ANALISIS KOMPARATIF METODE *DATA MINING* MULTISEKTOR PADA *DATASET* COVID-19, SAHAM BEI, DAN PERUSAHAAN GLOBAL

Satria Nur Fajriansyah[1)], Harsya Rafif Pramadhan[2)], Alaudin Safa[3)], Dandy Nurfajriansyah[4)], Muhammad Subaktiar Wijaya[5)], Yuda Samudra[6)]
Prodi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang

Correspondence author: Y.Samudra, dosen02623@unpam.ac.id, Tangerang Selatan, Indonesia

Abstract

The rapid advancement of information technology has significantly driven the adoption of *data mining* techniques across various sectors, primarily to uncover hidden patterns and support data-driven decision-making processes. This study aims to analyse and compare the effectiveness of several *data mining* methods—namely *K-Means Clustering*, *Decision tree*, *Principal component analysis (PCA)*, and *Bootstrapping*—in processing *dataset*s from three distinct domains: public health (COVID-19), finance (Indonesia Stock Exchange), and global business (multinational corporations). The *dataset*s utilised include COVID-19 data sourced from Kaggle, stock data listed on the Indonesia Stock Exchange, and corporate data comprising industry classifications and revenue attributes of global companies. The methodology adopted in this research encompasses several critical phases: *data preprocessing* to ensure consistency and reliability; implementation of classification and *clustering* algorithms; and model evaluation through accuracy metrics and visual analytics. Findings indicate that the *K-Means* algorithm performs effectively in *clustering* both COVID-19 spread regions and stock data based on numerical features. The *Decision tree* method demonstrates strong predictive capabilities in classifying risk categories within both COVID-19 *dataset*s and corporate profiles. PCA proves to be valuable in reducing data dimensionality while retaining essential information. Furthermore, *Bootstrapping* is employed to enhance the generalizability of the models, particularly in scenarios involving limited data samples. The study concludes that integrating multiple *data mining* approaches can yield comprehensive insights across sectors, although the level of effectiveness varies depending on the inherent characteristics of each *dataset*. Such a multidisciplinary and combined approach provides a robust framework for data-driven analysis and strategic decision support in diverse fields.

Keywords: *data mining, K-Means, Decision tree, covid-19, pca*

Abstrak

Kemajuan pesat dalam teknologi informasi telah memperluas pemanfaatan teknik *data mining* di berbagai bidang, khususnya dalam mengidentifikasi pola tersembunyi dan menunjang proses pengambilan keputusan berbasis data. Penelitian ini secara khusus mengkaji dan membandingkan efektivitas empat pendekatan *data mining* yakni *K-Means Clustering*, *Decision tree*, *Principal*

JURNAL REKAYASA INFORMASI
SWADHARMA (JRIS)
VOLUME 04 NO 01 JANUARI 2024
ISSN 2774 – 5759 | eISSN 2774-5732

component analysis (PCA), dan *Bootstrapping,* dalam mengolah data yang berasal dari tiga sektor strategis: sektor kesehatan (terkait COVID-19), sektor keuangan (pasar saham BEI), dan sektor bisnis global (perusahaan multinasional). *Dataset* yang digunakan bersumber dari berbagai platform terpercaya, termasuk data COVID-19 dari Kaggle, data saham perusahaan yang terdaftar di Bursa Efek Indonesia, serta informasi perusahaan multinasional yang mencakup variabel industri dan pendapatan tahunan. Rangkaian metodologi penelitian diawali dengan proses prapengolahan data (*data preprocessing*) untuk memastikan kualitas dan konsistensi data, dilanjutkan dengan penerapan algoritma klasifikasi dan pengelompokan (*clustering*), serta evaluasi performa model menggunakan metrik akurasi dan representasi visual. Dari hasil analisis yang dilakukan, ditemukan bahwa algoritma *K-Means* menunjukkan performa yang baik dalam mengelompokkan wilayah berdasarkan tingkat penyebaran COVID-19 serta dalam mengklasifikasikan saham berdasarkan indikator numerik. Sementara itu, metode *Decision tree* terbukti efektif dalam memprediksi kategori risiko, baik dalam konteks data kesehatan maupun data korporasi multinasional. PCA turut berkontribusi signifikan dalam mereduksi dimensi data tanpa kehilangan informasi utama yang relevan. Selain itu, teknik *Bootstrapping* diaplikasikan untuk meningkatkan kemampuan generalisasi model, terutama saat berhadapan dengan keterbatasan jumlah data. Secara keseluruhan, temuan penelitian ini menegaskan bahwa pendekatan kombinatif dalam *data mining* dapat menghasilkan wawasan mendalam yang lintas sektoral, dengan efektivitas yang bergantung pada karakteristik dan struktur data yang dianalisis. Pendekatan integratif semacam ini berpotensi memperkaya pemahaman dan mendukung pengambilan keputusan strategis di berbagai domain.

