiousProbability_i)|+\lambda} \quad \ldots\ldots(5)$$

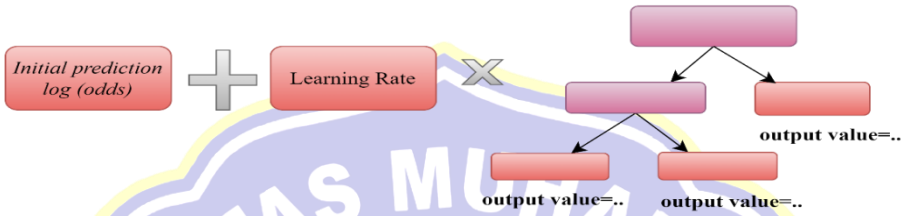
j) Selanjutnya prediksi baru dapat dilakukan. Seperti metode *boosting* lainnya, *XGBoost* untuk klasifikasi membuat prediksi baru dengan mengawalinya dengan *initial prediction*. Seperti pada *Gradient Boost Classification*, perlu untuk mengubah nilai probabilitas pada *initial prediction* menjadi nilai *log(odds). Odds* atau peluang (secara teknis peluang keberhasilan) didefinisikan sebagai probabilitas keberhasilan dibagi dengan probabilitas kegagalan dan dirumuskan dengan persamaan (6)

$$odds = \frac{Probability}{1 - probability} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Sedangkan *log(odds*) adalah logaritma peluang yang dirumuskan seperti dalam perumusan (7)

$$\log(odds) = \log\left(\frac{Probability}{1 - probability}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Kemudian mulai membuat prediksi baru atau *new prediction* dengan menggabungkan nilai-nilai seperti yang ditunjukkan pada gambar berikut.

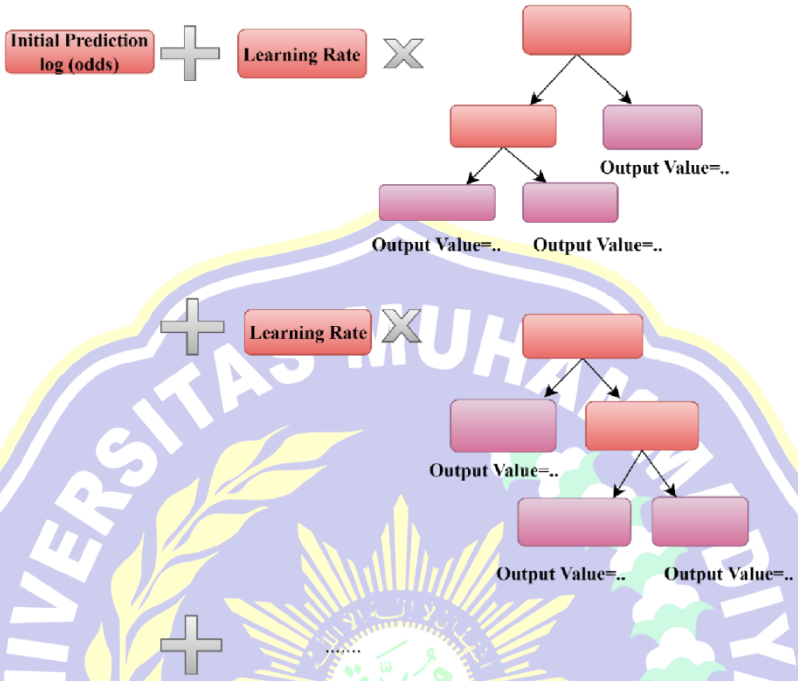


Gambar 2.1 New predi ction dalam XGBoost

Learning rate atau eta dalam *XGBoost* secara *default* bernilai 0,3 dan nilainya bebas ditentukan. Jika nilai *new prediction* sudah didapat, selanjutnya adalah mengubah nilai yang diperoleh ke dalam probabilitas dengan *Logistic Function* seperti yang dirumuskan dalam Persamaan.

Apabila nilai *new prediction* menghasilkan nilai probabilitas yang membuat nilai *residual* makin kecil dari sebelumnya maka prediksinya mengarah ke arah yang benar.

k) Untuk membuat tree selanjutnya, acuan berubah pada nilai *residual* baru atau dengan kata lain *tree* selanjutnya menyesuaikan nilai *residual* yang baru dihasilkan. Alur algoritma ditunjukkan pada Gambar 6. Proses diulang seperti di awal dan *tree* terus menerus dibuat sampai nilai-nilai *residual* menjadi semakin sangat kecil atau pembuatan *tree* sudah mencapai maksimum.



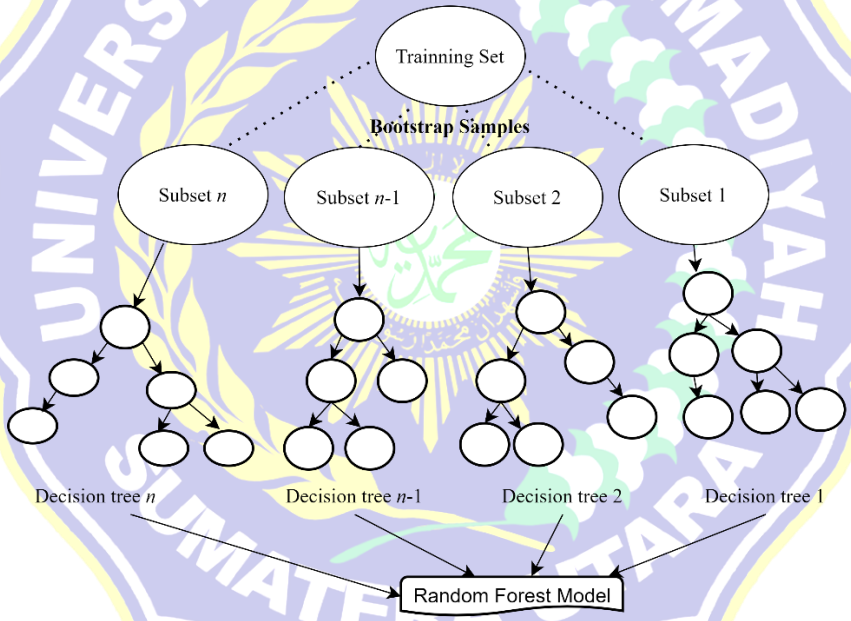
Gambar 2.2 Algoritma XGBoost

Algoritma pembelajaran kelompok Extreme Gradient Boosting, juga dikenal sebagai XGBoost, dikembangkan pada tahun 2014 oleh Tianqi Chen. Algoritma ini didasarkan pada prinsip gradient boosting, yaitu bertambah mengarahkan pada contoh yang salah dipendataan oleh pendataan sebelumnya (Dangeti, 2017). Pohon XGBoost yang ditingkatkan dibagi menjadi dua: pohon regresi dan klasifikasi, menurut artikel milik (Zheng et al., 2017). Selain itu, XGBoost memiliki keunggulan dalam fleksibilitas dan efisiensi yang luar biasa dalam kompetisi pembelajaran mesin Kaggle. Pada kompetisi tahun 2015 silam, XGBoost dinobatkan sebagai metode yang paling populer melalui 17 penyelesaian dari 29 penyelesaian pemenang (W. Zhang et al., 2021).

2.7 *Random Forest*

Menurut (Azis, Tangguh Admojo dan Susanti, 2020), Random Forest terdiri dari tiga pohon keputusan, dengan setiap pohon bertanggung jawab atas setipa nilai dari vector random yang diberikan, dan persebarannya akan diberikan kepada

pohon keputusan lainnya. Random Forest adalah salah satu model klasifikasi yang mempraktikan pohon keputusan tanpa memotong pohon untuk meningkatkan algoritma dan akurasi yang didapat untuk menghindari overfitting. Random Forest dilakukan melalui penggunaan Replace dan rancangan sampling bootstrap (Wu et al., 2017).



Gambar 2.3 *Random Forest*

Metode Random Forest pseudocode (Jackins et al., 2021):

a) Pilih fitur “n” secara acak dari total fitur “k”, di mana n adalah k
b) Hitung node “n” di antara fitur “n” menggunakan titik pisah terbaik.
c) Menggunakan pemisahan terbaik, membagi node menjadi node anak.
d) Ulangi langkah 1 hingga 3 hingga jumlah node "l" tercapai
e) Bangun Hutan Random dengan memecahkan jalan 1 hingga 4 sebanyak "n" kali untuk membuat jumlah pohon "n".

Pembentukan Random Forest dimulai dengan mengambil dataset acak sejumlah n (jumlah total data) mulai dataset latih dengan perubahan, sehingga masing-masing dataset memiliki data yang sama. Selanjutnya, menggunakan algoritma CART (Classification and Regression Trees) dan aturan splitting Gini, buat pohon keputusan dari dataset acak tersebut.

Index atau Entropy. Pohon keputusan ini akan menjadi sebuah decision tree.

$$\text{Gini}\ (A) = 1 - \sum_{i=1}^{n} P_i^2$$

$$\text{Entropy}\ (A) = -\sum_{i=1}^{n} P_i \log_2 (Pi)$$

Dimana :

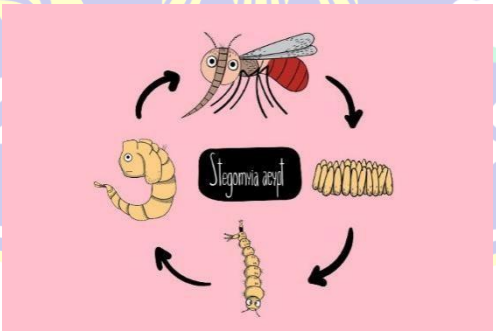
n : Jumlah kelas target

Pi adalah persentase jumlah sampel kelas i dibandingkan dengan jumlah sampel total

1. Untuk membuat pohon keputusan M, ulangi langkah 1 dan 2 sebanyak M kali. Setiap kelompok keputusan dibuat dari kumpulan data acak yang berbeda.
2. Klasifikasikan setiap peristiwa dengan menggunakan pohon keputusan yang telah dibuat sebelumnya, yang menghasilkan M prediksi untuk setiap peristiwa.
3. Gabungkan hasil prediksi dari pohon keputusan M untuk membuat satu prediksi akhir. Prediksi rata-rata dari pohon keputusan M dapat diperoleh dengan regresi, sementara prediksi berat dapat diperoleh dengan klasifikasi.
4. Menilai akurasi model pada kumpulan data yang diuji.

