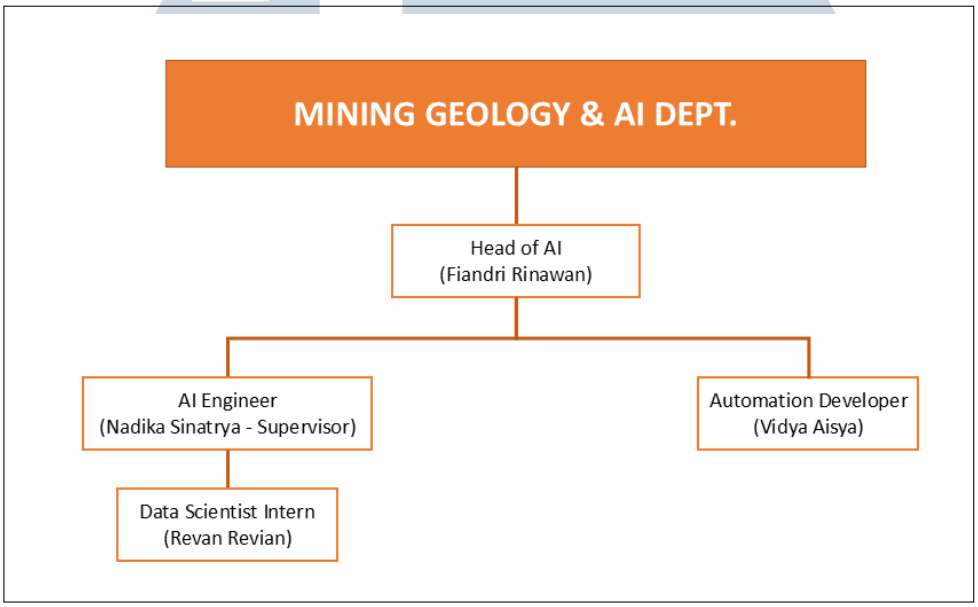
BAB 3
PELAKSANAAN KERJA MAGANG

3.1 Kedudukan dan Koordinasi



Gambar 3.1. Struktur organisasi Mining Geology & AI Department pada divisi CKBE

Selama pelaksanaan kerja magang di PT Berau Coal Energy, penempatan dilakukan di divisi CKBE pada Mining Geology & AI Department dengan posisi sebagai Data Scientist Intern. Struktur organisasi dari Mining Geology & AI Department ditampilkan pada Gambar 3.1, yang memperlihatkan posisi *intern* dalam hierarki departemen serta alur koordinasi antaranggota tim. Dalam kedudukan tersebut, tanggung jawab diberikan secara langsung oleh *supervisor* perusahaan yang berposisi sebagai AI Engineer, dan koordinasi dilakukan secara rutin untuk memastikan seluruh tugas berjalan sesuai arahan dan standar yang telah ditentukan.

Dengan sistem kerja yang bersifat *hybrid*, koordinasi dilakukan secara daring melalui platform Zoom, baik dalam bentuk percakapan teks maupun pertemuan video. Untuk menunjang kegiatan pemrograman dan analisis data, digunakan Google Colab sebagai lingkungan kerja utama. Selain itu, koordinasi tatap muka juga dilakukan secara berkala di kantor Jakarta saat pelaksanaan WFO.

3.2 Tugas yang Dilakukan

Selama pelaksanaan kerja magang, tugas utama yang dilakukan mencakup analisis data, pengembangan model *machine learning*, serta evaluasi hasil analisis untuk mendukung pengambilan keputusan strategis di perusahaan. Uraian lebih detail mengenai pekerjaan yang dilaksanakan setiap minggunya dapat dilihat pada Tabel 3.1.