Kata Kunci: *data mining, K-Means, Decision tree, covid-19, pca*

A. PENDAHULUAN

Di era digital yang ditandai dengan ledakan volume data dalam berbagai format dan sumber, *data mining* muncul sebagai pendekatan kunci untuk mengekstraksi informasi yang bernilai dari kumpulan data yang besar dan kompleks. Teknik ini tidak lagi terbatas pada satu bidang tertentu, melainkan telah meluas secara signifikan ke berbagai sektor, seperti kesehatan, keuangan, bisnis global, hingga industri manufaktur (Marisa et al., 2021). Kemampuan *data mining* dalam mengidentifikasi pola tersembunyi, melakukan klasifikasi, segmentasi, serta prediksi menjadikannya sebagai alat yang sangat penting dalam proses pengambilan keputusan yang berbasis data dan bukti empiris (Pradnyana et al., 2020).

Penelitian ini secara khusus memfokuskan diri pada penerapan teknik *data mining* di tiga sektor utama yang berbeda karakter namun memiliki relevansi tinggi, yakni sektor kesehatan (terkait pandemi COVID-19), sektor keuangan (melalui analisis saham di Bursa Efek Indonesia), serta sektor bisnis global (melalui data perusahaan multinasional). Selain ketiga sektor tersebut, aspek industri manufaktur juga dilibatkan melalui analisis data *Master Production Schedule* (MPS). Dalam konteks *dataset* berskala kecil, teknik *Bootstrapping* diterapkan untuk memperkuat hasil analisis dan meningkatkan keandalan estimasi statistik.

Penelitian ini memanfaatkan empat metode utama *data mining* yaitu *Decision tree*, *K-Means Clustering*, *Principal component analysis (PCA)*, dan *Bootstrapping*, dengan tujuan untuk mengevaluasi serta membandingkan efektivitas masing-masing metode dalam merespons tantangan struktur data yang beragam dan kebutuhan analisis lintas sektor.

Adapun sumber data yang digunakan dalam penelitian ini mencakup:

1. Data COVID-19: Diambil dari platform Kaggle, meliputi informasi jumlah kasus terkonfirmasi, jumlah kematian, serta jumlah pasien yang sembuh.
2. Data Saham BEI: Bersumber dari data historis Bursa Efek Indonesia, mencakup kode saham, nama perusahaan emiten, dan kontribusi terhadap indeks harga saham gabungan.
3. Data Perusahaan Global: Berisi informasi mengenai jenis industri dan pendapatan tahunan perusahaan dari berbagai negara.
4. Data Produksi Industri: Berasal dari dokumen internal MPS milik PT. XYZ, yang menggambarkan jadwal produksi mingguan untuk beberapa varian produk.
5. Data Pendapatan Individu: Digunakan untuk mendemonstrasikan penerapan teknik *Bootstrapping* dalam konteks klasifikasi tingkat pendapatan.

Merujuk pada literatur yang relevan, (Romero & Ventura, 2020) mendefinisikan *data mining* sebagai proses sistematis dalam menemukan pola dan hubungan tersembunyi dari kumpulan data berskala besar. Dalam praktiknya, *Decision tree* banyak digunakan untuk tujuan klasifikasi karena sifatnya yang mudah dipahami serta divisualisasikan (Kurnia et al., 2020). Metode *K-Means Clustering* sangat efektif dalam mengelompokkan data numerik berdasarkan tingkat kemiripan antar data (Chen, 2024). PCA, di sisi lain, merupakan teknik reduksi dimensi yang mampu menyederhanakan struktur data tanpa menghilangkan informasi penting yang terkandung di dalamnya (Hediyati & Suartana, 2021). Sementara itu, seperti dijelaskan oleh (Agustian & Bisri, 2019), *Bootstrapping* merupakan pendekatan statistik berbasis resampling yang berguna untuk meningkatkan stabilitas estimasi, khususnya saat data yang tersedia terbatas.

Melalui pendekatan analisis komparatif antar metode dan antar sektor, penelitian ini bertujuan untuk memberikan gambaran menyeluruh mengenai efektivitas penerapan teknik *data mining* lintas bidang. Hasil dari penelitian ini diharapkan dapat menjadi dasar dalam merumuskan rekomendasi metode terbaik yang sesuai dengan jenis data dan tujuan analisis yang berbeda, serta memperkaya pemahaman dalam praktik analitik berbasis data di berbagai sektor strategis.