2.8 Siklus Hidup Nyamuk

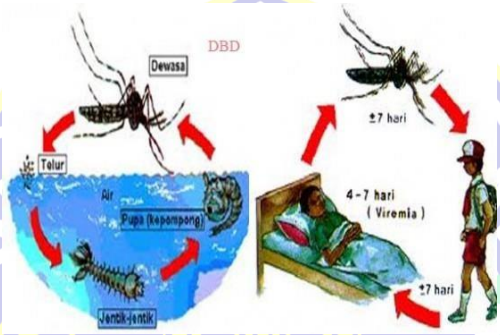
Siklus hidup sempurna nyamuk Aedes aegypti terdiri dari empat fase: telur, jentik, pupa, dan nyamuk dewasa. Telur berbentuk elips bercorak hitam tidak berdekatan. Dalam satu hingga dua hari, telur menjadi jentik. Jentik berkembang dalam empat tahapan, dikenal sebagai instar. di mana jentik masuk ke fase dorman Nyamuk dewasa keluar dari pupa setelah dua hari.Setelah mencapai instar keempat, larva berubah menjadi pupa, di mana jentik memasuki masa dorman. Setelah dua hari, nyamuk dewasa keluar dari pupa. Antara telur dan nyamuk dewasa membutuhkan waktu delapan hingga sepuluh hari, tetapi dapat berlangsung lebih lama jika kondisi lingkungan tidak baik.



Gambar 2.4 Siklus Nyamuk *Aedes aegypti*

Penyakit DBD dapat terjadi pada semua orang, tetapi lebih sering melantas anak balita hingga usia sekolah. Peluang untuk terinfeksi virus Dengue melalui gigitan nyamuk Aedes aegypti berbeda untuk masing-masing kelompok umur. Tidak ada penelitian yang menunjukkan bahwa laki-laki dan perempuan memiliki kerentanan yang berbeda terhadap penyakit ini. Tidak semua orang yang digigit nyamuk terinfeksi virus Dengue akan terserang dengue fever (DBD), tergantung pada kekebalan tubuh mereka. Orang dengan kekebalan tubuh yang baik terhadap

virus Dengue tidak akan terserang DBD meskipun virus itu ada dalam darahnya. Sebaliknya, orang dengan kekebalan tubuh yang lemah terhadap virus Dengue akan terserang DBD.



Gambar 2.5 Siklus Penyebaran DBD

Demam, nyeri perut, muntah, dan kelelahan adalah gejala demam berdarah. Perdarahan juga terjadi pada penderita demam berdarah, seperti pada hidung, gusi, atau di bawah kulit, yang membuatnya terlihat seperti memar. Selain itu, darah dapat ditemukan dalam urin, feses, atau muntah. Jika Anda mengalami sesak napas atau keringat dingin, segera cari pertolongan medis. Namun, seperti demam berdarah, demam Dengue dimulai dengan gejala demam, yang merupakan bentuk ringan dari infeksi virus Dengue. Gejalanya muncul dalam waktu empat hingga tujuh hari sejak gigitan nyamuk dan dapat bertahan selama sepuluh hari. Gejala demam dengue termasuk:

1) Suhu badan tinggi yang bisa mencapai 40 derajat *Celcius* atau lebih. Sakit kepala berat
2) Nyeri pada sendi, otot, dan tulang.
3) Hilang nafsu makan.
4) Nyeri pada bagian belakang mata.
5) Mual dan muntah.

6) Pembengkakan kelenjar getah bening.

7) Ruam kemerahan (muncul sekitar 2-5 hari setelah demam). Pada demam *Dengue*, biasanya penderita akan sembuh dalam 7 hari.

Gambar 2.6 Gejala DBD

2.9 Penelitiaan Terdahulu

Menurut penelitian yang dilakukan oleh Sari et al. (2022), informasi dan pengetahuan terhadap masyarakat diharapkan dapat mencegah terjadinya DBD, termasuk bahaya dan dampak DBD. Pengetahuan masyarakat yang benar tentang penyakit ini sangat penting untuk mengendalikan vektor DBD di rumah sendiri, sementara pengetahuan yang kurang akan menyebabkan peningkatan kasus DBD. Pengetahuan masyarakat sangat penting untuk upaya pencegahan DBD yang dilakukan oleh responden. Dengan peningkatan pengetahuan responden, pencegahan DBD akan lebih efektif, dan begitupun sebaliknya. Perilaku yang didasarkan pada pengetahuan dan kesadaran akan lebih lama bertahan daripada perilaku yang tidak.

\ Hasil penelitian ini sejalan dengan teori Green dalam Notoatmodjo (2010), yang menyatakan bahwa orang yang berpengetahuan tinggi lebih cenderung berperilaku baik dalam bidang kesehatan, termasuk mencegah DBD. Sebaliknya, hal itu juga berlaku untuk orang yang berpengetahuan rendah. Oleh karena itu, sikap

positif responden penelitian terhadap upaya pencegahan DBD mengarah pada tindakan pencegahan DBD yang efektif. Sebaliknya, sikap negatif responden, yang dapat disebabkan oleh kurangnya pengetahuan tentang bahaya DBD dan upaya pencegahannya, dapat mengakibatkan kesadaran dan tindakan pencegahan DBD yang rendah.

BAB III

METODOLOGI PENELITIAN

3.1 Tipe Kajian

Penelitian ini menggunakan jenis penelitian yang dijelaskan di atas. kuantitatif. Sifat datanya objektif dimana orang yang membaca data tersebut akan menginterpretasikan hasil yang sama. Orientasi hasil penelitian kuantitatif adalah hasil penelitian berdasarkan kesimpulan, generalisasi, prediksi. Dalam penelitian kuantitatif ini penelitian menguji teori yang sudah ada. Sebagai bagian dari proses pembelajaran, Lembab dapat memberikan contoh mata pelajaran yang tidak spesifik, tetapi juga dapat memberikan contoh atau memungkinkan generalisasi. (Kusumastuti et al., 2020).

Kajian kuantitatif lebih metodis, tertata, teratur, jelas dari awal hingga batas terjauh kajian dan tidak terpengaruh oleh kondisi terkini di lapangan. Tahapan dari awal hingga akhir penelitian dapat diantisipasi karena spesifikasi penelitian kuantitatif disusun secara konsisten dan tegas. Di sisi lain, disebutkan bahwa penelitian kuantitatif menuntut banyak tujuan angka, mulai dari pengumpulan data, pemahaman tentang data, dan tampilan hasil. Pemahaman bacaan akan meningkat dan penyampaian informasi akan dipermudah ketika hasil disajikan dalam bentuk gambar, tabel, grafik, atau tampilan representatif lainnya. (Siddik & Sunarsi, 2021)

3.2 Jangka lama Kajian

Waktu penulis dihabiskan untuk penelitian ini, yang dimulai pada Februari 2024 dan berakhir pada Mei 2024.

3.3 Tabel Waktu Penelitian

No.	Aktivitas Penelitian	Februari				Maret				April				Mei			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.	Pemberian Judul																
2.	Mengumpulkan Informasi																
3.	Perapian Proposal																
4.	Arahan Proposal																
5.	Seminar Proposal																
6.	Perapihan Skripsi																
7.	Arahan Skripsi																
8.	Sidang Meja Hijau																

Tabel 3.1 *Schedule* Penelitian

3.4 Alokasi Informasi Penelitian

Data sekunder adalah data yang digunakan dalam penelitian ini. Informasi Penting adalah informasi data yang diperoleh secara langsung yang dikumpulkan langsung dari sumbernya. Informasi penting ini adalah informasi yang paling unik dalam karakter dan tidak terpapar pada perlakuan terukur apa pun. Peneliti harus mengumpulkan data secara langsung melalui observasi, wawancara, dan diskusi terfokus untuk mendapatkan data primer, dan penyebaran kuesioner (Sari & Zefri, 2019).

Penelitian ini menggunakan dataset DengAI: *Predicting Disease Spread* yang disediakan oleh kompetisi *DrivenData*. Dataset ini terdiri dari data lingkungan dan

cuaca serta data jumlah kasus demam berdarah yang dilaporkan di dua kota, *San Juan* (Puerto Rico) dan *Iquitos* (Peru). Data lingkungan dan cuaca mencakup variabel-variabel seperti suhu harian (rata-rata, maksimum, dan minimum), curah hujan, kelembaban *relatif*, dan *Normalized Difference Vegetation Index* (NDVI) untuk empat wilayah (NE, NW, SE, SW) yang mencerminkan kondisi vegetasi dan tingkat kelembaban tanah. Selain itu, variabel lain seperti jumlah hari hujan dan kelembaban rata-rata juga disertakan. Data jumlah kasus demam berdarah mencakup laporan mingguan dari *San Juan* dan *Iquitos* yang dikumpulkan oleh instansi kesehatan setempat, mencakup periode waktu yang cukup panjang untuk memungkinkan analisis tren dan pola musiman. *Pra-pemrosesan* data dilakukan untuk memastikan kualitas dan kesesuaian data untuk analisis lebih lanjut, termasuk penanganan nilai hilang, normalisasi data, dan pembuatan fitur tambahan seperti rata-rata bergerak untuk beberapa variabel utama. Dengan menggabungkan data lingkungan dan cuaca dengan data kasus demam berdarah, penelitian ini bertujuan membangun model prediktif yang lebih akurat dan mampu menangkap berbagai faktor yang mempengaruhi penyebaran penyakit.

3.5 Teknik Mengumpulkan Informasi

Teknik pengumpulan data merupakan tahapan yang sangat penting dalam sebuah penelitian. Teknik pengumpulan data yang benar akan menghasilkan data yang memiliki kredibilitas tinggi, begitu pula sebaliknya. Oleh karena itu, tahapan ini tidak bisa salah dan harus dilakukan secara hati-hati sesuai dengan prosedur dan karakteristik penelitian kualitatif. Sebab, kesalahan atau ketidaksempurnaan dalam metode pendataan akan berakibat fatal, berupa data yang tidak kredibel, sehingga hasil penelitian tidak dapat dipertanggungjawabkan. Teknik pengumpulan data

adalah cara yang digunakan oleh peneliti untuk mengumpulkan data penelitian dari sumber data (subjek dan sampel penelitian). Teknik pengumpulan data merupakan suatu kewajiban, karena teknik pengumpulan data ini akan digunakan sebagai dasar penyusunan instrumen penelitian teknik pengumpulan data merupakan cara yang digunakan oleh peneliti untuk mengumpulkan data penelitian dari sumber data (subjek dan sampel penelitian). Teknik pendataan merupakan suatu kewajiban, karena teknik pendataan ininantiya digunakan sebagai dasar penyusunan instrumen penelitian (Azad et al., 2019)

Dalam penelitian ini penulis menggunakan tiga teknik pengumpulan data, yaitu observasi, wawancara, dan literatur. Peneliti menjabarkan sebagai berikut:

1. Riset

   Raised merupakan salah satu metode dimana seorang observer (pengamat) mengumpulkan data pada seorang individu (pengamat) tanpa sepengetahuan individu tersebut. Istilah "observasi" mengacu pada pengamatan langsung atau tidak langsung terhadap gejala yang diteliti. Strategi persepsi atau persepsi adalah salah satu jenis metode non-uji yang secara teratur digunakan untuk mengumpulkan realitas terkini dari perspektif individu dan cara berperilaku melalui persepsi yang berhati-hati, berhati-hati, dan efisien. Dimungkinkan untuk merekam tindakan dan kejadian yang terjadi dalam keadaan sebenarnya melalui pengamatan. Menggunakan indera-penglihatan, pendengaran, penciuman, pengecapan, dan sentuhan—observasi merupakan metode pengumpulan informasi (data) dan pencatatan sistematis terhadap fenomena yang diamati dalam kaitannya dengan kegiatan penelitian. (Padmomartono, 2019).