Tabel 3.1. Aktivitas Mingguan Selama Magang

Minggu Ke-	**Pekerjaan yang dilakukan**
1	Pengenalan proyek, pembagian tugas, serta pengumpulan data awal untuk analisis.
2-3	Pengembangan *pipeline* otomatis untuk pengumpulan data, analisis sentimen, serta evaluasi kualitas model analisis teks.
4-5	Integrasi dan optimalisasi model berbasis AI untuk analisis risiko dari data berita keuangan.
6	Riset, pengumpulan, serta eksplorasi data faktor-faktor pasar komoditas.
7	Pengumpulan data lanjutan serta analisis sentimen berita industri secara lebih luas.
8	Pengembangan dan evaluasi model prediksi harga komoditas dengan pendekatan statistik dan *machine learning*.
9-10	Eksplorasi, pengembangan fitur, serta evaluasi model *forecasting hybrid* untuk komoditas.
11-12	Optimasi, *debugging*, serta eksperimen tambahan untuk meningkatkan akurasi model *forecasting*.
13-14	Studi dan implementasi metode baru untuk meningkatkan performa *forecasting*.
15-16	Integrasi data eksternal dan pengembangan model lanjutan untuk prediksi harga komoditas, serta pengenalan data *well logging* geofisika.
17	Eksplorasi data *well logging* geofisika, *clustering*, serta evaluasi hasil identifikasi zona eksplorasi batubara.

Tabel 3.1 Aktivitas Mingguan Selama Magang (lanjutan)

Minggu Ke-	Pekerjaan yang dilakukan
18-19	Melakukan evaluasi klaster melalui *plotting* dan statistik sambil membandingkan pendekatan *clustering*, serta merumuskan dan memvalidasi fitur tambahan untuk analisis.

3.3 Uraian Pelaksanaan Magang

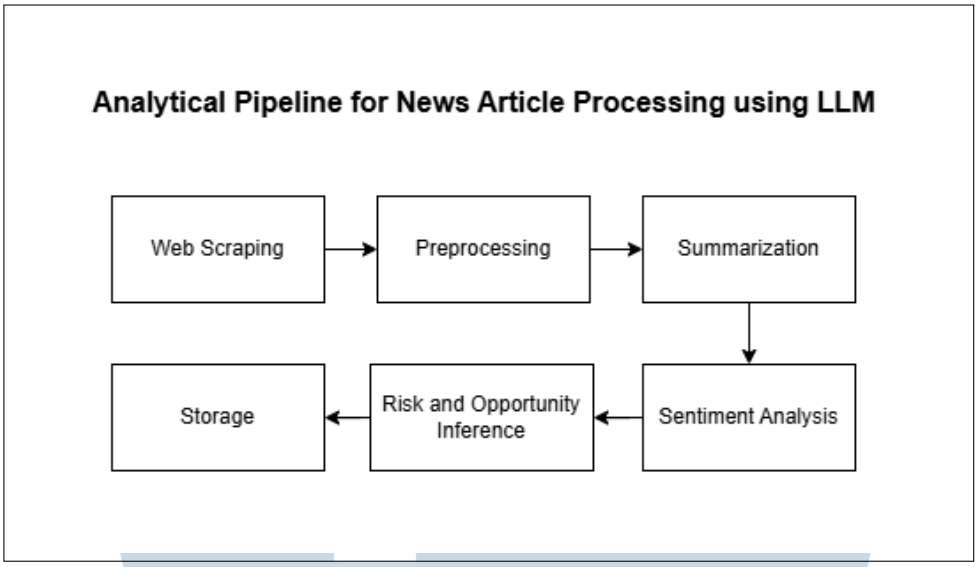
3.3.1 Implementasi LLM untuk Analisis Artikel Berita Daring Terkait Perusahaan Tambang Global

A *User Requirements*

Untuk mendukung proses pemantauan perkembangan industri pertambangan secara otomatis, tim strategi menetapkan beberapa spesifikasi teknis yang harus dipenuhi oleh sistem analitik yang dikembangkan. Sistem tersebut diharapkan mampu mengumpulkan dan mengolah data teks tidak terstruktur dari artikel berita secara otomatis, meliputi ekstraksi informasi relevan, *summarization*, *sentiment analysis*, serta identifikasi sinyal risiko maupun peluang bisnis secara *real time*.