B. METODE PENELITIAN

Sumber Data

Penelitian ini memanfaatkan sejumlah *dataset* yang merepresentasikan sektor-sektor berbeda untuk menguji efektivitas metode *data mining* secara lintas bidang. Adapun sumber data yang digunakan meliputi:

1. *Dataset* COVID-19: Bersumber dari Kaggle (Rajkumar, 2020), mencakup variabel numerik seperti jumlah kasus terkonfirmasi (*Confirmed*), jumlah kematian (*Deaths*), dan jumlah pasien sembuh (*Recovered*) berdasarkan wilayah geografis.
2. Data Saham BEI: Berisi informasi terkait kode saham, nama emiten, serta kontribusi masing-masing terhadap Indeks Harga Saham Gabungan (IHSG).
3. Data Perusahaan Global: Merupakan *dataset* internal yang dirancang untuk penelitian ini, mengandung atribut kategorikal seperti sektor industri dan variabel numerik berupa pendapatan

JURNAL REKAYASA INFORMASI
SWADHARMA (JRIS)
VOLUME 05 NO 02 JULI 2025
ISSN 2774 – 5759 | eISSN 2774-5732

tahunan. Target klasifikasi ditetapkan pada variabel Negara asal perusahaan.
4. Data Jadwal Produksi (MPS): Diambil dari dokumen internal perusahaan manufaktur (PT. XYZ), berisi jadwal produksi mingguan dari beberapa varian produk.
5. Data Pendapatan Individu: Digunakan secara khusus untuk demonstrasi teknik *Bootstrapping* dalam konteks klasifikasi tingkat pendapatan.

Pra-pemrosesan Data

Sebelum diterapkan ke model, data mengalami beberapa tahapan pra-pemrosesan, antara lain:
1. Standarisasi fitur numerik dilakukan dengan *StandardScaler* untuk menyamakan skala variabel, khususnya pada metode PCA dan *K-Means*.
2. Pengkodean variabel kategorikal, baik melalui *Label Encoding* maupun *One-Hot Encoding*, diterapkan agar data dapat diolah oleh model klasifikasi seperti *Decision tree*.
3. Penanganan *missing values* dan *outlier* dilakukan secara selektif untuk menjaga kualitas dan integritas data.
4. *Bootstrapping* digunakan untuk mereplikasi data dalam jumlah terbatas, guna menghasilkan distribusi sampel yang lebih representatif dan mengurangi variabilitas model.

Algoritma yang Digunakan

1. *K-Means Clustering*

Metode ini diterapkan untuk mengelompokkan data numerik berdasarkan tingkat kemiripan antar atribut. Pada *dataset* COVID-19 dan saham, *K-Means* digunakan untuk membentuk klaster wilayah atau saham dengan karakteristik yang serupa. Tahapan Algoritma (*Pseudocode*):
a. Tentukan jumlah klaster (k)
b. Inisialisasi *centroid* secara acak
c. Hitung jarak setiap data ke centroid
d. Kelompokkan data berdasarkan centroid terdekat
e. Perbarui posisi centroid berdasarkan rerata klaster
f. Ulangi proses hingga centroid stabil (konvergen)

2. *Decision tree*

Decision tree digunakan untuk klasifikasi, seperti dalam memprediksi kategori tingkat kematian COVID-19 atau negara asal perusahaan. Kelebihan utama dari algoritma ini adalah kemudahan interpretasi dan visualisasi pohon keputusan yang dihasilkan.

3. *Principal component analysis (PCA)*

PCA berperan dalam mereduksi dimensi data tanpa kehilangan informasi penting. Metode ini digunakan untuk menyederhanakan data numerik kompleks, khususnya pada *dataset* COVID-19 dan saham, serta meningkatkan efektivitas visualisasi.

4. *Bootstrapping*

Digunakan dalam konteks *dataset* berskala kecil, seperti data pendapatan individu. Teknik ini menghasilkan sampel ulang secara acak (dengan pengembalian) untuk meningkatkan jumlah observasi dan mengurangi variabilitas estimasi model. *Pseudocode Bootstrapping*:
a. Ambil sampel acak dengan pengembalian dari *dataset* asli
b. Latih model pada data hasil sampel
c. Ulangi proses beberapa kali dan rata-rata hasil performa model

Evaluasi Model

Evaluasi efektivitas model dilakukan berdasarkan tipe metode yang digunakan:
1. *Clustering* (*K-Means*): Dinilai menggunakan Silhouette Score dan visualisasi distribusi klaster.
2. Klasifikasi (*Decision tree*): Dievaluasi menggunakan metrik seperti akurasi, confusion matrix, dan classification report.

3. PCA: Diperiksa melalui explained variance ratio, yaitu proporsi variansi yang dijelaskan oleh tiap komponen utama.

Tools dan Perangkat Lunak

Seluruh eksperimen dilakukan menggunakan Python 3.x, dengan dukungan beberapa pustaka dan framework berikut:

1. pandas dan numpy : Untuk manipulasi dan analisis data
2. scikit-learn : Untuk penerapan algoritma *data mining* dan evaluasi model
3. matplotlib dan seaborn : Untuk visualisasi data dan hasil analisis
4. statsmodels : Untuk kebutuhan analisis statistik tambahan, terutama pada *Bootstrapping*.