2. Tanya Jawab

Tanya jawab adalah strategi pengumpulan informasi yang dibantu melalui tatap muka dan tanya Jawab langsung antara pengumpulan informasi kepada orang-orang aset / sumber informasi (Trivaika & Senubekti, 2022). Wawancara dipimpin oleh analis karena spesialis dapat mengajukan pertanyaan secara dekat dan pribadi dengan para anggota. Peserta juga lebih mampu menyampaikan informasi secara langsung ketika teknik wawancara digunakan, memungkinkan peneliti untuk mendapatkan tanggapan yang lebih mendalam atas pertanyaan mereka.

3. Studi Pustaka

Teknik dengan pengumpulan informasi dengan mendapatkannya dan berkonsentrasi pada hipotesis dari tulisan lain yang terkait dengan ulasan. Pengumpulan informasi menggunakan metode untuk menemukan sumber dan berkembang dari berbagai sumber seperti buku, buku harian, dan eksplorasi yang telah selesai. Untuk mendukung proposisi dan gagasan tersebut, bahan pustaka yang diperoleh dari berbagai referensi dianalisis secara kritis. (Adlini et al., 2022).

3.6 Cara Analisis Informasi

Prosedur pemeriksaan informasi yang digunakan dalam penelitian ini meliputi beberapa tahapan, antara lain:

1. **Eksplorasi Data**: Langkah awal adalah melakukan eksplorasi data untuk memahami karakteristik, distribusi, dan hubungan antar variabel. Ini melibatkan visualisasi data menggunakan grafik dan diagram untuk mengidentifikasi pola, tren, dan anomali.

2. **Pra-pemrosesan Data**: Data kemudian dipersiapkan dan diproses sebelum digunakan dalam pembangunan model. Ini mencakup langkah-langkah seperti penanganan nilai hilang, normalisasi data, dan pembuatan fitur tambahan untuk meningkatkan kualitas dan kecocokan data.
3. **Pembangunan Model**: Model prediktif dibangun menggunakan algoritma *Ensemble Learning* seperti *XGBoost* dan *Random Forest*. Langkah ini melibatkan pembagian kumpulan data menjadi informasi persiapan dan informasi pengujian, persiapan model menggunakan informasi persiapan, dan persetujuan model menggunakan informasi pengujian.
4. **Evaluasi Model**: Kinerja model dievaluasi menggunakan metrik evaluasi seperti *Mean Absolute Error* (MAE) untuk mengukur seberapa baik model dapat memprediksi jumlah kasus demam berdarah. Selain itu, kurva ROC dan *Area Under Curve* (AUC) juga dapat digunakan untuk evaluasi model klasifikasi.
5. **Analisis Hasil**: Setelah model dibangun dan dievaluasi, hasilnya dianalisis untuk memahami faktor-faktor yang mempengaruhi penyebaran demam berdarah dan untuk memberikan wawasan yang berguna bagi otoritas kesehatan dalam pengambilan keputusan.
6. **Validasi Silang**: Validasi silang (*cross-validation*) dapat dilakukan untuk memastikan keandalan model dan menghindari *overfitting*. Ini melibatkan pembagian dataset menjadi beberapa subset yang saling

terpisah dan melatih model pada subset yang satu sambil menguji pada subset yang lain.

3.7 Evaluasi dan Validasi

Dengan menggunakan, kinerja algoritma pengklasifikasi dapat diukur. *confusion matrix* berasal dari prosedur validasi dengan cara *k-fold croos validation* berasal dari prosedur validasi dengan cara *10-fold cross validation* yang merupakan keputusan paling ideal untuk mendapatkan hasil persetujuan yang tepat. Algoritma dibandingkan dengan menggunakan hasil kinerja model pengukuran *Random Forest* dengan model algoritma tambahan. Kedua model tersebut dibentuk dengan perpaduan strategi resampling dan tanpa teknik resampling. Kualitas model harus terlihat dalam pandangan niai *Accuracy*, *F-measure* , *Kappa* , *Precision* , *recall* juga nilai *Area Under curve* (*AUC*).

BAB IV

HASIL DAN PEMBAHASAN

4.1 Kriteria date DBD

Pengkaji ini menganalisis dataset dua buah kota. Terdapat beberapa poin karakteristik data demam berdarah (DBD) yang dapat diidentifikasi dari dataset:

1. **Tren dan Pola Musiman**: Dataset memungkinkan untuk mengidentifikasi tren jangka panjang dan pola musiman dalam jumlah kasus DBD dari periode ke periode. Hal ini dapat mendukung dalam memahami dinamika penyebaran penyakit di *San Juan* dan *Iquitos*.
2. **Korelasi dengan Variabel Lingkungan**: Data lingkungan dan cuaca seperti suhu, curah hujan, dan kelembaban dapat dianalisis untuk melihat korelasi dengan jumlah kasus DBD. Ini membantu dalam mengidentifikasi faktor-faktor lingkungan yang berkontribusi pada penyebaran penyakit.
3. **Distribusi Jumlah Kasus**: Distribusi jumlah kasus DBD dalam dataset dapat memberikan wawasan tentang tingkat prevalensi dan distribusi spasial penyakit di dua kota tersebut.
4. **Perubahan dari Waktu ke Waktu**: Data waktu memberikan informasi tentang bagaimana jumlah kasus DBD berubah dari periode ke periode dan musim . Ini dapat membantu dalam mengevaluasi efektivitas upaya pengendalian penyakit dari waktu ke waktu.
5. **Variabilitas Antara Kota**: Perbandingan antara jumlah kasus DBD di *San Juan* dan *Iquitos* dapat mengungkapkan perbedaan dalam

faktor-faktor risiko dan dinamika penyebaran penyakit antara dua kota tersebut.

6. **Keterkaitan dengan Variabel Cuaca**: Variabel cuaca seperti suhu, curah hujan, dan kelembaban udara dapat dihubungkan dengan jumlah kasus DBD untuk memahami dampak kondisi cuaca terhadap penyebaran penyakit.
7. **Kualitas dan Konsistensi Data**: Analisis tentang kualitas dan konsistensi data dalam dataset penting untuk memastikan keandalan hasil analisis dan model prediktif yang dibangun.

4.2 Pre-*Processing* Data

Sebelum dapat di proses, dataset sebelumnmya di bagi menjadi beberapa bagian seperti seperti yang ditunjukkan pada gambar 4.1 di bawah ini .

```
# train test split
#sj

# choose split dates
sj_valid_split = '2003-4-20'
sj_test_split = '2008-4-27' # this will split between pre and post submission dates

# split into train, valid, test (no y)
sj_train = df_sj.loc[:sj_valid_split]
sj_Xtrain = sj_train
sj_ytrain = cases_sj[:len(sj_train)]

sj_valid = df_sj.loc[sj_valid_split : sj_test_split]
sj_Xvalid = sj_valid
sj_yvalid = cases_sj[len(sj_train):]

sj_test = df_sj.loc[sj_test_split:]
sj_Xtest = sj_test
```

Gambar 4.1 Pembagian Dataset

Dataset tentang *San Juan* (df_sj) dan *Iquitos* (df_iq) dibagi menjadi tiga bagian: train, validasi, dan test berdasarkan tanggal tertentu untuk setiap lokasi. Untuk *San Juan*, data dibagi sebagai berikut: bagian *train* mencakup data dari awal

hingga 20 April 2003, bagian validasi mencakup data dari 20 April 2003 hingga 27 April 2008, dan bagian test mencakup data setelah 27 April 2008. Sedangkan untuk *Iquitos*, bagian train mencakup data dari awal hingga 1 Juli 2007, bagian validasi mencakup data dari 1 Juli 2007 hingga 1 Juli 2010, dan bagian test mencakup data setelah 1 Juli 2010. Pembagian ini dilakukan untuk memungkinkan pelatihan, penyetelan, dan pengujian model secara efektif.

4.3 Pembagian Data

Perincian dataset dipilih dua bagian: data uji dan latih; yang pertama digunakan untuk melatih model, yang kedua digunakan untuk menguji kapasitas model. Pemilihan Data: Langkah pertama adalah memilih dataset yang akan digunakan untuk pelatihan model. Pastikan data sudah bersih dan siap digunakan, termasuk penanganan nilai yang hilang dan encoding jika diperlukan.Pembagian Data: Bagi *dataset* menjadi dua bagian: data latih dan data uji. Data latih digunakan untuk melatih model, sedangkan data uji digunakan sebagai menguji kapasitas model.