Teknologi yang digunakan berbasis LLM, yang mampu memproses data teks dalam jumlah besar dengan akurasi tinggi dan waktu inferensi cepat. Sistem harus dapat mendukung *sentiment analysis* dalam kategori positif, negatif, dan netral, serta mampu menghasilkan ringkasan teks yang ringkas namun komprehensif untuk tiap artikel. Selain itu, sistem wajib memiliki kapabilitas untuk mendeteksi dan menandai sinyal risiko maupun peluang berdasarkan pola linguistik yang dipelajari dari data historis.

Pipeline dari sistem analisis berita ditunjukkan pada Gambar 3.2, yang memuat tahapan mulai dari *scraping* artikel daring hingga penyimpanan hasil analisis risiko dan peluang.



Gambar 3.2. *Pipeline* untuk menganalisis artikel berita daring terkait pertambangan menggunakan LLM

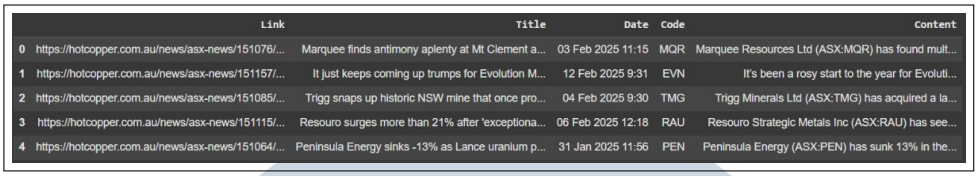
B Implementasi

B.1 *Web Scraping*

Langkah pertama dalam *pipeline* adalah pengumpulan artikel berita dari situs https://hotcopper.com.au, yang dipilih karena secara konsisten mempublikasikan berita terkini seputar industri pertambangan global, termasuk sektor mineral, energi, dan logam mulia. Situs ini dinilai relevan dan kaya informasi untuk keperluan analitik strategis.

Proses pengambilan data dilakukan dengan menggunakan kombinasi *library* Selenium dan BeautifulSoup. Selenium digunakan untuk mengontrol *browser* secara otomatis, seperti memuat halaman utama kategori berita pertambangan dan menekan tombol `Load More` sebanyak tiga kali untuk memperluas cakupan artikel yang diambil. Setelah seluruh konten dimuat, BeautifulSoup digunakan untuk memproses struktur HTML dan mengekstrak elemen penting dari setiap artikel, yaitu tautan, judul, tanggal publikasi, kode saham, dan isi artikel. Konten kemudian melalui proses sanitasi untuk menghapus *tag* HTML, simbol aneh, dan karakter tidak relevan lainnya. Artikel duplikat juga diidentifikasi dan dihapus berdasarkan kemiripan judul dan tautan. Hasil akhir disimpan dalam format file CSV agar dapat digunakan pada tahap pemrosesan berikutnya.

Contoh hasil dari proses *web scraping* disajikan dalam Gambar 3.3 berikut.



	Link	Title	Date	Code	Content
0	https://hotcopper.com.au/news/asx-news/151076/...	Marquee finds antimony aplenty at Mt Clement a...	03 Feb 2025 11:15	MQR	Marquee Resources Ltd (ASX:MQR) has found mult...
1	https://hotcopper.com.au/news/asx-news/151157/...	It just keeps coming up trumps for Evolution M...	12 Feb 2025 9:31	EVN	It's been a rosy start to the year for Evoluti...
2	https://hotcopper.com.au/news/asx-news/151085/...	Trigg snaps up historic NSW mine that once pro...	04 Feb 2025 9:30	TMG	Trigg Minerals Ltd (ASX:TMG) has acquired a la...
3	https://hotcopper.com.au/news/asx-news/151115/...	Resouro surges more than 21% after 'exceptiona...	06 Feb 2025 12:18	RAU	Resouro Strategic Metals Inc (ASX:RAU) has see...
4	https://hotcopper.com.au/news/asx-news/151064/...	Peninsula Energy sinks -13% as Lance uranium p...	31 Jan 2025 11:56	PEN	Peninsula Energy (ASX:PEN) has sunk 13% in the...