C. HASIL DAN PEMBAHASAN

Dataset dan Hasil COVID – 19

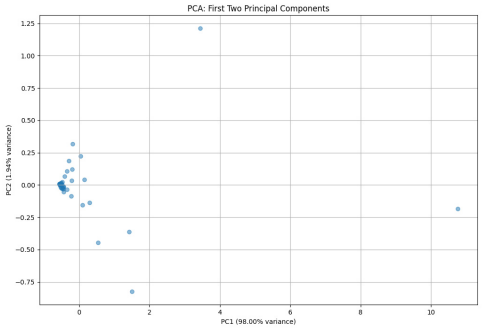
Tabel 1. *Dataset* Covid-19

Province/State	*Country/Region*	*Confirmed*	*Deaths*	*Recovered*
Quindio	Colombia	2941	87	1905
Quintana Roo	Mexico	11478	1598	9485
Rajasthan	India	118793	1367	98812
Reunion	France	3501	15	2482
Rheinland	Germany	10246	248	9271
Rhode Island	US	24177	1102	0
Rio Grande	Brazil	67459	2355	39924

Dataset terkait penyebaran COVID-19 yang dianalisis dalam penelitian ini diperoleh dari platform Kaggle, dan berisi informasi komprehensif mengenai jumlah kasus terkonfirmasi (*Confirmed*), jumlah kematian (*Deaths*), serta jumlah pasien yang dinyatakan sembuh (*Recovered*) di berbagai wilayah dunia. Secara keseluruhan, *dataset* ini terdiri atas 54 entri, di mana masing-masing entri merepresentasikan kombinasi unik antara kolom *Province/State* dan *Country/Region*, yang mencerminkan cakupan geografis distribusi data.

Setiap baris pada *dataset* menggambarkan kondisi epidemiologis wilayah tertentu pada saat data dikumpulkan. Kolom *Confirmed* menunjukkan akumulasi kasus positif yang telah terverifikasi, *Deaths* mencatat total kematian yang diakibatkan oleh virus COVID-19, sementara *Recovered* mengindikasikan jumlah pasien yang telah dinyatakan sembuh. Perlu dicatat bahwa beberapa entri memiliki nilai nol pada kolom *Recovered*, yang disebabkan oleh belum tersedianya data kesembuhan pada waktu pengambilan data, atau karena laporan resmi dari wilayah tersebut belum dirilis.

Dataset ini dijadikan dasar penerapan beberapa metode analitik. Algoritma *K-Means Clustering* digunakan untuk mengelompokkan wilayah berdasarkan tingkat keparahan penyebaran virus, dengan mempertimbangkan kombinasi variabel numerik yang tersedia. Selanjutnya, *Decision tree* diterapkan untuk melakukan klasifikasi wilayah ke dalam kategori tingkat kematian, yang dihitung berdasarkan rasio antara jumlah kematian dan jumlah kasus terkonfirmasi. Selain itu, *Principal component analysis* (PCA) dimanfaatkan sebagai teknik reduksi dimensi untuk menyederhanakan data menjadi dua komponen utama, dengan tujuan mempermudah visualisasi dan pemahaman pola penyebaran pandemi secara global. Berikut adalah hasil programnya :



JURNAL REKAYASA INFORMASI
SWADHARMA (JRIS)
VOLUME 05 NO 02 JULI 2025
ISSN 2774 – 5759 | eISSN 2774-5732

Gambar 1. Visualisasi *Dataset* COVID – 19, PCA Metode

Gambar tersebut menyajikan hasil visualisasi dua dimensi dari proses *Principal component analysis* (PCA) yang diterapkan pada *dataset* COVID-19. Analisis ini dilakukan terhadap tiga variabel utama, yaitu *Confirmed* (jumlah kasus terkonfirmasi), *Deaths* (jumlah kematian), dan *Recovered* (jumlah kesembuhan). Sebelum dilakukan transformasi PCA, seluruh fitur numerik telah melalui proses standarisasi menggunakan metode *StandardScaler*, untuk memastikan kesetaraan skala antar variabel. Dalam grafik yang dihasilkan, sumbu horizontal (PC1) merepresentasikan komponen utama pertama yang menjelaskan sekitar 98,00% dari total variansi dalam *dataset*, sedangkan sumbu vertikal (PC2) hanya menjelaskan sekitar 1,94% variansi. Temuan ini mengindikasikan bahwa hampir seluruh informasi yang signifikan dalam data dapat direpresentasikan hanya oleh satu dimensi utama, yaitu PC1. Secara statistik, besar kemungkinan bahwa dimensi ini didominasi oleh variabel *Confirmed*, mengingat variabel ini memiliki skala nilai tertinggi dan kontribusi terbesar terhadap variasi antar wilayah. Setiap titik pada grafik mewakili satu entri wilayah, yang merupakan kombinasi antara *Province/State* dan *Country/Region* dalam *dataset*. Titik-titik yang terdistribusi dekat dengan pusat koordinat (nilai mendekati nol pada PC1 dan PC2) menunjukkan wilayah-wilayah dengan jumlah kasus, kematian, dan kesembuhan yang cenderung rendah atau relatif mendekati rata-rata. Sebaliknya, titik-titik yang terletak jauh dari pusat—terutama pada sisi kanan atas atau kanan bawah grafik menunjukkan *outlier*, yaitu wilayah dengan jumlah kasus atau tingkat kematian yang jauh lebih tinggi dibandingkan wilayah lainnya.