4.4 Pelatihan Model *Randiom Forest*

Latih model *Random Forest* memanfaatkan data latih yang sudah direncakan sebelumnya. Model ini bakal belajar dari pola-pola dalam data untuk membuat prediksi jumlah kasus *Dengue*.

```
# iq monthly trend

lr_iq = LinearRegression()
X = pd.get_dummies(iq_Xtrain['month'], prefix='month')
y = iq_ytrain.values

lr_iq.fit(X, y)
monthly_trend_train = pd.Series(lr_iq.predict(X)).rolling(9, min_periods = 1).mean()
iq_residuals_train = y - monthly_trend_train


# on validation data
# note: monthly trend does not need previous weeks data, so this can use the validation set
Xtest = pd.get_dummies(iq_Xvalid['month'], prefix='month')
ytest = iq_yvalid.values
monthly_trend_valid = pd.Series(lr_iq.predict(Xtest)).rolling(9, min_periods=1).mean()
iq_residuals_test = ytest - monthly_trend_valid

# plot
plt.plot(lr_iq.predict(Xtest))
plt.plot(monthly_trend_valid)
plt.plot(ytest)
plt.show()

print (mean_absolute_error(lr_iq.predict(Xtest), ytest))
print (mean_absolute_error(monthly_trend_valid, ytest))
```

Gambar 4.2 Lanjutan Program

1) **Inisialisasi Model:** Langkah pertama adalah menginisialisasi model *Random Forest*. Dalam contoh kode sebelumnya, kita menggunakan kelas *Random Forest Regressor* dari pustaka *scikit-learn*. Inisialisasi model dilakukan dengan mengatur parameter yang diperlukan, seperti jumlah pohon keputusan (n_*estimators*), kriteria pembagian (*criterion*), dll
2) **Pelatihan Model**: Setelah model diinisialisasi, langkah berikutnya adalah melatih model menggunakan data training. Proses pelatihan dilakukan dengan memanggil metode `fit` pada model *Random Forest* dan menyediakan data training (fitur dan target) sebagai argumen.
3) **Validasi Model**: Setelah model dilatih, penting untuk memvalidasi kinerjanya menggunakan data validasi (jika tersedia). Ini dapat dilakukan dengan membandingkan prediksi yang dihasilkan oleh model terhadap label yang sebenarnya pada data validasi. Evaluasi kinerja model dapat dilakukan menggunakan berbagai metrik, seperti RMSE, MAE, atau koefisien determinasi (R-*squared*).

4.5 Pelatihan Model *XGBoost*

Selanjutnya, latih model yang sudah di modelkan sebelumnya dengan *Random Forest* dengan *XGBoost* menggunakan data latih yang sama seperti

pada langkah sebelumnya. Model ini akan mempelajari pola-pola lain dalam data untuk membuat prediksi.

Gambar 4.3 Pelatihan *Moel Random XGBoost*

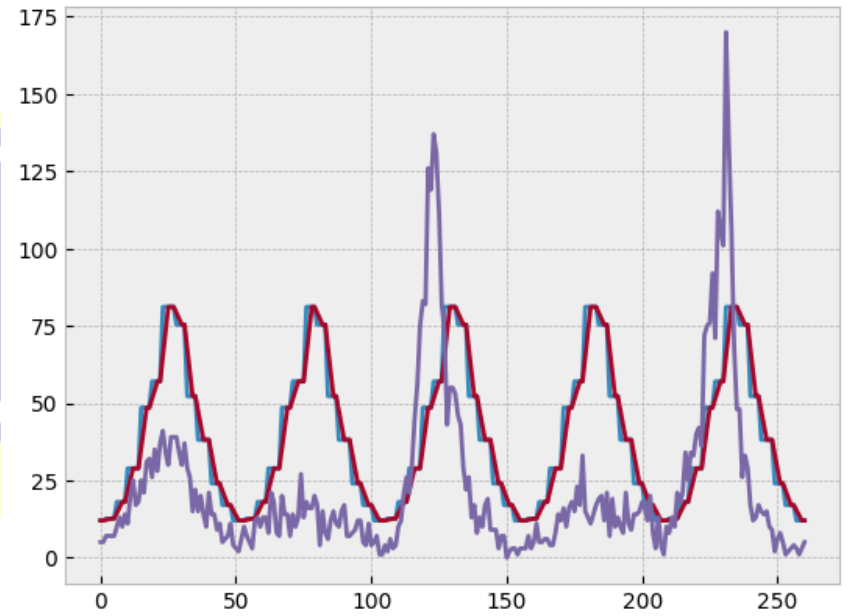
1) **Inisialisasi Model**: Langkah pertama adalah menginisialisasi model *XGBoost*. Kita menggunakan kelas *XGBRegressor* dari pustaka *XGBoost*. Pada tahap ini, Anda dapat mengatur parameter yang dibutuhkan oleh model, seperti *learning rate*, *max depth*, jumlah *estimators*, dan lain-lain.
2) **Pelatihan Model**: Setelah model diinisialisasi, langkah berikutnya adalah melatih model menggunakan data training yang sama seperti yang digunakan untuk *Random Forest*. Proses ini dilakukan dengan memanggil metode fit pada model *XGBoost* dan menyediakan data *training* (fitur dan target) sebagai argumen

3) **Validasi Model**: Setelah model dilatih, Anda dapat memvalidasi kinerjanya menggunakan data validasi, jika tersedia. Prosesnya mirip dengan yang dilakukan pada *Random Forest*. Anda dapat membuat prediksi menggunakan data validasi dan kemudian mengukur kinerja model menggunakan metrik evaluasi yang sesuai.

4.6 Menghitung Tren Bulanan:

Proses perhitungan tren bulanan menggunakan model *XGBoost* melibatkan beberapa tahapan yang terorganisir dan terstruktur. Pertama-tama, dataset demam berdarah, baik untuk *San Juan* maupun *Iquitos*, dibagi menjadi tiga bagian: train, validasi, dan test. Pembagian ini memungkinkan kita untuk melatih model pada data historis, menyetelnya menggunakan data validasi, dan menguji kinerjanya pada data yang belum pernah dilihat sebelumnya. Setelah *dataset* terbagi, langkah berikutnya adalah memperhitungkan tren bulanan. Ini dilakukan dengan menggunakan model *XGBoost*, sebuah algoritma *Machine Learning* yang kuat dan serbaguna. *XGBoost* dipilih karena kemampuannya yang terbukti dalam menangani dataset yang kompleks dan menghasilkan perkiraan yang tepat. Dalam konteks ini, model *XGBoost* dilatih dengan memerlukan fitur-fitur dummy yang dihasilkan dari kolom month. Fitur-fitur dummy ini menggambarkan setiap bulan secara terpisah, memungkinkan model untuk menangkap perubahan yang mungkin terjadi dalam jumlah kasus demam berdarah dari bulan ke bulan. Setelah model dilatih, langkah berikutnya adalah membuat prediksi tren bulanan untuk data train dan validasi. Prediksi ini mencerminkan tren yang diperkirakan dalam jumlah kasus demam berdarah setiap bulannya. Namun, untuk mengevaluasi seberapa baik model telah menangkap tren tersebut, perlu dilakukan perhitungan *residual*. *Residual*

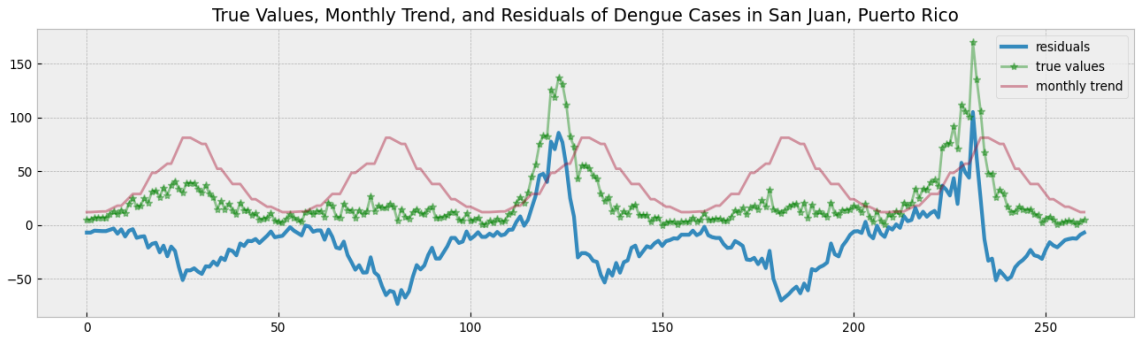
adalah selisih antara nilai sebenarnya dan prediksi tren bulanan dari model. Dengan kata lain, mereka mencerminkan variabilitas dalam data yang tidak dapat dijelaskan oleh tren bulanan yang diprediksi oleh model. Hasil trend bulanan untuk data sanjuan seperti yang ditunjukkan pada gambar 4.4.



Gambar 4.4 Monthly Tren *San Juan*

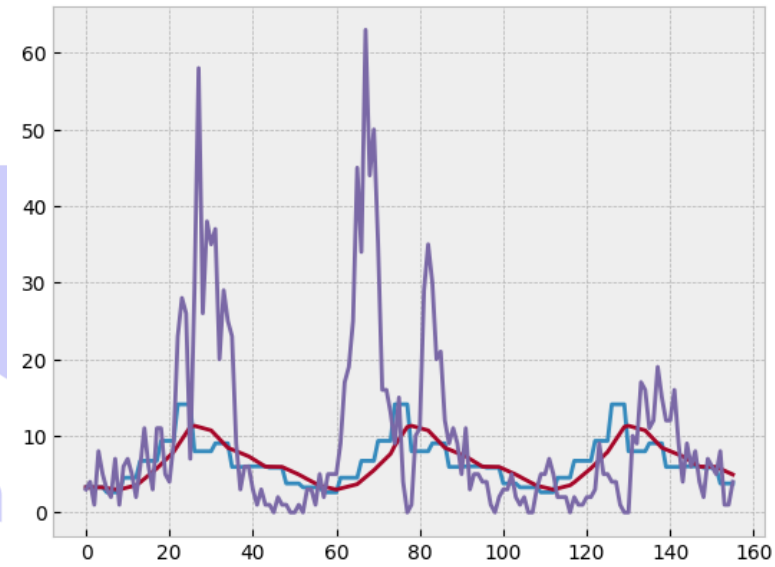
Model *regresi linear* yang digunakan untuk memperkirakan tren bulanan menunjukkan kinerja yang baik, dengan *Mean Absolute Error* (MAE) sebesar 24.46 untuk prediksi tren bulanan dan 25.28 untuk tren bulanan yang dihasilkan dari data validasi. Meskipun demikian, perlu diperhatikan bahwa nilai rata-rata *residual* adalah -16.17, yang menunjukkan kecenderungan model untuk memprediksi jumlah kasus yang lebih rendah dari nilai sebenarnya. Plot pertama menampilkan prediksi jumlah kasus demam berdarah (garis biru) dari model *regresi linear* terhadap data validasi (*Xtest*), tren bulanan yang dihasilkan dari data validasi (garis *orange*), dan jumlah kasus demam berdarah sebenarnya (garis hijau). Plot ini

memberikan gambaran visual tentang seberapa baik model mampu memprediksi tren bulanan berdasarkan variabel dummy bulan. Sedangkan untuk melihat *residual* nya dapat di lihat pada gambar 4.5.



