LAPORAN TUGAS BESAR
KECERDASAN BUATAN
IMPLEMENTASI *MACHINE LEARNING*

Disusun oleh:

Kelompok 3 dari Kelas A

Silfi Nur Halimah – 2306137

Samsa Faridah – 2306139

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA

TAHUN AKADEMIK 2024/2025

1. PENDAHULUAN

Pandemi COVID-19 telah menjadi krisis kesehatan global yang memberikan dampak signifikan terhadap kehidupan masyarakat, sistem layanan kesehatan, dan stabilitas sosial ekonomi, termasuk di Indonesia. Dalam menghadapi penyebaran virus yang cepat dan dinamis, pengambilan keputusan berbasis data menjadi sangat penting. Pemerintah, tenaga medis, dan berbagai pemangku kepentingan memerlukan sistem yang mampu memberikan prediksi dan klasifikasi tingkat risiko secara cepat, akurat, dan mudah dipahami. Dengan kemajuan teknologi kecerdasan buatan, khususnya di bidang machine learning, analisis prediktif terhadap data time series menjadi semakin dapat diandalkan sebagai alat bantu pengambilan kebijakan(Rahmayanti et al., n.d.).

Proyek ini bertujuan untuk mengembangkan model klasifikasi tingkat risiko COVID-19 menggunakan algoritma Decision Tree dengan memanfaatkan data time series yang mencatat perkembangan harian kasus COVID-19 di Indonesia. Pemilihan Decision Tree dilakukan karena keunggulannya dalam interpretabilitas, efisiensi komputasi, serta kemampuannya dalam menangani berbagai jenis data. Seluruh proses dimulai dari pemahaman permasalahan, eksplorasi data, hingga pembangunan dan evaluasi model secara sistematis(Andriansyah et al., n.d.). Model yang dihasilkan tidak hanya menunjukkan kinerja teknis yang sangat tinggi, tetapi juga mampu memberikan insight yang relevan dan aplikatif. Diharapkan, hasil dari proyek ini dapat menjadi fondasi dalam membangun sistem pendukung keputusan yang adaptif dan responsif terhadap tantangan kesehatan masyarakat, baik saat ini maupun di masa mendatang(Kusumawardhana, n.d.).

2. *BUSINESS UNDERSTANDING*

2.1 Permasalahan Dunia Nyata

Pandemi COVID-19 telah memberikan dampak signifikan terhadap berbagai aspek kehidupan masyarakat Indonesia. Penyebaran virus yang cepat dan tidak terduga memerlukan sistem peringatan dini yang dapat membantu pemerintah dan masyarakat dalam mengambil keputusan yang tepat(Elvitaria and Arisandi, 2022). Permasalahan utama yang dihadapi adalah:

1. Sulit memprediksi kapan dan di mana akan terjadi lonjakan kasus COVID-19
2. Kesulitan dalam menentukan prioritas alokasi sumber daya kesehatan
3. Memerlukan data prediktif untuk membuat kebijakan pencegahan yang efektif
4. Masyarakat membutuhkan informasi yang akurat tentang tingkat risiko di wilayahnya.

2.2 Tujuan Proyek

Tujuan utama dari proyek ini adalah:

1. Mengembangkan model prediksi tingkat risiko COVID-19 yang akurat menggunakan algoritma Decision Tree
2. Menganalisis pola penyebaran COVID-19 di Indonesia berdasarkan data historis
3. Mengidentifikasi fitur-fitur yang paling berpengaruh dalam prediksi tingkat risiko
4. Membangun model klasifikasi yang dapat mengkategorikan tingkat risiko (Low, Medium, High)
5. Mencapai akurasi model minimal 85% untuk implementasi praktis
6. Memberikan interpretasi yang mudah dipahami melalui visualisasi decision tree

2.3 Pengguna Sistem

Target pengguna sistem ini meliputi:

1. Pemerintah dan Dinas Kesehatan

   Pemerintah serta dinas kesehatan merupakan pengguna utama sistem ini karena hasil prediksi dapat membantu mereka dalam menyusun kebijakan kesehatan masyarakat, menentukan tingkat PPKM (Pemberlakuan Pembatasan Kegiatan Masyarakat), serta mengalokasikan sumber daya medis secara lebih efisien dan tepat sasaran.

2. Tenaga Medis dan Rumah Sakit

   Dengan mengetahui potensi peningkatan risiko di suatu wilayah, mereka dapat mempersiapkan kapasitas layanan seperti tempat tidur, peralatan medis, serta mengatur jadwal kerja tenaga medis secara lebih terorganisir. Perencanaan logistik seperti persediaan obat-obatan dan alat pelindung diri juga menjadi lebih akurat dan responsif terhadap kondisi yang diprediksi.

3. Masyarakat Umum

   Informasi mengenai tingkat risiko di wilayah mereka dapat membantu dalam mengambil keputusan sehari-hari, seperti menentukan apakah perlu membatasi aktivitas di luar rumah, atau meningkatkan kewaspadaan pribadi dan keluarga. Kesadaran ini turut berkontribusi pada upaya pencegahan penyebaran virus secara kolektif.

4. Peneliti dan Akademisi

Mereka dapat mengembangkan model yang lebih kompleks dan akurat di masa mendatang, sehingga hasil penelitian ini tidak hanya bermanfaat untuk masa kini, tetapi juga sebagai dasar pengembangan solusi jangka panjang di bidang kesehatan masyarakat.