Visualisasi ini berfungsi sebagai alat yang sangat efektif untuk mereduksi kompleksitas dimensi data dan secara eksploratif mengidentifikasi pola-pola penyebaran yang ekstrem atau tidak umum. Selain membantu pemahaman terhadap struktur data secara keseluruhan, hasil PCA ini juga dapat dijadikan landasan awal untuk proses pengelompokan wilayah (clustering) atau penetapan prioritas dalam penanganan pandemi berdasarkan profil statistik masing-masing wilayah.

Dataset & Hasil Saham BEI

Tabel 2. *Dataset* Saham BEI

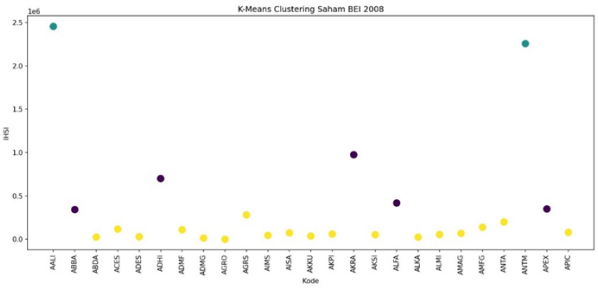
Kode	Nama Emiten	IHSI	Kode
AALI	Astra Agro Lestari Tbk	2454187	AALI
ABBA	Abdi Bangsa Tbk	342105	ABBA
ABDA	Asuransi Bina Dana Arta Tbk	24732	ABDA
ACES	Ace Hardware Indonesia Tbk	117073	ACES
ADES	Ades Waters Indonesia Tbk	29091	ADES

Dataset saham yang digunakan dalam penelitian ini diperoleh dari daftar perusahaan yang terdaftar secara resmi di Bursa Efek Indonesia (BEI). *Dataset* ini terdiri atas 25 entri data, dengan masing-masing baris merepresentasikan satu perusahaan publik atau emiten yang telah mencatatkan sahamnya di pasar modal. Setiap entri mengandung sejumlah informasi penting yang menjadi dasar dalam proses analisis, di antaranya:

1. Kode Saham: Merupakan singkatan resmi dari masing-masing saham yang terdiri dari empat huruf kapital, dan berfungsi sebagai identitas unik di sistem perdagangan BEI.

2. Nama Emiten: Menunjukkan nama lengkap dari perusahaan yang menerbitkan saham tersebut.
3. IHSI (Indeks Harga Saham Individual): Menggambarkan besarnya kontribusi saham terhadap pergerakan indeks harga secara individu. Nilai IHSI dalam *dataset* bersifat kuantitatif dan digunakan untuk mengukur seberapa besar pengaruh suatu saham terhadap dinamika fluktuasi indeks di pasar modal.
4. Nilai IHSI digunakan sebagai indikator utama dalam menilai bobot relatif dari masing-masing saham dalam pergerakan pasar. Saham dengan nilai IHSI yang lebih tinggi memiliki kontribusi yang lebih signifikan terhadap perubahan indeks agregat, sehingga dianggap lebih dominan dalam mempengaruhi arah pasar. Sebaliknya, saham dengan nilai IHSI yang rendah cenderung memiliki pengaruh yang lebih kecil terhadap kinerja indeks.

Dalam konteks penelitian ini, *dataset* saham dimanfaatkan sebagai landasan untuk menerapkan dua metode analitik, yaitu *K-Means Clustering* dan *Principal component analysis* (PCA). Analisis clustering dilakukan untuk mengelompokkan saham-saham berdasarkan karakteristik kontribusinya terhadap indeks, sehingga diperoleh klasifikasi saham dalam kategori berpengaruh tinggi, sedang, dan rendah. Sementara itu, PCA digunakan untuk mereduksi kompleksitas data ke dalam dua dimensi utama, guna memudahkan visualisasi pola distribusi kontribusi saham secara lebih intuitif. Teknik ini dapat memberikan insight yang berguna bagi investor maupun analis pasar dalam mengidentifikasi klaster saham dengan dampak signifikan terhadap fluktuasi indeks, serta memfasilitasi pengambilan keputusan investasi yang lebih terarah. Berikut adalah hasil programnya :



Gambar 2. Visualisasi *Dataset* Saham BEI, K-Means Metode

Gambar yang ditampilkan merepresentasikan hasil penerapan algoritma *K-Means Clustering* terhadap data saham perusahaan publik yang terdaftar di Bursa Efek Indonesia (BEI) pada tahun 2008. *Dataset* yang dianalisis mencakup 25 entri, dengan setiap titik data merepresentasikan satu emiten saham. Atribut utama yang digunakan dalam proses pengelompokan adalah nilai IHSI (Indeks Harga Saham Individual), yang menggambarkan kontribusi relatif masing-masing saham terhadap fluktuasi indeks pasar.