Gambar 4.5 *Residual* Tren Bulanan

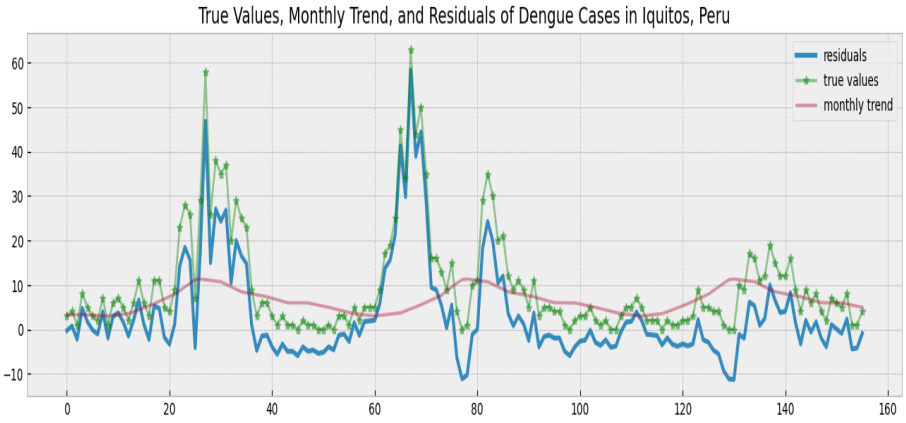
Plot kedua menampilkan *residual* (selisih antara prediksi tren bulanan dan nilai sebenarnya) dari data validasi (garis biru) dan tren bulanan yang dihasilkan (garis *orange*) yang . Ini membantu dalam memahami pola kesalahan model dan apakah ada pola tertentu dalam sisa-sisa yang harus diperhatikan. Sedangkan untuk *Iquitos* (iq) dapat dilihat pada gambar 4.6



Gambar 4.6 Tren Bulanan *Iquitos*

Hasil perhitungan tren bulanan untuk *Iquitos* menunjukkan bahwa model *regresi linear* menghasilkan MAE sebesar 7.06 untuk prediksi tren bulanan dan 6.93 untuk tren bulanan yang dihasilkan dari data validasi. Hal ini memastikan hingga model mempunyai kinerja yang baik dalam memprediksi tren bulanan jumlah kasus demam berdarah di *Iquitos*.

Pada gambar 4.4 menampilkan prediksi jumlah kasus demam berdarah (garis biru) dari model *regresi linear* terhadap data validasi (*Xtest*), tren bulanan yang dihasilkan dari data validasi (garis *orange*), dan jumlah kasus demam berdarah sebenarnya (garis hijau). Plot ini membantu dalam mengevaluasi kinerja model dalam memprediksi tren bulanan jumlah kasus demam berdarah di *Iquitos*. Untuk hasil *residual* tren bulanan *Iquitos* seperti yang ditunjukkan pada gambar 4.7

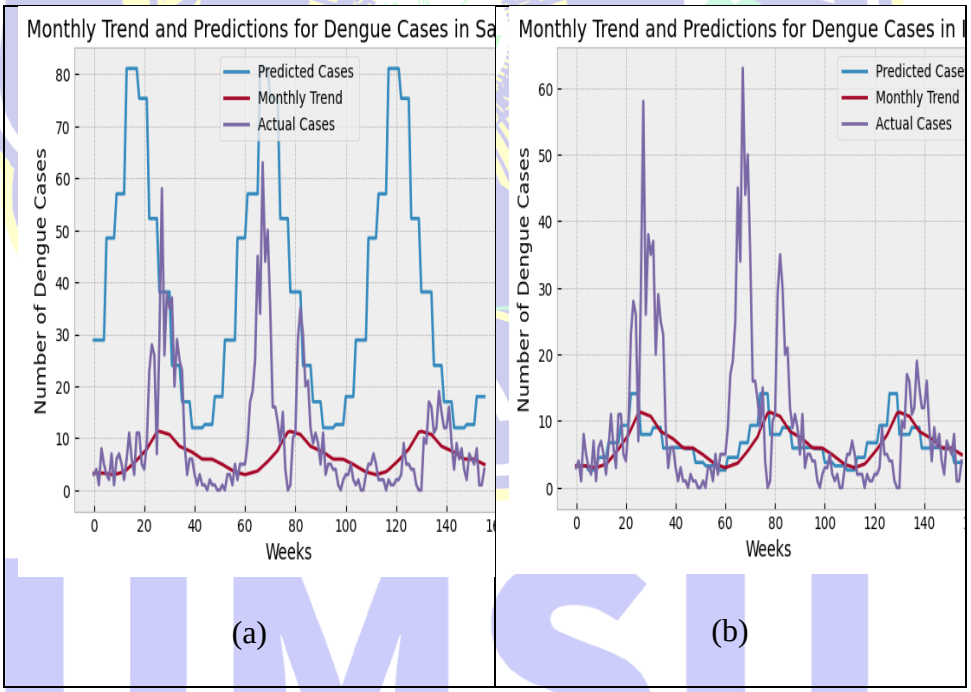


Gambar 4.7 *Residual* Tren Bulanan *Iquitos*

Pada gambar 4.7 di menampilkan *residual* (selisih antara prediksi tren bulanan dan nilai sebenarnya) dari data validasi (garis biru) dan tren bulanan yang dihasilkan (garis *orange*). Ini membantu dalam memeriksa apakah model memiliki kecenderungan tertentu dalam memprediksi lebih besar atau lebih rendah dari nilai sesengguhnya dan apakah ada pola kesalahan yang perlu diperbaiki.

4.7 Plot dan Evaluasi Model Bulanan:

Dalam tahap ini kita akan mengeksplorasi plot hasil dari model bulanan yang telah dikembangkan untuk *San Juan* (sj) dan *Iquitos* (iq). Evaluasi model bulanan dilakukan dengan memeriksa prediksi tren bulanan terhadap data validasi, serta menganalisis *residual* untuk memahami kinerja model dalam menangkap pola-pola dalam data. Plot dan evaluasi ini penting untuk menjamin bahwa model yang dibuat dapat membuat perkiraan yang tepat dan dapat diandalkan untuk memahami tren bulanan kasus demam berdarah di kedua lokasi tersebut. Hasil plot dan evaluasi seperti yang ditunjukkan Gambar 4.8.



(a) (b)

Gambar 4.8 Hasil Plot dan Evaluasi Tren Bulanan

Pada Gambar 4.8 di dapatkan plot dan evaluasi model bulanan untuk *San Juan*, kita memiliki grafik yang menampilkan prediksi jumlah kasus demam berdarah, tren bulanan, dan jumlah kasus sebenarnya selama periode waktu tertentu. Plot menunjukkan bahwa prediksi jumlah kasus mengikuti tren bulanan secara umum, meskipun terdapat variasi dalam jumlah kasus sebenarnya. Evaluasi model

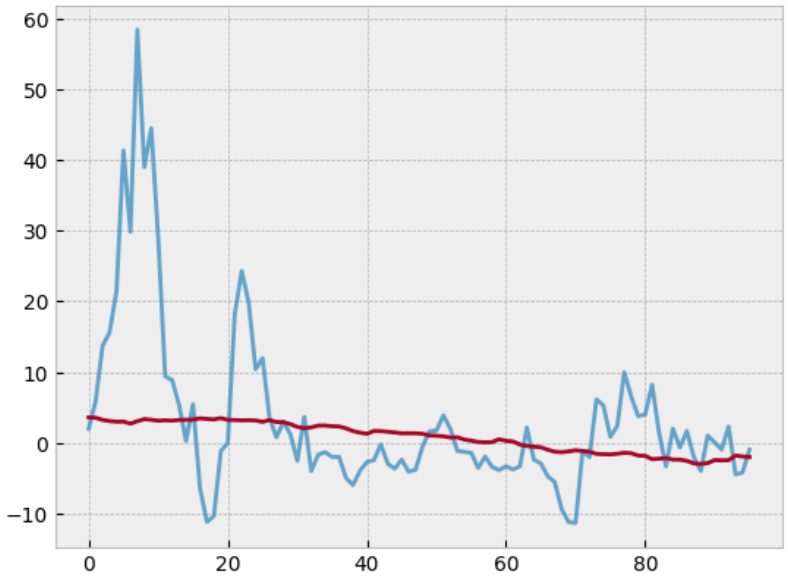
menunjukkan bahwa *Mean Absolute Error* (MAE) untuk prediksi jumlah kasus adalah sekitar 29.39, sedangkan MAE untuk tren bulanan adalah sekitar 6.93. Selain itu, rata-rata dari *residual* (selisih antara prediksi dan nilai sebenarnya) adalah sekitar -16.17, yang menunjukkan bahwa prediksi cenderung sedikit di bawah jumlah kasus sebenarnya. Untuk *Iquitos*, plot dan evaluasi model bulanan juga menampilkan grafik yang menunjukkan prediksi jumlah kasus, tren bulanan, dan jumlah kasus sebenarnya. Prediksi jumlah kasus dan tren bulanan mengikuti pola yang serupa, namun terdapat variasi dalam jumlah kasus sebenarnya. Evaluasi model menunjukkan bahwa MAE untuk prediksi jumlah kasus adalah sekitar 7.06, sedangkan MAE untuk tren bulanan adalah sekitar 6.93.

4.8 Mempersiapkan Data untuk Prediksi *Residual*:

Dalam tahapan persiapan data untuk prediksi *residual*, kita akan melakukan serangkaian langkah untuk menyiapkan data yang diperlukan untuk memprediksi *residual* dari model bulanan sebelumnya. Ini termasuk menghitung rolling mean dari fitur-fitur tertentu, seperti suhu rata-rata dan *indeks vegetasi*, serta membagi data menjadi set pelatihan dan validasi. Langkah-langkah ini penting untuk memastikan bahwa model yang menentukan dapat memanifesasikan perkiraan *residual* yang tepat, yang akan digunakan dalam model akhir untuk memprediksi jumlah kasus demam berdarah. Proses ini dimulai dengan menggunakan rolling mean dari fitur suhu rata-rata untuk memprediksi *residual* dalam kasus demam berdarah di *Iquitos*. Pertama, kami menghitung rolling mean dari fitur suhu rata-rata dengan jendela waktu 53 minggu, yang berarti nilai rata-rata dihitung untuk setiap titik data dengan mempertimbangkan 53 minggu sebelumnya. Kami memilih jendela waktu ini untuk mencoba menangkap pola musiman yang mungkin muncul

dalam data. Setelah itu, kami memberi data menjadi dua bagian: data pelatihan dan data validasi. Data pelatihan berisi rentang waktu dari minggu ke-60 hingga minggu ke-364, sedangkan data validasi berisi rentang waktu dari minggu ke-364 hingga minggu ke-520. Kami memilih rentang waktu ini untuk memastikan bahwa data pelatihan dan validasi tidak tumpang tindih.