Gambar 3.3. Contoh hasil proses *web scraping* terhadap artikel berita daring

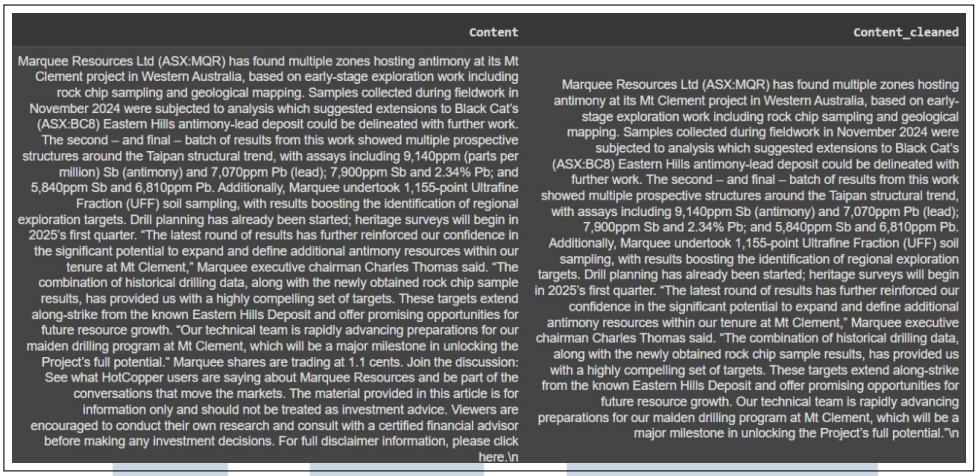
B.2 *Preprocessing*

Setelah artikel berita dikumpulkan melalui proses *web scraping*, tahap selanjutnya adalah melakukan *preprocessing* teks untuk memastikan bahwa data *input* memiliki format yang bersih, konsisten, dan sesuai dengan kebutuhan model. Tahap ini sangat penting untuk meningkatkan kualitas representasi teks dan menghindari gangguan dalam proses pemodelan.

Preprocessing dimulai dengan pembersihan elemen-elemen non-teks seperti *tag* HTML, simbol khusus, karakter *unicode* aneh, dan *whitespace* berlebih. Karakter-karakter seperti *newline*, tabulasi, atau *emoji* juga dihapus agar teks lebih seragam dan mudah diproses. Selain itu, dilakukan proses eksplisit untuk mengecualikan bagian-bagian artikel yang tidak relevan terhadap konten utama, seperti nama penulis, *footer* otomatis, tautan *internal*, *legal disclaimer*, dan ajakan interaksi pengguna (misalnya: *Join the discussion*). Penyaringan ini penting untuk memastikan bahwa hanya isi utama artikel yang dianalisis oleh model, tanpa gangguan dari elemen tambahan yang sering muncul pada situs berita daring.

Beberapa artikel yang memiliki isi sangat pendek atau kosong juga difilter secara otomatis untuk menjaga kualitas *input*. Selanjutnya, panjang teks dibatasi agar tetap berada dalam batas *token* maksimum yang dapat diproses oleh model *transformer*, yaitu antara 512 hingga 1024 *token*. Jika artikel melebihi batas ini, maka hanya bagian awal (*leading segment*) yang diambil karena umumnya mengandung informasi paling penting dan ringkasan utama dari berita. Hasil akhir dari tahap ini berupa teks yang bersih, terstruktur, dan siap untuk dimasukkan ke dalam *pipeline* analitik berbasis LLM pada tahap berikutnya.

Gambar 3.4 memperlihatkan contoh hasil teks sebelum dan sesudah proses preprocessing dilakukan.