Sebelum proses klasterisasi dilakukan, data IHSI terlebih dahulu distandarisasi menggunakan metode *StandardScaler* untuk memastikan kesetaraan skala antar data. Selanjutnya, algoritma K-Means diterapkan dengan parameter jumlah klaster (k) sebanyak tiga, yang ditentukan berdasarkan hasil analisis menggunakan *Elbow Method*, sebuah pendekatan visual yang umum digunakan untuk menentukan jumlah klaster optimal.

Dalam visualisasi tersebut, setiap titik menggambarkan satu saham, dengan posisi vertikal yang mencerminkan nilai IHSI setelah distandarisasi, dan warna yang berbeda menandakan keanggotaan klaster masing-masing. Tiga klaster utama yang terbentuk dapat dijelaskan sebagai berikut:
1. Klaster kuning mencakup saham-saham dengan nilai IHSI yang rendah, yang berarti kontribusinya terhadap indeks pasar relatif kecil.

2. Klaster ungu tua terdiri atas saham-saham dengan kontribusi menengah, berada di antara nilai ekstrem bawah dan atas.
3. Klaster biru kehijauan (cyan) menampung saham dengan nilai IHSI yang sangat tinggi dan tergolong sebagai *outlier*, seperti saham AALI (Astra Agro Lestari) yang secara signifikan menyimpang dari pola umum distribusi.

Distribusi hasil klasterisasi ini menunjukkan bahwa sebagian besar saham yang dianalisis memiliki pengaruh kecil hingga sedang terhadap pergerakan indeks pasar, sementara hanya segelintir saham yang memberikan kontribusi besar dan dominan. Informasi ini sangat berguna dalam proses segmentasi saham berdasarkan bobot pengaruhnya, yang dapat dijadikan dasar untuk strategi investasi yang lebih terfokus, penyusunan portofolio, maupun pengambilan keputusan berdasarkan tingkat risiko dan dominasi pasar dari masing-masing saham.

Dataset dan Hasil Perusahaan Global

Tabel 3. *Dataset* Perusahaan Global

Perusahaan	Negara	Industri	Pendapatan dalam miliar
América Móvil	Meksiko	Layanan telekomunikasi	17
Cemex	Meksiko	Bahan bangunan	15.3
China Mobile	Cina	Layanan telekomunikasi	30.1
CNOOC	Cina	Minyak dan gas	8.7
CVRD	Brasil	Pertambangan	15.1

Dataset perusahaan global yang digunakan dalam penelitian ini terdiri atas data dari 25 perusahaan multinasional yang berasal dari berbagai negara berkembang. Tujuan utama dari penggunaan *dataset* ini adalah untuk melakukan proses klasifikasi negara asal masing-masing perusahaan, berdasarkan dua atribut utama, yaitu jenis industri tempat perusahaan beroperasi dan pendapatan tahunan yang dinyatakan dalam satuan miliar dolar Amerika Serikat.

Struktur *dataset* mencakup empat kolom utama sebagai berikut:

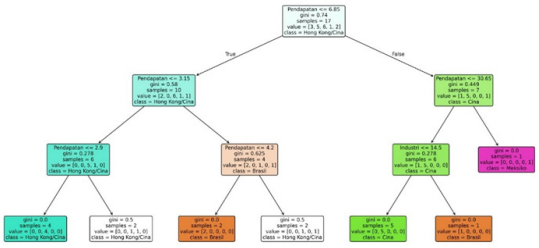
1. Perusahaan: Nama entitas bisnis multinasional yang menjadi objek pengamatan.
2. Negara: Negara tempat perusahaan tersebut berdiri dan beroperasi secara hukum, yang menjadi target klasifikasi.
3. Industri: Sektor ekonomi utama yang menjadi fokus operasional perusahaan, seperti energi, teknologi, manufaktur, atau jasa keuangan.
4. Pendapatan dalam miliar: Total pendapatan tahunan perusahaan dalam satuan miliar USD, yang merepresentasikan ukuran skala bisnis dari masing-masing entitas.

Dataset ini digunakan sebagai dasar dalam eksperimen klasifikasi menggunakan algoritma *Decision tree*, dengan tujuan untuk mengevaluasi sejauh mana informasi tentang sektor industri dan nilai pendapatan mampu memprediksi negara asal perusahaan. Penggunaan *Decision tree* dipilih karena kemampuannya dalam menangani data kategorikal dan numerik secara bersamaan serta menghasilkan model yang dapat divisualisasikan dan diinterpretasikan dengan mudah.

Selain itu, untuk keperluan analisis eksploratif, metode *Principal component analysis* (PCA) juga dapat diterapkan pada fitur numerik, khususnya pendapatan, guna menyederhanakan representasi data dan mengidentifikasi pola distribusi atau keberadaan *outlier* di antara perusahaan-perusahaan tersebut. PCA memungkinkan pemetaan perusahaan dalam ruang dua dimensi sehingga perbedaan karakteristik berdasarkan pendapatan lebih mudah diamati secara visual.