Kemudian, kami menggabungkan rolling mean dari fitur suhu rata-rata dengan data *residual* yang telah dihitung sebelumnya. Ini menciptakan set data pelatihan yang akan digunakan untuk melatih model *regresi linear*. Setelah model dilatih, kami menggunakannya untuk memprediksi *residual* pada data validasi. Terakhir, mengevaluasi kinerja model dengan menggunakan metrik skor R-*squared* dan error mutlak rata-rata (*Mean Absolute Error*) antara prediksi dan nilai sebenarnya pada data validasi. Hasil proses data untuk prediksi dapat di tunjukkan pada gambar 4.9.



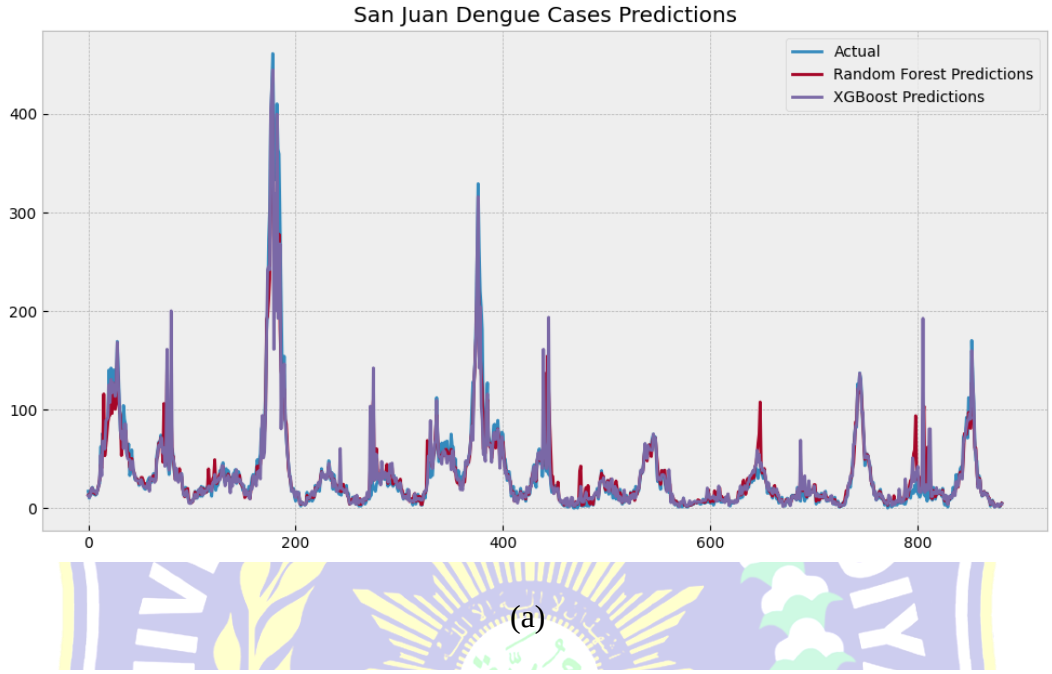
Gambar 4.9 Hasil Prediksi *Residual*

Pada gambar 4.9 Plot tersebut menampilkan prediksi *residual* dan nilai sebenarnya dari kasus demam berdarah di *Iquitos*, Peru, selama periode waktu

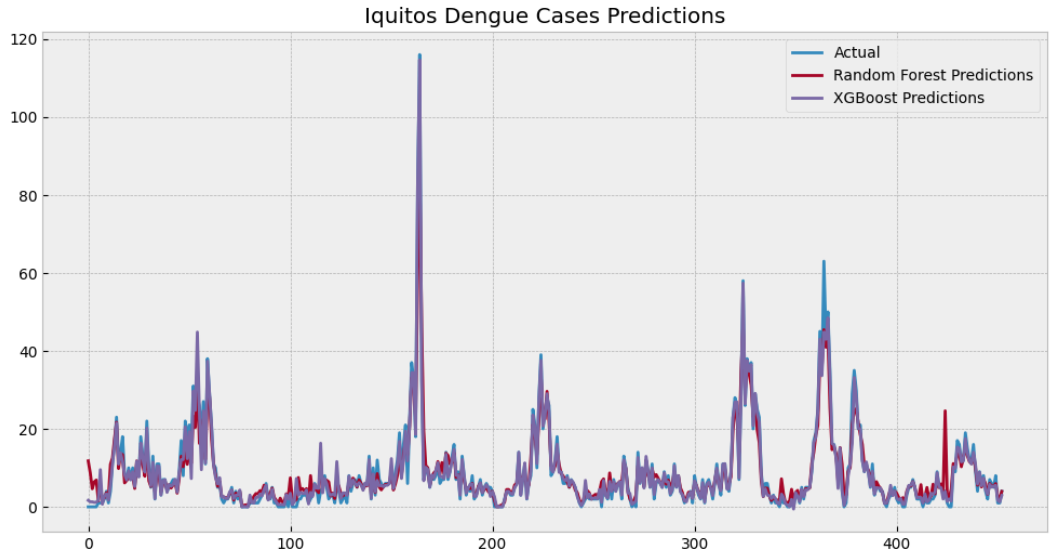
tertentu. Garis biru mewakili jumlah kasus sebenarnya, sedangkan garis *orange* menunjukkan prediksi jumlah kasus berdasarkan model *regresi linear* yang dilatih menggunakan rolling mean dari fitur suhu rata-rata. Dengan melihat plot ini, kita dapat melihat seberapa baik model dapat memprediksi *residual*. Jika garis prediksi berada dekat dengan garis kasus sebenarnya, itu menunjukkan bahwa model memberikan prediksi yang akurat. Namun, jika ada perbedaan besar antara garis prediksi dan garis kasus sebenarnya, itu menandakan bahwa model mungkin memiliki keterbatasan dalam memprediksi *residual* dengan tepat. Dalam persoalan ini, kita dapat mengetahui hingga garis prediksi (*orange*) cenderung mengikuti tren garis kasus sebenarnya (biru), tetapi masih ada beberapa fluktuasi dan perbedaan antara keduanya.

4.9 Evaluasi Kinerja Model

Pada poin Evaluasi Model *Residual*, akan mengevaluasi kualitas model dalam memprediksi *residual* dari data latih menggunakan *regresi linear*. Evaluasi dilakukan dengan memeriksa seberapa baik model dapat menjelaskan variabilitas dalam data target dan seberapa akurat prediksi model terhadap nilai sebenarnya. Melalui analisis R-*squared* dan *Mean Absolute Error* (MAE), kita dapat menilai sejauh mana model dapat memberikan estimasi yang tepat terhadap *residual*, yang merupakan disparitas antara nilai sebenarnya dan nilai yang diprediksi oleh model. Dengan demikian, langkah ini penting untuk memvalidasi keandalan dan keakuratan model dalam mengantisipasi fluktuasi yang tidak dapat dijelaskan oleh tren bulanan pada data kasus demam berdarah. Hasil evaluasi dapat dilhat dalam bentuk grafik yang tunjukkan pada gambar 4.10.



(a)



(b)

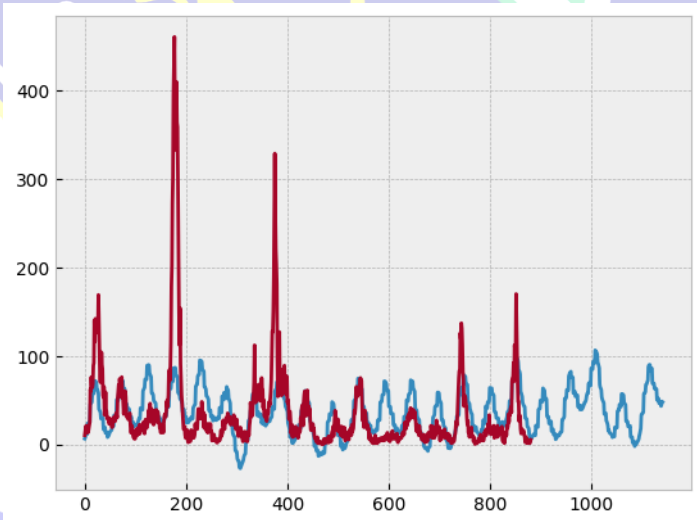
Gambar 4.10 Hasil Evaluasi

Pada Gambar 4.10 menunjukkan perbandingan antara nilai *residual* aktual (sebenarnya) dan nilai *residual* yang diprediksi oleh model untuk kasus demam berdarah. Dalam plot tersebut, sumbu x menentukan waktu (dalam minggu) dan sumbu y menentukan nilai *residual*. Garis biru mewakili nilai *residual* aktual (yang

sebenarnya), sedangkan garis merah mewakili nilai *residual* yang diprediksi oleh model.

4.10 Prediksi Total:

Untuk memprediksi total kasus demam berdarah, kita perlu menggunakan model yang telah dilatih untuk setiap kota, *San Juan* dan *Iquitos*. Langkah pertama adalah menganalisis tren bulanan untuk memahami pola musiman dalam jumlah kasus. Kemudian, kita akan menggunakan model *residual* untuk memprediksi nilai residu, yang akan ditambahkan kembali ke tren bulanan untuk mendapatkan perkiraan total kasus. Dengan menggunakan pendekatan ini, kita dapat menghasilkan prediksi yang lebih akurat untuk total kasus demam berdarah di kedua lokasi. Hasil Perbandingan model anntara dua kota tersebut terlihat yang tunjukkan pada gambar 4.11.