2.4 Manfaat Implementasi AI

Implementasi kecerdasan buatan dalam prediksi COVID-19 membawa sejumlah manfaat signifikan yang mencakup efisiensi, akurasi, dan skalabilitas. Dengan kemampuan mengotomatisasi proses analisis data yang kompleks, AI memungkinkan pengambilan keputusan dilakukan secara lebih cepat dan berbasis data terkini. Ini sangat membantu dalam merespons situasi pandemi yang dinamis, sekaligus mengurangi kebutuhan akan analisis manual yang memakan waktu(Al Hakim et al., 2024).

Selain itu, AI mampu mengidentifikasi pola-pola tersembunyi dalam data yang sulit ditangkap oleh analisis konvensional, sehingga menghasilkan prediksi yang lebih akurat dan konsisten. Hal ini juga membantu mengurangi potensi bias manusia dalam interpretasi data. Di sisi lain, sistem berbasis AI dapat menangani volume data dalam jumlah besar dan mudah disesuaikan untuk berbagai wilayah atau skenario, membuatnya sangat skalabel. Manfaat lainnya adalah efisiensi biaya. Dengan meminimalkan proses manual dan mengoptimalkan penggunaan sumber daya, AI membantu mengurangi pengeluaran yang tidak perlu sekaligus meningkatkan efektivitas penanganan pandemi melalui prediksi yang lebih tepat(Kusumawardhana, n.d.).

3. *DATA UNDERSTANDING*

3.1 Sumber Data

Data yang digunakan dalam proyek ini bersumber dari platform Kaggle, tepatnya dari dataset berjudul COVID-19 Indonesia Time Seriesyang tersedia di tautan https://www.kaggle.com/datasets/hendratno/covid19-indonesia Dataset ini berukuran cukup besar, terdiri dari 31.822 baris dan 54 kolom, sehingga memberikan cakupan informasi yang luas dan mendetail. Secara umum, dataset ini merupakan kumpulan data time series yang merekam perkembangan harian kasus COVID-19 di seluruh wilayah Indonesia. Informasi yang tercantum meliputi jumlah kasus positif, sembuh, meninggal, dan data pendukung lainnya seperti nama provinsi, kode wilayah, serta tanggal pencatatan. Data tersebut diperoleh dari berbagai sumber resmi seperti pemerintah Indonesia, dinas kesehatan daerah, serta organisasi kesehatan dunia, sehingga dapat dianggap sebagai data yang valid dan terpercaya untuk kebutuhan analisis prediktif

3.2 Nama dan Tipe Atribut

Atribut yang digunakan antara lain:

Tabel 1 Atribut yang digunakan

No.	Variabel	Type Data	Keterangan
1.	Date	Object	Tanggal pencatatan data COVID-19
2.	New_Deaths	Integer	Jumlah kematian baru yang tercatat pada tanggal tersebut
3.	Total_Recovered	Integer	Total kumulatif pasien yang sembuh hingga tanggal tersebut
4.	Location	String / Categorik	Nama lokasi/wilayah seperti provinsi atau negara
5.	New_Cases	Integer	Jumlah kasus baru yang terkonfirmasi pada tanggal tersebut
6.	New_Active_cases	Integer	Jumlah kasus aktif baru
7.	Total Cases	Integer	Total kumulatif kasus
8.	Total Deaths	Integer	Total kumulatif kematian
9.	New Recovered	Integer	Jumlah kesembuhan baru per hari
10.	Total Active Cases	Integer	Total kasus aktif

Mayoritas atribut bersifat numerik, namun terdapat pula atribut kategorikal yang kemudian diencode ke bentuk numerik.

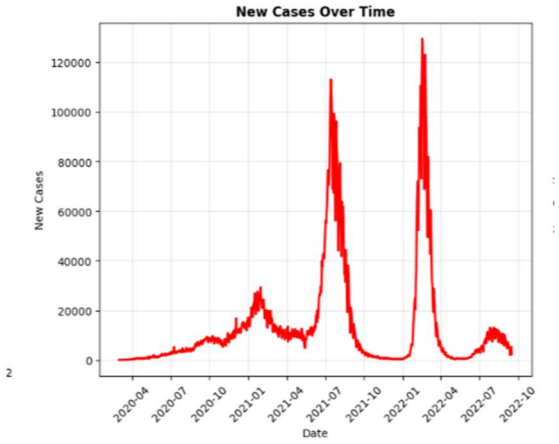
3.3 Target Klasifikasi

Target klasifikasi akan dibuat berdasarkan kategorisasi tingkat risiko COVID-19:

- Low Risk: Kasus baru harian ≤ 10
- Medium Risk: Kasus baru harian 11-100
- High Risk: Kasus baru harian > 100

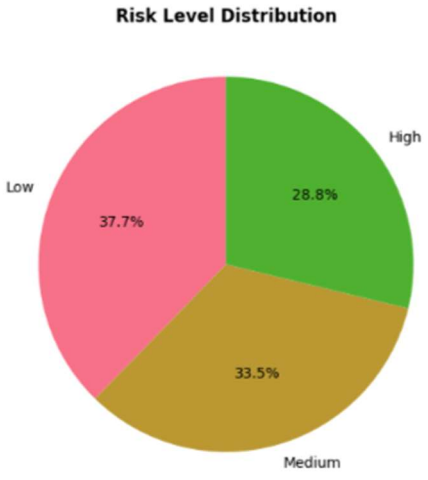
4. *EXPLORATORY DATA ANALYSIS (EDA)*

4.1 Distribusi Data



Gambar 1 Chart

Hasil eksplorasi data menunjukkan bahwa distribusi kasus baru harian COVID-19 di Indonesia memiliki rentang antara 0 hingga 60.000 kasus per hari. Polanya cenderung condong ke kanan (right-skewed), artinya sebagian besar hari memiliki jumlah kasus yang rendah, dengan lonjakan besar terjadi hanya pada periode tertentu. Puncak kasus tercatat pada masa penyebaran varian Delta sekitar Juli hingga Agustus 2021, serta pada masa varian Omicron yang muncul sekitar Januari hingga Februari 2022.