Walaupun ukuran *dataset* ini tergolong kecil, keberagaman sektor industri serta variasi yang cukup signifikan dalam nilai pendapatan memberikan ruang untuk

pengujian metode klasifikasi dalam skala terbatas. Dengan demikian, *dataset* ini dapat berfungsi sebagai representasi awal dalam mengkaji keterkaitan antara karakteristik ekonomi perusahaan dengan negara asalnya, sekaligus membuka peluang untuk penelitian lanjutan dengan cakupan data yang lebih luas. Berikut adalah hasil programnya :



Gambar 3. Visualisasi *Dataset* Perusahaan Global, *Decision tree* Metode

Gambar di atas menampilkan hasil visualisasi model *Decision tree* yang dirancang untuk melakukan klasifikasi negara asal perusahaan global berdasarkan dua fitur utama, yaitu Pendapatan tahunan (dalam miliar USD) dan jenis industri. *Dataset* yang digunakan mencakup 25 perusahaan multinasional yang berasal dari beberapa negara berkembang, seperti Meksiko, Tiongkok, Brasil, dan Hong Kong.

Tujuan utama dari penerapan *Decision tree* ini adalah untuk mengevaluasi sejauh mana kombinasi antara atribut numerik (Pendapatan) dan kategorikal (Industri) dapat digunakan untuk memprediksi negara asal suatu perusahaan. Struktur pohon dibentuk secara hierarkis, dengan proses pemisahan (*splitting)* dimulai dari fitur yang paling informatif. Dalam hal ini, Pendapatan muncul sebagai atribut yang paling dominan, berperan sebagai *root node* dalam pembentukan pohon keputusan.

Beberapa temuan penting dari hasil visualisasi antara lain:

1. Node akar (*root node*) membagi *dataset* berdasarkan ambang nilai pendapatan sebesar 6,85 miliar USD. Perusahaan dengan pendapatan di bawah ambang ini sebagian besar berasal dari wilayah Hong Kong atau Tiongkok, sedangkan perusahaan dengan pendapatan di atas ambang cenderung berasal dari Tiongkok dan Meksiko.
2. Pada cabang kanan pohon, proses pemisahan dilanjutkan dengan pertimbangan nilai pendapatan yang lebih tinggi dan jenis industri tertentu, yang mengarah pada klasifikasi ke negara-negara seperti Brasil, Meksiko, dan Tiongkok.
3. Nilai *Gini impurity* yang rendah, bahkan hingga 0,0 pada beberapa node daun, menunjukkan bahwa pada node-node tersebut hanya terdapat satu kelas (negara) dalam data, yang menandakan tingkat kemurnian klasifikasi yang sangat tinggi.

Sebagai contoh, terdapat satu node daun dengan gini = 0,0 dan samples = 1, yang secara akurat mengklasifikasikan satu entitas perusahaan ke dalam negara Meksiko, berdasarkan kriteria pendapatan yang melebihi 30,65 miliar USD.

Struktur hierarkis pohon ini mengungkapkan bahwa pendapatan perusahaan merupakan indikator kunci dalam membedakan negara asal perusahaan multinasional, terutama ketika dikombinasikan dengan sektor industrinya. Hasil ini tidak hanya memperlihatkan akurasi model klasifikasi yang dibangun, tetapi juga menunjukkan potensi kuat metode *Decision tree* dalam menganalisis dan memetakan karakteristik ekonomi perusahaan terhadap asal geografisnya. Temuan ini dapat dimanfaatkan untuk penelitian lebih lanjut dalam bidang analitik bisnis internasional maupun pengambilan keputusan strategis berbasis data lintas negara

D. PENUTUP

Penelitian ini telah melakukan analisis komparatif terhadap sejumlah metode *data*

mining dengan menerapkannya pada beragam *dataset* yang berasal dari tiga sektor berbeda, yakni sektor kesehatan (data penyebaran COVID-19), sektor keuangan (data saham dari Bursa Efek Indonesia), serta sektor bisnis global (data perusahaan multinasional). Metode-metode yang digunakan mencakup *K-Means Clustering* untuk proses segmentasi, *Decision tree* untuk klasifikasi, *Principal component analysis (PCA)* untuk reduksi dimensi dan visualisasi data, serta *Bootstrapping* guna memperkuat performa model, khususnya dalam konteks *dataset* yang berskala kecil.

Berdasarkan hasil eksperimen, diperoleh beberapa temuan penting: (1) *K-Means Clustering* terbukti efektif dalam mengelompokkan data berdasarkan pola numerik yang ada. Pada kasus COVID-19, metode ini berhasil membentuk tiga klaster wilayah berdasarkan intensitas kasus dan kematian. Sementara pada data saham BEI, metode ini mampu mengidentifikasi perbedaan antara saham yang berpengaruh besar terhadap indeks dan saham dengan kontribusi yang lebih rendah; (2) *Decision tree* memberikan hasil klasifikasi yang cukup akurat. Pada data kesehatan, algoritma ini mampu memprediksi tingkat kematian berdasarkan kombinasi antara kasus terkonfirmasi dan jumlah kesembuhan. Dalam konteks data perusahaan global, *Decision tree* berhasil memetakan negara asal perusahaan berdasarkan sektor industri dan pendapatan tahunan; (3) *Principal component analysis (PCA)* secara signifikan berhasil mereduksi dimensi data menjadi dua komponen utama dengan total variansi yang terjaga di atas 98%, tanpa kehilangan informasi esensial. Hal ini sangat membantu dalam menyederhanakan struktur data COVID-19 dan saham, sekaligus mempermudah proses visualisasi dan interpretasi; (4) *Bootstrapping* memberikan kontribusi terhadap peningkatan stabilitas evaluasi model, terutama pada *dataset* kecil seperti klasifikasi tingkat pendapatan individu, melalui teknik resampling yang menghasilkan distribusi data yang lebih representatif.