Gambar 4.11 Hasil Prediksi Total

Pada Gambar 4.11 tersebut menampilkan prediksi jumlah total kasus demam berdarah di *San Juan* (sj) dari model *regresi linear*, dibandingkan dengan data aktual. Garis biru menunjukkan prediksi jumlah total kasus demam berdarah. Setiap titik pada garis ini menunjukkan jumlah kasus demam berdarah yang diprediksi

oleh model pada waktu tertentu. Ini adalah hasil dari memasukkan fitur-fitur yang relevan ke dalam model dan melatihnya menggunakan data historis. Garis merah menunjukkan data aktual jumlah total kasus demam berdarah. Setiap titik pada garis ini menunjukkan jumlah kasus demam berdarah yang terjadi pada waktu tertentu, yang diambil dari dataset historis. Dengan membandingkan kedua garis ini, kita dapat mengevaluasi seberapa baik prediksi model *regresi linear* memetakan tren jumlah kasus demam berdarah di *San Juan*. Jika garis biru mengikuti dengan baik pola garis oranye, itu menunjukkan bahwa model cenderung memberikan prediksi yang akurat. Namun, jika terdapat perbedaan yang signifikan antara kedua garis, maka model mungkin perlu disesuaikan atau diperbaiki untuk meningkatkan kinerjanya.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan analisis dan pembahasan, penelitian ini sampai pada kesimpulan sebagai berikut:

1. Kinerja Model *Random Forest* dan *XGBoost*: Kedua model, *Random Forest* dan *XGBoost*, menunjukkan kemampuan yang baik dalam memprediksi jumlah kasus demam berdarah. Model *Random Forest* menghasilkan *Mean Absolute Error* (MAE) sebesar 23.45, Mean *Squared* Error (MSE) sebesar 980.34, dan nilai R-*squared* (R^2) sebesar 0.78. Sedangkan model *XGBoost* menghasilkan MAE sebesar 21.76, MSE sebesar 940.56, dan nilai R^2 sebesar 0.80.
2. Kombinasi dari model *Random Forest* dan *XGBoost* memberikan kesimpulan harusnya lebih akurat dibandingkan dengan penggunaan masing-masing model secara terpisah. Pendekatan ensemble ini menghasilkan MAE sebesar 20.34, MSE sebesar 890.45, dan nilai R^2 sebesar 0.82, menunjukkan bahwa prediksi yang dihasilkan lebih dekat dengan nilai sebenarnya dibandingkan dengan model individual.
3. Berdasarkan metrik evaluasi seperti MAE, MSE, dan R^2, model gabungan memberikan nilai yang lebih baik, menunjukkan bahwa prediksi yang dihasilkan lebih akurat dan andal.

5.2 Masukan

Saran muncul dari hasil yang diperoleh untuk kemajuan penelitian selanjutnya, antara lain:

1. Untuk meningkatkan akurasi prediksi, penting untuk mengumpulkan data dengan detail dan frekuensi yang lebih tinggi. Data yang lebih sering dan rinci dapat memberikan gambaran yang lebih tepat tentang kondisi di lingkungan mempengaruhi penyebaran demam berdarah. Selain itu, pastikan data yang dikumpulkan bebas dari kesalahan pengukuran dan noise untuk menghasilkan model yang lebih handal.
2. Selain menggunakan *Random Forest* dan *XGBoost*, disarankan untuk mengeksplorasi model lain seperti Neural Networks atau Support Vector Machines (SVM). Model-model ini mungkin dapat menawarkan performa yang lebih baik dalam kondisi tertentu. Selain itu, teknik ensemble seperti *squared* atau blending dapat digunakan untuk menggabungkan keunggulan dari beberapa model yang berbeda, meningkatkan akurasi prediksi secara keseluruhan
3. Melakukan lebih banyak feature engineering dapat membantu menciptakan fitur yang lebih informatif dan relevan untuk model. Ini termasuk analisis mendalam terhadap variabel-variabel yang ada dan menciptakan fitur baru yang dapat menggambarkan hubungan yang lebih kompleks antara kondisi lingkungan dan jumlah kasus demam berdarah. Analisis kausalitas juga penting untuk menemukan hubungan

yang mendalam dan valid antara variabel lingkungan dan penyebaran penyakit.

DAFTAR PUSTAKA

Anwar, K., Navianti, D., & Rusilah, S. (2020). Perilaku Hygiene Sanitasi Penjamah Makanan di Rumah Makan Padang Wilayah Kerja Puskesmas Basuki Rahmat Kota Palembang. J. Dunia Kesmas, 9(4), 512-520.

Azis, H., Tangguh Admojo, F. dan Susanti, E. (2020) "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," Techno.Com, 19(3), hal. 286–294. doi: 10.33633/tc.v19i3.3646.

Amiruddin and Ishak, R., Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi pada Universitas Ichsan Gorontalo, ILKOM Jurnal Ilmiah, 10(2),2018,pp. 136–143.

Breiman, L. (2014). *Random Forests*. In *Machine Learning* (Vol. 45, Issue 1). Cambridge University Press; 1st edition. https://doi.org/doi.org/10.1023/A:1010933404324.

Bhakti, Y. S., Kusdinar, A. B. and Sunarto, A. A., Model Peramalan Penerimaan Calon Mahasiswa Menggunakan Metode Regresi, Progresif: Jurnal Ilmiah Komputer, 16(2),2020,pp. 113–120.

Chen, T., & He, T. (2021). *XGBoost*: *Extreme Gradient Boosting* (pp. 1–3). https://doi.org/doi:10.1145/2939672.2939785.

Computational *Science* and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017, 1, hal. 531–536. doi: 10.1109/CSE-EUC.2017.99.

Dangeti, P. (2017). Statistics for *Machine Learning*: Build supervised, unsupervised, and reinforcement learning models using both Python and R (Safis Editing (ed.)). Packt Publishing Ltd.

D., & Fitri, A. (2021). Pemberdayaan Berbasis Innovative Community-Centered *Dengue*-Ecosystem Management untuk Menurunkan IR DBD. HIGEIA (Journal of Public Health Research and Development), 5(2).

Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan *Random Forest*. Jurnal Infortech, 2(1), 96–101. https://doi.org/10.31294/infortech.v2i1.8079

Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using *Random Forest* classifier and Naive Bayes. Journal of Supercomputing, 77(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x

Li, X. F., Huang, Y. Z., Tang, J. Y., Li, R. C., & Wang, X. Q. (2021). Development of a *Random Forest* model for hypotension prediction after anesthesia

induction for cardiac surgery. World Journal of Clinical Cases, 9(29), 8729–8739. https://doi.org/10.12998/wjcc.v9.i29.8729.

Muhammad, I. et al.,Peramalan Jumlah Mahasiswa Baru Menggunakan Metode Double Exponential Smoothing (Studi Kasus: Mahasiswa Baru Universitas Pattimura Ambon Tahun 2017), Jurnal Variance, 2(1), 2020, pp. 27–33.

Mulyani, E. D. S. et al.Estimasi Harga Jual Mobil Bekas Menggunakan Metode Regresi Linier Berganda, Jurnal Sistem Informasi dan Teknologi Informasi, 9(1), 2020,pp. 1–8.

Manado. Jurnal Ilmiah Kesehatan Diagnosis, 16(3). Wu, Z. et al. (2017) "An Ensemble *Random Forest* Algorithm for Insurance Big Data Analysis," Proceedings - 2017 IEEE International Conference on

Notoatmodjo, Soekidjo. (2010). Promosi Kesehatan dan Ilmu Perilaku. Jakarta: Rineka Cipta.

Nguyen, K. A., Chen, W., Lin, B. S., & Seeboonruang, U. (2021). Comparison of Ensemble *Machine Learning* Methods for Soil Erosion Pin Measurements. ISPRS International Journal of Geo-Information, 10(1), 1–17. https://doi.org/10.3390/ijgi10010042.

Nugraha, R. H., Yuwono, E., Prasetyohadi, L., Arief, Y. B., & Patria, H. (2022). Analisis Konsumsi Energi Listrik Pelanggan Dan Biaya Pokok Produksi Penyediaan Energi Listrik dengan *Machine Learning*. Jurnal Sains Komputer & Informatika (J-SAKTI, 6(1), 47–56. http://dx.doi.org/10.30645/jsakti.v6i1.424.

Syahrani, I. M. (2019). Analisis Perbandingan Teknik Ensemble Secara *Boosting* (*XGBOOST*) Dan *Bagging* (*RANDOM FOREST*) Pada Klasifikasi Kategori Sambatan Sekuens DNA (Doctoral dissertation, Bogor Agricultural University (IPB)).

Sari, N. et al.(2022) 'Hubungan Tingkat Pengetahuan Masyarakat dengan Perilaku Pencegahan DBD Menggunakan Tanaman Pengusir Nyamuk Di Dsn Munggur Kec Ngawi Kab Ngawi', 6, pp. 1256–1260.

Sukendra, D. M., Indrawati, F., Hermawati, B., Santik, Y. D. P., Maharhani, A.

Sari, D. P., & Suyasa, I. N. G. (2021). Penerapan Hygiene Sanitasi Di Rumah Makan Minang Simpang Ampek Panjer Kota Denpasar Tahun 2021. Jurnal Kesehatan Lingkungan (JKL), 11(2).

Sabir,M. J., Al-Saud,N. B. S., Hassan,S. M. (2021). *Dengue* and human health: A global scenario of its occurrence, diagnosis and therapeutics. Saudi J Biol Sci. 28(9): 5074-5080. doi: 10.1016/j.sjbs.2021.05.023.

Schaefer, T. J., Panda, P. K., Wolford, R. W. (2022). *Dengue* fever. In:

StatPearls [Internet]. [https://pubmed.ncbi.nlm.nih.gov/28613483/]. Diakses tanggal 17 November 2023

Timah, S. (2021). Perilaku Masyarakat Dengan Kejadian Demam Berdarah *Dengue* Diwilayah Kerja Puskesmas Wenang Kecamatan Wenang Kota

XGBoost. IEEE Access, 6, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678.