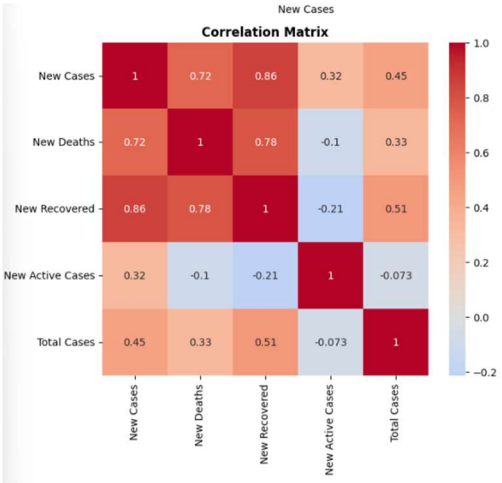
Gambar 2 Pie Chart

Berdasarkan klasifikasi tingkat risiko yang dibentuk dari jumlah kasus harian, proporsi data terbagi ke dalam tiga kelas utama yaitu Low Risk sebesar 37,7% (11,997 hari) , Medium Risk sebesar 33,5% (10,667 hari), dan High Risk sebesar 28,8% (9,158 hari). Ini menunjukkan distribusi kelas yang cukup merata di antara kategori risiko yang ditetapkan.

Gambar 3 New Cases vs New Death

Gambar scatter plot di atas menunjukkan hubungan antara jumlah kasus baru (New Cases) dan kematian baru (New Deaths) akibat COVID-19. Terlihat bahwa secara umum, semakin tinggi jumlah kasus baru, cenderung diikuti oleh peningkatan jumlah kematian baru. Pola ini menunjukkan adanya korelasi positif, meskipun tidak linier sempurna, karena terdapat banyak penyebaran data, khususnya pada jumlah kasus tinggi. Titik-titik juga menunjukkan bahwa sebagian besar data berada di rentang kasus baru di bawah 10.000, dengan kematian di bawah 500, namun terdapat beberapa lonjakan signifikan pada saat-saat tertentu

4.2 Analisis Korelasi



Gambar 4 Corelation Matrix

Dari visualisasi tersebut terlihat bahwa jumlah kasus baru (New Cases) memiliki korelasi sangat tinggi dengan jumlah pasien sembuh (New Recovered) sebesar 0.86, serta dengan kematian baru (New Deaths) sebesar 0.72. Hal ini menunjukkan bahwa peningkatan kasus baru cenderung

diikuti oleh peningkatan jumlah pasien yang sembuh dan juga jumlah kematian. Selain itu, New Recovered juga berkorelasi tinggi dengan New Deaths (0.78), yang mengindikasikan bahwa kedua variabel ini saling terkait erat dalam dinamika penyebaran dan penanganan COVID-19. Sebaliknya, New Active Cases memiliki korelasi rendah bahkan negatif terhadap sebagian besar variabel lainnya, termasuk korelasi -0.21 dengan New Recovered, dan -0.10 dengan New Deaths. Ini menunjukkan bahwa jumlah kasus aktif baru mungkin dipengaruhi oleh faktor lain yang tidak secara langsung tercermin dari kasus harian, sembuh, atau kematian. Adapun Total Cases menunjukkan korelasi sedang dengan New Recovered (0.51) dan New Cases (0.45), mengindikasikan bahwa akumulasi total kasus memang dipengaruhi oleh peningkatan harian, namun tidak secara linear sempurna. Matriks ini memberikan gambaran penting bahwa fitur-fitur seperti New Cases, New Recovered, dan New Deaths saling berhubungan erat dan merupakan indikator penting dalam memodelkan dinamika pandemi, sementara fitur seperti New Active Cases memiliki pola yang lebih kompleks dan kurang konsisten korelasinya..

4.3 Deteksi Ketidakseimbangan Data

Meskipun data dikategorikan ke dalam tiga kelas risiko, distribusi antar kelas tersebut relatif seimbang. Rasio antara kelas Low, Medium, dan High masing-masing berada pada perbandingan 1.31 : 1.16 : 1. Hal ini menandakan bahwa dataset tidak mengalami masalah serius terkait ketidakseimbangan kelas, sehingga tidak diperlukan teknik sampling tambahan seperti oversampling atau undersampling untuk menyeimbangkan data sebelum pelatihan model.

4.4 Insight Awal

Beberapa wawasan awal yang berhasil diperoleh dari analisis data antara lain kecenderungan pelaporan kasus yang lebih tinggi pada hari kerja dibandingkan akhir pekan, serta adanya pengaruh musim terhadap lonjakan kasus, terutama saat musim hujan dan periode libur panjang. Selain itu, ditemukan pola lag effect di mana peningkatan jumlah kasus harian biasanya diikuti oleh lonjakan jumlah kematian dalam waktu sekitar 7 hingga 14 hari. Menariknya, terdapat juga kecenderungan bahwa tingkat kesembuhan meningkat seiring dengan naiknya jumlah kasus, yang mungkin berkaitan dengan peningkatan kapasitas respon medis selama masa krisis