Temuan dalam penelitian ini memiliki sejumlah potensi penerapan praktis dalam berbagai sektor: (1) Kesehatan masyarakat: Hasil klasterisasi wilayah berbasis tingkat penyebaran COVID-19 dapat dijadikan acuan dalam menetapkan prioritas alokasi sumber daya dan intervensi medis; (2) Pasar modal dan investasi: Pengelompokan saham berdasarkan kontribusi terhadap indeks membantu investor dalam mengidentifikasi kelompok saham dengan performa serupa, sehingga dapat meningkatkan efisiensi strategi diversifikasi portofolio; (3) Analisis bisnis global: Proses klasifikasi terhadap perusahaan multinasional berkontribusi pada pemahaman lebih lanjut mengenai karakteristik ekonomi negara berkembang, terutama dari sisi sektor industri dominan dan kapasitas pendapatan; (4) Manajemen produksi industri: Pemanfaatan data Master Production Schedule (MPS) dan analisis klaster memungkinkan optimalisasi penjadwalan produksi dan pengelolaan sumber daya secara lebih adaptif dan efisien.

Meski telah menunjukkan hasil yang menjanjikan, penelitian ini memiliki sejumlah keterbatasan: (1) Beberapa *dataset*, seperti data perusahaan global dan data produksi industri (MPS), memiliki jumlah entri yang terbatas, sehingga dapat memengaruhi akurasi model serta keterbatasan dalam proses generalisasi hasil; (2) Faktor-faktor eksternal yang bersifat temporal atau demografis belum sepenuhnya terintegrasi dalam proses analisis, padahal variabel-variabel tersebut dapat memperkaya konteks interpretasi; (3) Evaluasi performa model belum mencakup perbandingan komprehensif lintas

algoritma alternatif (misalnya, Random Forest, Support Vector Machine, atau Gradient Boosting) yang berpotensi memberikan hasil lebih optimal.

Sebagai arahan untuk penelitian lanjutan, beberapa rekomendasi yang dapat dipertimbangkan antara lain: (1) Menggunakan *dataset* yang lebih besar dan beragam, dengan penambahan atribut waktu, lokasi geografis, atau data demografi guna memperluas cakupan dan kedalaman analisis; (2) Melakukan perbandingan sistematis antar algoritma, baik untuk klasifikasi maupun *clustering*, guna memperoleh gambaran yang lebih holistik tentang performa masing-masing metode; (3) Mengembangkan dashboard visual interaktif atau sistem pendukung keputusan (Decision Support System/DSS) untuk membantu implementasi hasil analisis ke dalam konteks dunia nyata; (4) Melibatkan pendekatan analitik prediktif lanjutan, seperti time-series forecasting dan deep learning, untuk mengolah data berskala besar dan kompleks secara lebih adaptif dan akurat.

E. DAFTAR PUSTAKA

Agustian, A. A., & Bisri, A. (2019). Data Mining Optimization Using Sample Bootstrapping and Particle Swarm Optimization in the Credit Approval Classification. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *2*(1), 18–27. https://doi.org/10.24014/ijaidm.v2i1.6299

Chen, Q. (2024). Application of K-Means Algorithm in Marketing. *Advances in Economics, Management and Political Sciences*, *71*, 178–184. https://doi.org/10.54254/2754-1169/71/20241485

Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *JIEET : Journal of Information Engineering and Educational Technology*, *5*(2), 49–54. https://doi.org/10.26740/jieet.v5n2.p49-54

Kurnia, A., Mirza, A. H., & Andri, A. (2020). Penerapan Decision Tree Data Mining Pada Produksi Kelapa Sawit PT Hindoli Di Sungai Lilin Kabupaten Musi Banyuasin. *Jurnal Pengembangan Sistem Informasi Dan Informatika*, *1*(2), 84–99. https://doi.org/10.47747/jpsii.v1i2.168

Marisa, F., Maukar, A. L., & Akhriza, T. M. (2021). *Data Mining Konsep dan Penerapannya*. Yogyakarta : Deepublish.

Pradnyana, G. A., Darmawiguna, I. G. M., & Wijaya, I. N. S. W. (2020). *Data Mining: Menemukan Pengetahuan dalam Data*. Depok : PT Raja Grafindo Persada.

Rajkumar, S. (2020). *COVID-19 in India: Dataset on Novel Corona virus disease 2019 in India*.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wires: Data Mining & Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355