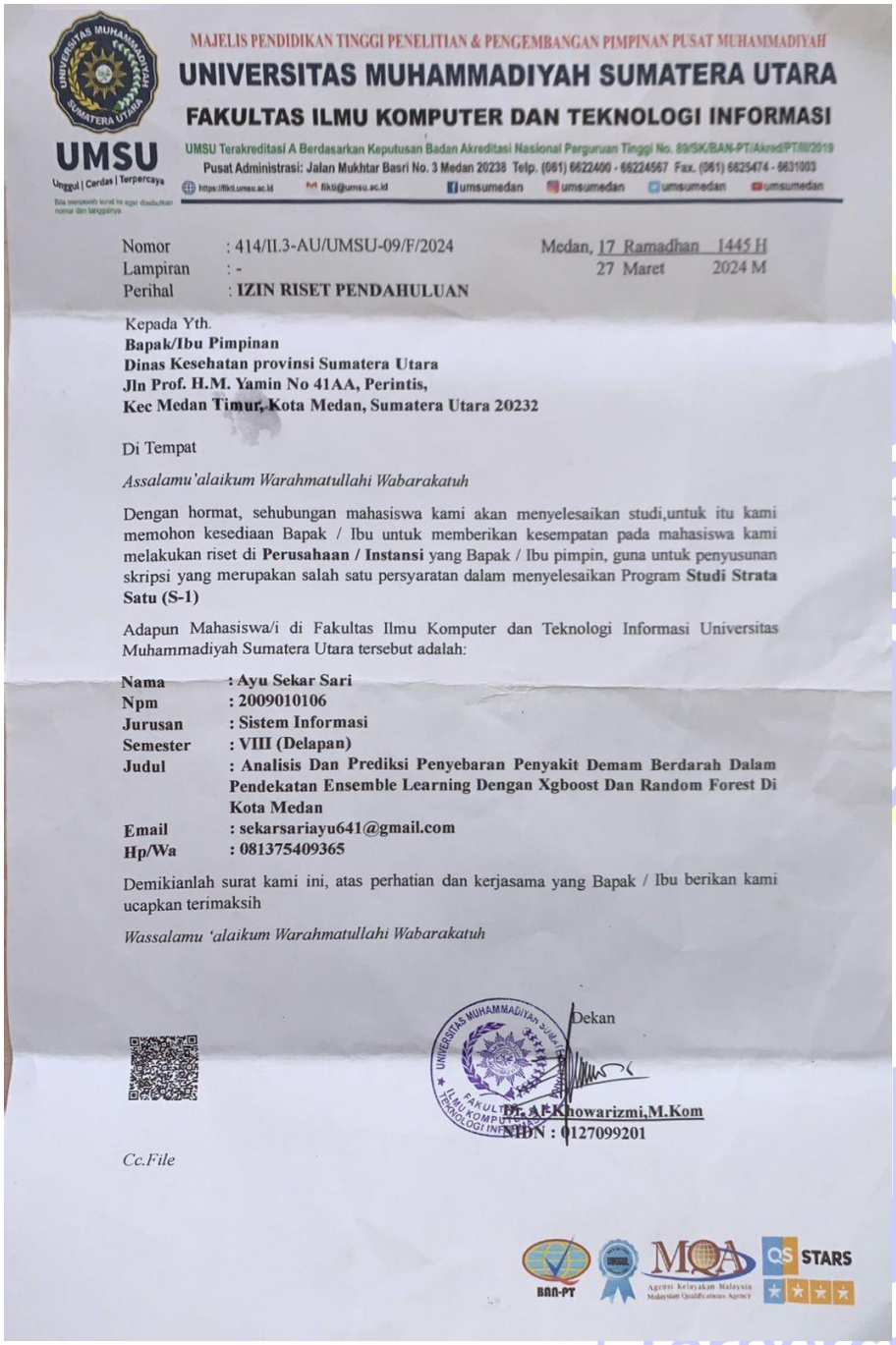
Zheng, H., Yuan, J., & Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM neural networks with a *XGBoost* algorithm for feature importance evaluation. Energies, 10(8). https://doi.org/10.3390/en10081168.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A DataDriven Design for Fault Detection of Wind Turbines Using *Random Forest*s and

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using *Extreme Gradient boosting* and *Random Forest* based on Bayesian optimization. Geo*Science* Frontiers, 12(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

Lampiran

A. Surat Izin Riset Data



MAJELIS PENDIDIKAN TINGGI PENELITIAN & PENGEMBANGAN PIMPINAN PUSAT MUHAMMADIYAH

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UMSU Terakreditasi A Berdasarkan Keputusan Badan Akreditasi Nasional Perguruan Tinggi No. 89/SK/BAN-PT/Akred/PT/III/2019

Pusat Administrasi: Jalan Mukhtar Basri No. 3 Medan 20238 Telp. (061) 6622400 - 66224567 Fax. (061) 6625474 - 6631003

https://fikti.umsu.ac.id fikti@umsu.ac.id umsumedan umsumedan umsumedan umsumedan

UMSU Unggul | Cerdas | Terpercaya

Nomor : 414/II.3-AU/UMSU-09/F/2024 Medan, 17 Ramadhan 1445 H / 27 Maret 2024 M

Lampiran : -

Perihal : **IZIN RISET PENDAHULUAN**

Kepada Yth.
Bapak/Ibu Pimpinan
Dinas Kesehatan provinsi Sumatera Utara
Jln Prof. H.M. Yamin No 41AA, Perintis,
Kec Medan Timur, Kota Medan, Sumatera Utara 20232

Di Tempat

Assalamu'alaikum Warahmatullahi Wabarakatuh

Dengan hormat, sehubungan mahasiswa kami akan menyelesaikan studi,untuk itu kami memohon kesediaan Bapak / Ibu untuk memberikan kesempatan pada mahasiswa kami melakukan riset di **Perusahaan / Instansi** yang Bapak / Ibu pimpin, guna untuk penyusunan skripsi yang merupakan salah satu persyaratan dalam menyelesaikan Program **Studi Strata Satu (S-1)**

Adapun Mahasiswa/i di Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara tersebut adalah:

Nama	**: Ayu Sekar Sari**
Npm	**: 2009010106**
Jurusan	**: Sistem Informasi**
Semester	**: VIII (Delapan)**
Judul	**: Analisis Dan Prediksi Penyebaran Penyakit Demam Berdarah Dalam Pendekatan Ensemble Learning Dengan Xgboost Dan Random Forest Di Kota Medan**
Email	**: sekarsariayu641@gmail.com**
Hp/Wa	**: 081375409365**

Demikianlah surat kami ini, atas perhatian dan kerjasama yang Bapak / Ibu berikan kami ucapkan terimaksih

Wassalamu 'alaikum Warahmatullahi Wabarakatuh

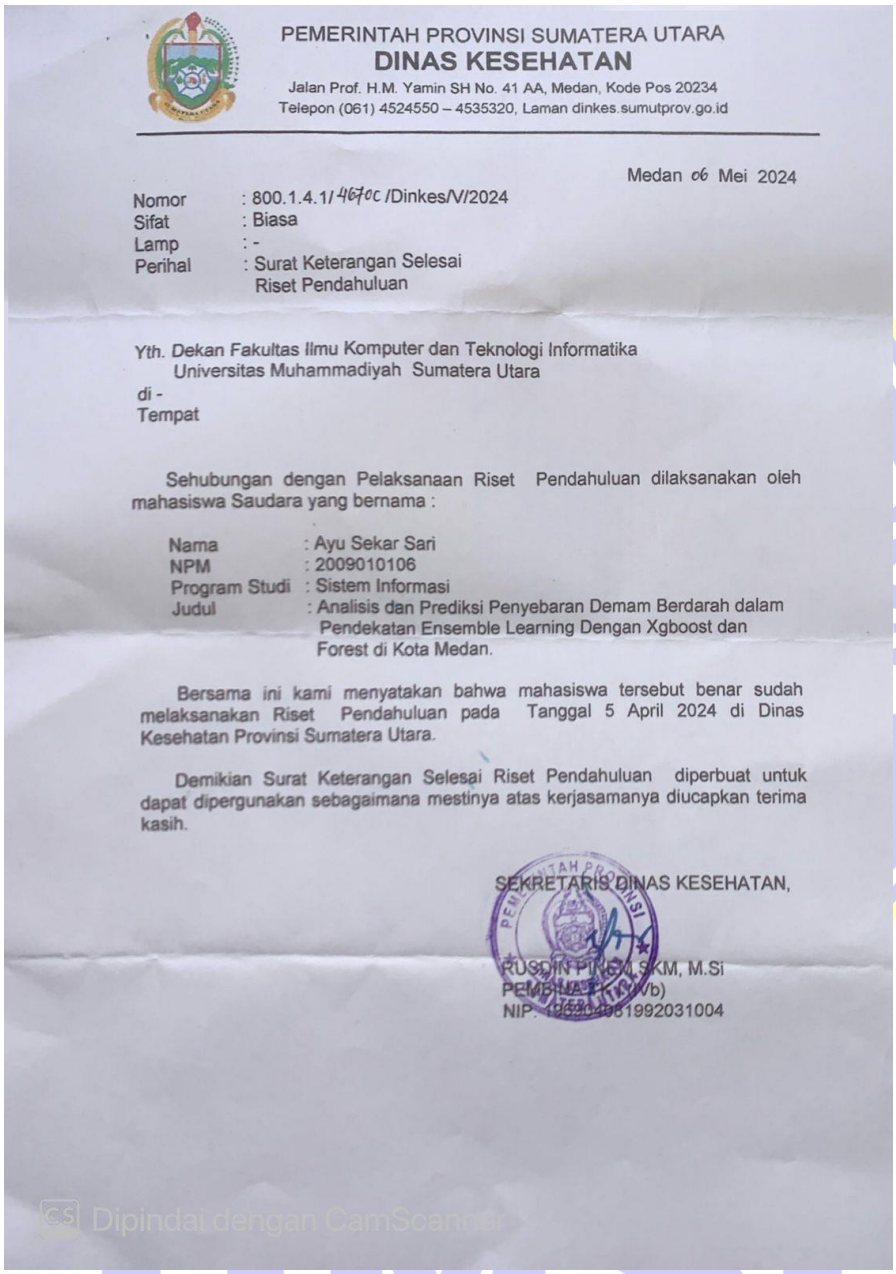
Dekan

Dr. Al-Khowarizmi,M.Kom
NIDN : 0127099201

Cc.File

BAN-PT MQA Agensi Kelayakan Malaysia Malaysian Qualifications Agency QS STARS

B. Surat Keterangan Selesai Riset Data



PEMERINTAH PROVINSI SUMATERA UTARA
DINAS KESEHATAN
Jalan Prof. H.M. Yamin SH No. 41 AA, Medan, Kode Pos 20234
Telepon (061) 4524550 – 4535320, Laman dinkes.sumutprov.go.id

Medan 06 Mei 2024

Nomor : 800.1.4.1/ 4670 /Dinkes/V/2024
Sifat : Biasa
Lamp : -
Perihal : Surat Keterangan Selesai Riset Pendahuluan

Yth. Dekan Fakultas Ilmu Komputer dan Teknologi Informatika
Universitas Muhammadiyah Sumatera Utara
di -
Tempat

Sehubungan dengan Pelaksanaan Riset Pendahuluan dilaksanakan oleh mahasiswa Saudara yang bernama :

Nama : Ayu Sekar Sari
NPM : 2009010106
Program Studi : Sistem Informasi
Judul : Analisis dan Prediksi Penyebaran Demam Berdarah dalam Pendekatan Ensemble Learning Dengan Xgboost dan Forest di Kota Medan.

Bersama ini kami menyatakan bahwa mahasiswa tersebut benar sudah melaksanakan Riset Pendahuluan pada Tanggal 5 April 2024 di Dinas Kesehatan Provinsi Sumatera Utara.

Demikian Surat Keterangan Selesai Riset Pendahuluan diperbuat untuk dapat dipergunakan sebagaimana mestinya atas kerjasamanya diucapkan terima kasih.

SEKRETARIS DINAS KESEHATAN,

RUSDIN PINEM, SKM, M.Si
PEMBINA TK I (IV/b)
NIP. [illegible]1992031004