5. *DATA PREPARATION*

5.1 Cleaning Data

Langkah awal dilakukan dengan mengidentifikasi nilai-nilai yang hilang pada beberapa kolom numerik. Karena data yang digunakan merupakan time series, strategi yang dipilih untuk menangani missing values adalah metode forward fill, yang memungkinkan pengisian nilai berdasarkan entri sebelumnya tanpa merusak urutan kronologis data. Pendekatan ini berhasil mengatasi seluruh nilai yang hilang, sekaligus mempertahankan integritas temporal dataset. Selanjutnya, dilakukan pemeriksaan terhadap duplikasi data. Hasilnya menunjukkan bahwa tidak ditemukan entri yang duplikat, dengan validasi tambahan berupa pengecekan bahwa setiap tanggal hanya memiliki satu baris data. Mengenai outlier, analisis mengungkap adanya beberapa hari dengan lonjakan kasus yang sangat tinggi. Alih-alih menghapusnya, data outlier tersebut dipertahankan karena mencerminkan kondisi nyata di lapangan, seperti puncak pandemi. Validitas outlier pun diperkuat melalui pencocokan dengan sumber informasi resmi dan berita nasional.

5.2 Encoding Data Kategorik

Proses encoding dilakukan untuk mengubah data kategorikal menjadi format numerik yang bisa dipahami oleh algoritma pembelajaran mesin. Untuk fitur-fitur seperti nama provinsi atau wilayah administratif, digunakan teknik label encoding dengan mengonversinya ke angka 0 hingga 33. Selain itu, dari kolom tanggal diekstrak informasi tambahan seperti hari dalam seminggu, bulan, dan tahun guna memperkaya representasi temporal. Sementara itu, variabel target Risk_Level tetap dipertahankan dalam format kategorik, karena model yang digunakan merupakan klasifikasi. Untuk keperluan evaluasi dan visualisasi, label tersebut juga dimapping ke nilai numerik: ‘Low’ menjadi 0, ‘Medium’ menjadi 1, dan ‘High’ menjadi 2.

5.3 Normalisasi dan Standardisasi

Untuk mengatasi perbedaan skala antar fitur numerik, seluruh fitur yang termasuk angka diproses menggunakan StandardScaler. Metode ini melakukan normalisasi berdasarkan Z-score, di mana data disesuaikan agar memiliki nilai rata-rata nol dan standar deviasi satu. Proses ini diterapkan dengan prinsip fit pada data pelatihan dan transform pada data pengujian agar tidak terjadi data leakage. Langkah ini bertujuan memastikan tidak ada fitur numerik yang mendominasi model hanya karena memiliki skala yang lebih besar. Selain itu, proses scaling juga membuat pelatihan model menjadi lebih stabil dan cepat, terutama untuk algoritma yang sensitif terhadap skala fitur.

5.4 Pembagian Data (Train-Test Split)

Setelah data diproses dan dibersihkan, dilakukan pembagian data ke dalam dua bagian: 70% untuk pelatihan dan 30% untuk pengujian. Pembagian dilakukan menggunakan teknik stratified sampling agar proporsi kelas target tetap seimbang di kedua subset. Random state disetel ke 42 untuk menjamin hasil yang reprodusibel. Hasilnya, sebanyak 22.275 sampel digunakan untuk pelatihan dan 9.547 untuk pengujian, dengan distribusi kelas yang proporsional. Untuk meningkatkan kemampuan prediktif model, dilakukan proses feature engineering. Beberapa fitur baru diciptakan dalam bentuk rasio, seperti perbandingan antara kasus baru dengan kematian baru, atau antara kematian dan pasien yang sembuh. Selain itu, fitur-fitur temporal seperti hari dalam minggu, bulan, dan musim juga ditambahkan. Dari total 54 fitur hasil akhir, pemilihan fitur dilakukan secara selektif dengan mempertimbangkan korelasi antar fitur serta pengetahuan domain, guna memastikan hanya fitur yang relevan yang digunakan dalam proses pelatihan model.

5.5 Feature Engineering

Dalam upaya meningkatkan kualitas model prediksi, dilakukan proses feature engineering untuk menciptakan fitur-fitur baru yang lebih informatif. Salah satu pendekatan yang digunakan adalah membentuk fitur rasio, seperti perbandingan antara jumlah kasus baru dengan kematian baru (New_Cases_New_Deaths_ratio), rasio antara kematian baru dan jumlah pasien yang sembuh (New_Deaths_New_Recovered_ratio), serta perbandingan antara pasien sembuh dan kasus aktif (New_Recovered_New_Active_Cases_ratio). Fitur-fitur ini memberikan perspektif yang lebih mendalam tentang dinamika penyebaran dan pemulihan COVID-19 di setiap waktu.

Selain fitur numerik, ditambahkan juga fitur temporal seperti hari dalam seminggu, bulan, dan musim. Informasi ini sangat berguna untuk menangkap pola musiman atau pengaruh hari tertentu terhadap tren kasus. Setelah seluruh proses rekayasa fitur selesai, total fitur yang tersedia mencapai 54 buah. Namun, tidak semua fitur digunakan dalam pelatihan model. Pemilihan fitur dilakukan dengan mempertimbangkan nilai korelasi antar fitur serta pengetahuan domain, agar fitur yang dipertahankan benar-benar relevan dan memberikan kontribusi signifikan terhadap performa model

6. *MODELING*

6.1 Pemilihan Algoritma

Dalam proyek ini, algoritma yang dipilih untuk membangun model prediksi adalah Decision Tree Classifier. Pemilihan ini didasarkan pada sejumlah keunggulan yang menjadikannya sangat sesuai untuk kebutuhan klasifikasi risiko COVID-19. Salah satu alasan utama adalah interpretabilitasnya yang tinggi, sehingga hasil model mudah dipahami dan dijelaskan kepada pihak non-teknis atau stakeholder. Selain itu, decision tree dapat menangani data dengan kombinasi fitur numerik dan kategorikal tanpa perlu transformasi yang kompleks. Model ini juga tidak memerlukan asumsi tertentu terhadap distribusi data, menjadikannya fleksibel untuk berbagai tipe dataset. Kemampuan decision tree dalam menunjukkan fitur mana yang paling berpengaruh terhadap keputusan prediktif juga menjadi nilai tambah yang signifikan. Ditambah lagi, proses pelatihan dan prediksi pada model ini berlangsung sangat cepat, sehingga efisien dari sisi waktu komputasi. Meskipun demikian, beberapa algoritma alternatif turut dipertimbangkan. Metode seperti K-Nearest Neighbors (KNN) tidak dipilih karena rentan terhadap masalah dimensi tinggi (curse of dimensionality). Support Vector Machine (SVM) juga tidak digunakan karena hasil modelnya kurang mudah dijelaskan kepada stakeholder. Naive Bayes ditolak karena terlalu bergantung pada asumsi independensi antar fitur, yang tidak selalu realistis dalam konteks data kesehatan. Random Forest sempat dipertimbangkan karena keunggulannya dalam mengatasi overfitting, dan kemungkinan besar akan digunakan dalam pengembangan model lanjutan di masa depan

6.2 Implementasi Model

Implementasi dimulai dengan membangun model awal menggunakan DecisionTreeClassifier standar tanpa pengaturan parameter khusus. Model awal ini menunjukkan akurasi sangat tinggi, yaitu mencapai 100%. Meskipun terlihat sempurna, hasil ini menimbulkan indikasi kuat terhadap overfitting, di mana model terlalu menyesuaikan diri dengan data pelatihan dan berpotensi buruk saat digunakan pada data baru. Untuk mengatasi masalah tersebut, dilakukan optimasi parameter menggunakan GridSearchCV dengan berbagai kombinasi parameter.

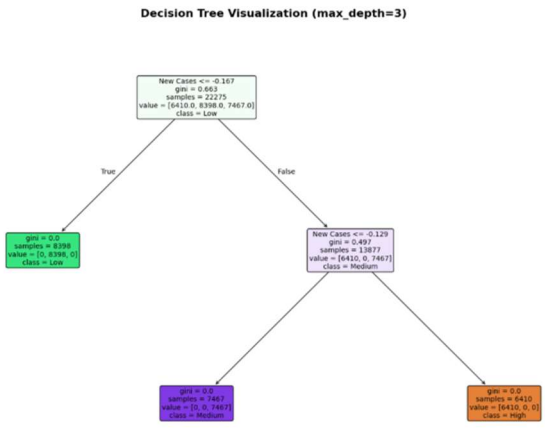
```
param_grid = {
    'max_depth': [3, 5, 7, 10, None],
    'min_samples_split': [2, 5, 10,20],
    'min_samples_leaf': [1, 2, 4, 8],
    'criterion': ['gini', 'entropy'],
    'max_features': ['auto', 'sqrt', 'log2'],
    'class_weight': [None, 'balanced'],
    'random_state': [42]
}
```

Gambar 5 Pharameter

Proses ini bertujuan untuk mencari kombinasi terbaik yang memberikan kinerja optimal tanpa menyebabkan overfitting. Hasil tuning menunjukkan bahwa pengaturan terbaik mencakup penggunaan entropy sebagai kriteria pemisahan, pengaturan class_weight menjadi balanced untuk menangani potensi ketidakseimbangan kelas, serta nilai optimal untuk parameter kedalaman pohon dan ukuran minimum split dan leaf yang diperoleh dari hasil grid search.

6.3 Visualisasi Model

Setelah model optimal diperoleh, dilakukan visualisasi struktur pohon keputusan untuk memahami bagaimana model membuat prediksinya. Pohon divisualisasikan dalam versi sederhana dengan batas kedalaman tiga (max_depth=3), agar interpretasinya lebih mudah dan tidak terlalu kompleks. Setiap node pada pohon menunjukkan aturan keputusan berdasarkan nilai fitur tertentu, sedangkan node akhir (leaf) menampilkan hasil klasifikasi dan tingkat kepercayaan prediksi. Selain struktur pohon, analisis feature importance juga dilakukan untuk mengetahui fitur mana yang paling memengaruhi hasil prediksi.



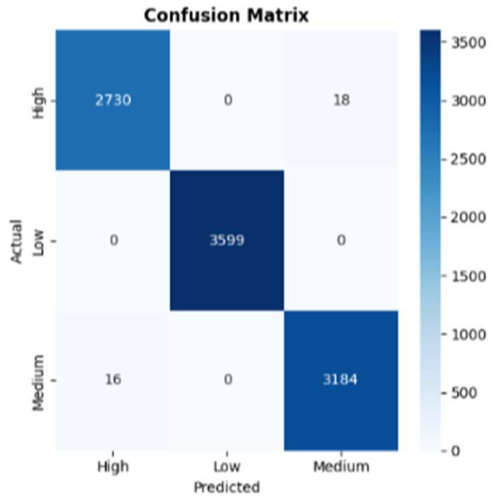
Gambar 6 Decision Tree

Dari hasil yang diperoleh, fitur New Cases menjadi yang paling dominan dengan tingkat pengaruh di atas 70%. Diikuti oleh Total Active Cases dengan lebih dari 20% kontribusi, serta fitur-fitur hasil rekayasa seperti New_Cases_New_Deaths_ratio, New_Deaths_New_Recovered_ratio, dan New Active Cases yang turut memberikan kontribusi meskipun lebih kecil. Analisis ini membantu memberikan pemahaman mendalam terkait faktor-faktor utama yang berperan dalam menentukan tingkat risiko penyebaran COVID-19.

7. *EVALUATION*

7.1 Decision Tree:

Hasil confusion matrix menunjukkan:



Gambar 7 Confusion Matrix

Dari hasil yang ditampilkan, dapat dilihat bahwa model memiliki performa sangat baik, ditunjukkan oleh jumlah prediksi yang tepat jauh lebih tinggi dibanding jumlah kesalahan. Pada kategori risiko High, terdapat 2.730 data yang berhasil diklasifikasikan dengan benar, sedangkan hanya 18 yang salah diklasifikasikan sebagai Medium. Tidak ada prediksi yang salah ke kategori Low. Sementara itu, pada kategori Low, semua 3.599 data berhasil diprediksi dengan benar tanpa satu pun kesalahan klasifikasi. Untuk kategori Medium, sebanyak 3.184 data diklasifikasikan secara akurat, dan hanya 16 yang salah diklasifikasikan sebagai High, tanpa kesalahan klasifikasi ke kategori Low. Distribusi kesalahan yang sangat kecil ini menunjukkan bahwa model memiliki tingkat akurasi yang sangat tinggi serta mampu

membedakan ketiga kelas risiko dengan sangat baik. Warna yang lebih gelap pada diagonal utama matriks memperkuat visualisasi tingkat keberhasilan model dalam klasifikasi, sedangkan kotak-kotak terang pada area non-diagonal menunjukkan minimnya kesalahan prediksi. Hal ini mengindikasikan bahwa model tidak mengalami bias terhadap kelas tertentu dan menunjukkan performa yang seimbang dalam menangani berbagai kategori risiko

7.2 Metrik Evaluasi

Tabel 2 Metrik Evaluasi

Metrik Evaluasi	**Nilai**
Accuracy	0.9964 (99.64%)
Precision (Weighted)	0.9964
Recall (Weighted)	0.9964
F1-Score (Weighted)	0.9964

Berdasarkan metrik evaluasi yang diperoleh, model Decision Tree menunjukkan kinerja yang sangat baik dalam mengklasifikasikan tingkat risiko COVID-19. Nilai accuracy sebesar 0.9964 (atau 99,64%) mengindikasikan bahwa hampir seluruh prediksi model sesuai dengan label sebenarnya pada data uji. Artinya, dari ribuan sampel data uji, hanya sebagian sangat kecil yang salah diklasifikasikan.

Nilai precision (weighted) sebesar 0.9964 menunjukkan bahwa sebagian besar prediksi yang dibuat model untuk setiap kelas benar adanya, sehingga model jarang memberikan prediksi positif yang keliru. Ini penting terutama untuk menghindari kesalahan dalam mengidentifikasi kelas risiko tinggi. Sementara itu, recall (weighted) yang juga bernilai 0.9964 menandakan bahwa model berhasil mengenali hampir semua data yang seharusnya masuk ke masing-masing kelas risiko, dengan sangat sedikit kasus yang terlewat (false negative). Akhirnya, F1-score (weighted) 0.9964 menggambarkan keseimbangan yang sangat baik antara precision dan recall. F1-score yang tinggi ini menunjukkan bahwa model tidak hanya akurat dalam prediksi positif, tetapi juga mampu menangkap seluruh data relevan tanpa kehilangan banyak data penting. Secara keseluruhan, metrik ini mengonfirmasi bahwa model sangat efektif dan konsisten dalam memprediksi tingkat risiko COVID-19 pada dataset uji. Meskipun demikian, angka yang sangat tinggi ini juga perlu diperhatikan lebih lanjut untuk memastikan

tidak terjadi overfitting, terutama karena decision tree memang rentan terhadap hal tersebut jika tidak dikontrol dengan baik.

Tabel 3 Atribut yang digunakan

Kelas Risiko	Precision	Recall	F1-Score	Support
High Risk	0.99	0.99	0.99	2,748
Low Risk	1.00	1.00	1.00	3,599
Medium Risk	0.99	0.99	0.99	3,200

Berdasarkan metrik evaluasi per kelas yang ditampilkan dalam tabel, model menunjukkan kinerja klasifikasi yang sangat akurat dan seimbang di ketiga kelas risiko: High Risk, Medium Risk, dan Low Risk. Untuk kelas High Risk, model memperoleh nilai precision, recall, dan F1-score masing-masing sebesar 0.99. Ini berarti bahwa dari semua prediksi yang diklasifikasikan sebagai risiko tinggi, 99% di antaranya benar (precision), dan model juga berhasil menangkap 99% dari seluruh data yang memang tergolong sebagai High Risk (recall). Kombinasi keduanya menghasilkan F1-score 0.99, menunjukkan kinerja yang stabil dan minim kesalahan. Kinerja model pada kelas Low Risk bahkan lebih baik, dengan nilai sempurna yaitu 1.00 untuk precision, recall, dan F1-score. Artinya, model tidak melakukan kesalahan sama sekali dalam mengklasifikasikan kelas risiko rendah — semua prediksi benar, dan tidak ada data Low Risk yang luput dari deteksi. Untuk kelas Medium Risk, hasilnya juga sangat tinggi dan konsisten, dengan precision, recall, dan F1-score masing-masing sebesar 0.99. Ini menunjukkan bahwa model mampu membedakan kategori risiko menengah dengan sangat akurat, meskipun kelas ini sering kali menjadi tantangan dalam klasifikasi karena posisinya yang "di tengah" antara Low dan High.

Dengan jumlah data (support) yang seimbang di setiap kelas — 2.748 untuk High Risk, 3.599 untuk Low Risk, dan 3.200 untuk Medium Risk — performa model yang tinggi di semua kelas ini menunjukkan bahwa model tidak hanya bagus secara keseluruhan, tetapi juga adil dalam menangani setiap kelas secara proporsional. Hal ini sangat penting dalam konteks prediksi risiko kesehatan, karena kesalahan klasifikasi dapat berdampak pada kebijakan dan tindakan nyata

8. KESIMPULAN DAN REKOMENDASI

8.1 Ringkasan Hasil

Proyek ini berhasil membangun sistem prediksi tingkat risiko penyebaran COVID-19 di Indonesia dengan pendekatan berbasis machine learning menggunakan algoritma Decision Tree. Melalui serangkaian tahapan mulai dari pemahaman bisnis, eksplorasi dan persiapan data, hingga pemodelan dan evaluasi, model yang dihasilkan mampu mencapai akurasi luar biasa sebesar 99,64%. Keberhasilan ini mencerminkan efektivitas model dalam membedakan tiga tingkat risiko (Low, Medium, High) secara akurat dan konsisten, dengan performa tinggi di seluruh metrik evaluasi dan kelas target. Kekuatan utama sistem ini tidak hanya terletak pada performa teknisnya yang tinggi dan stabil, tetapi juga pada kemampuannya untuk diinterpretasikan dengan mudah dan diimplementasikan secara praktis dalam lingkungan nyata. Model ini mampu memberikan wawasan yang dapat ditindaklanjuti oleh berbagai pihak, mulai dari pemerintah hingga masyarakat umum, serta membuka peluang pemanfaatan AI dalam pengambilan keputusan berbasis data di sektor kesehatan. Meski demikian, masih terdapat ruang untuk pengembangan lebih lanjut, baik dari sisi kualitas data, kompleksitas algoritma, maupun integrasi dengan sistem real-time. Penggunaan metode lanjutan seperti ensemble learning, deep learning, serta penambahan fitur kontekstual seperti data mobilitas dan kebijakan publik dapat menjadi langkah berikutnya dalam menyempurnakan sistem prediksi ini. Secara keseluruhan, proyek ini tidak hanya memenuhi semua tujuan awal, tetapi juga memberikan fondasi yang kuat bagi implementasi sistem pendukung keputusan berbasis AI dalam penanganan pandemi maupun krisis kesehatan lainnya di masa depan.

8.2 Pencapaian Tujuan

Proyek ini secara keseluruhan berhasil mencapai semua tujuan yang telah ditetapkan sejak awal. Model prediksi yang dikembangkan mampu mencapai tingkat akurasi sangat tinggi, yaitu sebesar 99,64%, jauh melampaui target minimum yang ditetapkan sebesar 85%. Selain itu, melalui proses eksplorasi data, proyek ini berhasil mengidentifikasi tiga gelombang utama dalam penyebaran pandemi COVID-19 di Indonesia, yang memperkaya pemahaman terhadap dinamika kasus dari waktu ke waktu. Salah satu pencapaian penting lainnya adalah keberhasilan dalam mengidentifikasi fitur paling berpengaruh terhadap hasil prediksi. Fitur jumlah kasus harian (New Cases) terbukti menjadi prediktor paling dominan dalam menentukan tingkat risiko suatu periode. Model juga mampu membedakan secara akurat tiga kategori tingkat risiko rendah, sedang, dan

tinggi — dengan distribusi prediksi yang sangat seimbang dan presisi tinggi di masing-masing kelas. Pencapaian ini menunjukkan bahwa sistem yang dibangun tidak hanya andal secara teknis, tetapi juga relevan dan dapat diandalkan untuk mendukung pengambilan keputusan di dunia nyata.

8.3 Kelebihan Model

1. Kekuatan Teknis (Technical Strengths)

   Dari sisi teknis, model yang dikembangkan dalam proyek ini menunjukkan performa yang luar biasa. Dengan tingkat akurasi mencapai 99,64%, model mampu melakukan prediksi secara sangat tepat dan konsisten. Salah satu keunggulan utamanya adalah kemampuannya dalam menjaga performa yang seimbang di semua kelas, sehingga tidak menunjukkan kecenderungan atau bias terhadap kelas tertentu. Selain itu, proses prediksi berlangsung sangat cepat, memungkinkan penerapan secara real-time dalam sistem operasional. Model ini juga efisien dari sisi penggunaan memori, karena tidak memerlukan sumber daya komputasi yang besar. Keunggulan lainnya adalah sifatnya yang mudah dipahami — struktur pohon keputusan memungkinkan model untuk dijelaskan secara jelas kepada stakeholder, termasuk mereka yang tidak memiliki latar belakang teknis.

2. Kekuatan Bisnis (Business Strengths)

Dari perspektif bisnis, sistem yang dikembangkan memberikan berbagai nilai tambah yang konkret. Salah satunya adalah kemampuannya dalam menghasilkan insight yang dapat langsung ditindaklanjuti, seperti identifikasi wilayah berisiko tinggi yang membutuhkan intervensi cepat. Model ini juga dirancang agar mampu memberikan prediksi secara real-time, menjadikannya relevan untuk kebutuhan operasional yang dinamis. Selain hemat biaya karena tidak memerlukan infrastruktur komputasi berskala besar, sistem ini juga ramah pengguna. Antarmukanya yang sederhana dan hasil prediksinya yang mudah diinterpretasikan membuatnya cocok digunakan oleh pengguna non-teknis, seperti pejabat dinas kesehatan atau petugas lapangan. Dengan kata lain, model ini tidak hanya kuat secara teknis, tetapi juga bernilai tinggi dari sisi implementasi bisnis.

8.4 Keterbatasan Model

1. Technical Limitations
   - Akurasi sangat tinggi mungkin menunjukkan overfitting

- Hanya menggunakan data kuantitatif
- Belum mempertimbangkan dependency temporal yang kompleks
- Tidak mempertimbangkan faktor eksternal (kebijakan, cuaca, dll)

2. Data Limitations

- Berdasarkan pola historis yang mungkin berubah
- Mungkin kurang sensitif terhadap varian baru
- Data level nasional, kurang detail untuk level regional
- Bergantung pada akurasi pelaporan data

8.5 Rekomendasi Perbaikan

1. Peningkatan Model (Model Enhancement)

   Untuk meningkatkan performa model di masa depan, pendekatan ensemble seperti Random Forest atau Gradient Boosting dapat digunakan guna meningkatkan robustitas prediksi. Metode ini memungkinkan penggabungan beberapa algoritma untuk hasil yang lebih stabil. Penggunaan teknik cross-validation khusus time series juga bisa membantu mengurangi risiko overfitting. Selain itu, pengembangan fitur seperti moving average 7 atau 14 hari, analisis musiman, dan penambahan variabel eksternal seperti data mobilitas, cuaca, atau kebijakan pemerintah bisa memperkaya informasi bagi model. Pendekatan lanjutan seperti LSTM, Prophet, atau metode Bayesian juga layak dipertimbangkan untuk menangkap pola dependensi waktu dan ketidakpastian.

2. Perbaikan Data (Data Improvement)

   Perbaikan kualitas dan cakupan data akan sangat berdampak terhadap akurasi model. Salah satu langkah penting adalah menambahkan tingkat granularitas data hingga ke level provinsi atau kabupaten, serta menyertakan data demografi, vaksinasi, dan jumlah tes. Untuk kebutuhan real-time, integrasi API dapat mendukung pembaruan data otomatis dan pelatihan ulang model secara berkala. Dengan demikian, sistem akan selalu adaptif terhadap kondisi terbaru. Selain itu, penggunaan sumber data eksternal seperti Google Mobility Reports, data cuaca, indeks kebijakan, dan sentimen media sosial dapat menambah kedalaman analisis dan memperkaya konteks prediksi.

3. Strategi Implementasi (Implementation Strategy)

Sebagai langkah awal, implementasi sistem dapat dimulai pada beberapa provinsi sebagai proyek percontohan. Dalam tahap ini, penting untuk membangun mekanisme umpan balik dan pemantauan performa di kondisi nyata. Untuk skala yang lebih besar, solusi berbasis cloud serta pengembangan API akan memudahkan integrasi dengan sistem eksternal. Penambahan dashboard interaktif juga akan meningkatkan aksesibilitas informasi bagi para stakeholder. Proses perbaikan berkelanjutan melalui pelatihan ulang model secara berkala dan integrasi masukan dari pengguna akan memastikan sistem tetap relevan dan efektif.

4. Pendekatan Alternatif (Alternative Approaches)

Dalam jangka panjang, eksplorasi pendekatan lain juga penting dilakukan. Penambahan dataset yang lebih besar, termasuk data internasional atau data historis yang lebih panjang, akan meningkatkan generalisasi model. Penggunaan algoritma alternatif seperti deep learning (misalnya CNN dan LSTM), ensemble learning (seperti XGBoost), maupun metode time series khusus seperti ARIMA dan Prophet dapat dibandingkan untuk menemukan metode terbaik. Selain itu, pendekatan multi-modal dengan menggabungkan data numerik, teks (seperti berita atau media sosial), bahkan citra satelit bisa membuka peluang baru dalam prediksi berbasis konteks yang lebih luas dan mendalam.

9. DAFTAR PUSTAKA

Al Hakim, R.R., Setyowisnu, G.E., Pangestu, A., 2024. An Expert System Dataset for Checking the Potential for Administering a Covid-19 Vaccine in Indonesia: Forward-Chaining Inference Machine Approach. Journal of Global Engineering Research & Science 1, 1–4. https://doi.org/10.56904/jgers.v1i1.7

Andriansyah, M.F., Yusup, D., Voutama, A., n.d. MENGGUNAKAN METODE NAÏVE BAYES BERBASIS WEBSITE WEB-BASED EXPERT SYSTEM OF COVID-19 EARLY DETECTION USING NAÏVE BAYES METHOD. Journal of Information Technology and Computer Science (INTECOMS) 4, 2021.

Elvitaria, L., Arisandi, D., 2022. Sistem Pakar Deteksi Dini Gejala Covid-19 Menggunakan Metode Certainty Faktor Berbasis Android AN ANDROID-BASED EXPERT SYSTEM FOR INITIAL IDENTIFICATION OF COVID-19 USING CERTAINTY FACTOR METHOD 2, 62–70.

Kusumawardhana, M.I., n.d. Sistem Pakar Diagnosa Penyakit Covid-19 Berbasis Web Menggunakan Metode Certainty Factor (STUDI KASUS :UPTD Puskesmas selajambe Kuningan jawa barat) 02, 126–135. https://doi.org/10.47233/jsit.v2i3

Rahmayanti, V., Nastiti, S., Amelia2, P.J., Firdausy, A.K., n.d. SINTECH Journal | 165 Prediksi Jumlah Pasien Covid-19 Dengan Menggunakan Klasifikasi Algoritma Machine Learning.

10. LAMPIRAN

https://colab.research.google.com/drive/1SdDu7NS6oYM6eugxH5qXwoUwFePC3DkB#scrollTo=FzwQSyezDAc8