Content	Content_cleaned
Marquee Resources Ltd (ASX:MQR) has found multiple zones hosting antimony at its Mt Clement project in Western Australia, based on early-stage exploration work including rock chip sampling and geological mapping. Samples collected during fieldwork in November 2024 were subjected to analysis which suggested extensions to Black Cat's (ASX:BC8) Eastern Hills antimony-lead deposit could be delineated with further work. The second – and final – batch of results from this work showed multiple prospective structures around the Taipan structural trend, with assays including 9,140ppm (parts per million) Sb (antimony) and 7,070ppm Pb (lead); 7,900ppm Sb and 2.34% Pb; and 5,840ppm Sb and 6,810ppm Pb. Additionally, Marquee undertook 1,155-point Ultrafine Fraction (UFF) soil sampling, with results boosting the identification of regional exploration targets. Drill planning has already been started; heritage surveys will begin in 2025's first quarter. "The latest round of results has further reinforced our confidence in the significant potential to expand and define additional antimony resources within our tenure at Mt Clement," Marquee executive chairman Charles Thomas said. "The combination of historical drilling data, along with the newly obtained rock chip sample results, has provided us with a highly compelling set of targets. These targets extend along-strike from the known Eastern Hills Deposit and offer promising opportunities for future resource growth. "Our technical team is rapidly advancing preparations for our maiden drilling program at Mt Clement, which will be a major milestone in unlocking the Project's full potential." Marquee shares are trading at 1.1 cents. Join the discussion: See what HotCopper users are saying about Marquee Resources and be part of the conversations that move the markets. The material provided in this article is for information only and should not be treated as investment advice. Viewers are encouraged to conduct their own research and consult with a certified financial advisor before making any investment decisions. For full disclaimer information, please click here.\n	Marquee Resources Ltd (ASX:MQR) has found multiple zones hosting antimony at its Mt Clement project in Western Australia, based on early-stage exploration work including rock chip sampling and geological mapping. Samples collected during fieldwork in November 2024 were subjected to analysis which suggested extensions to Black Cat's (ASX:BC8) Eastern Hills antimony-lead deposit could be delineated with further work. The second – and final – batch of results from this work showed multiple prospective structures around the Taipan structural trend, with assays including 9,140ppm Sb (antimony) and 7,070ppm Pb (lead); 7,900ppm Sb and 2.34% Pb; and 5,840ppm Sb and 6,810ppm Pb. Additionally, Marquee undertook 1,155-point Ultrafine Fraction (UFF) soil sampling, with results boosting the identification of regional exploration targets. Drill planning has already been started; heritage surveys will begin in 2025's first quarter. "The latest round of results has further reinforced our confidence in the significant potential to expand and define additional antimony resources within our tenure at Mt Clement," Marquee executive chairman Charles Thomas said. "The combination of historical drilling data, along with the newly obtained rock chip sample results, has provided us with a highly compelling set of targets. These targets extend along-strike from the known Eastern Hills Deposit and offer promising opportunities for future resource growth. Our technical team is rapidly advancing preparations for our maiden drilling program at Mt Clement, which will be a major milestone in unlocking the Project's full potential."\n

Gambar 3.4. Contoh teks artikel sebelum dan sesudah tahap *preprocessing* dilakukan

B.3 *Summarization*

Tahap selanjutnya adalah melakukan *summarization* terhadap artikel-artikel yang telah diperoleh. Tujuannya adalah untuk menyajikan informasi penting dalam bentuk ringkas namun tetap representatif terhadap isi utama artikel, guna mempermudah analisis dan pemahaman oleh pengguna akhir.

Seluruh model yang digunakan dalam proyek ini diambil dari Hugging Face Model Hub, sebuah platform terbuka yang menyediakan ribuan model *pre-trained* untuk berbagai tugas *Natural Language Processing* (NLP) dan *machine learning*. Dalam implementasinya, proyek ini menggunakan lima model *summarization* berbasis arsitektur *transformer*, yaitu sebagai berikut.

1. `facebook/bart-large-cnn`: model BART yang telah di-*fine-tune* pada dataset CNN/DailyMail dan cocok untuk peringkasan multi-kalimat.
2. `facebook/bart-large-xsum`: varian BART dengan peringkasan satu kalimat utama, cocok untuk gaya *headline*.
3. `human-centered-summarization/financial-summarization-pegasus`: model PEGASUS yang dioptimalkan untuk domain keuangan dan bisnis.
4. `sshleifer/distilbart-cnn-12-6`: model ringan dari BART dengan ukuran lebih kecil namun tetap kompeten.
5. `t5-base`: model serbaguna dengan pendekatan berbasis instruksi.

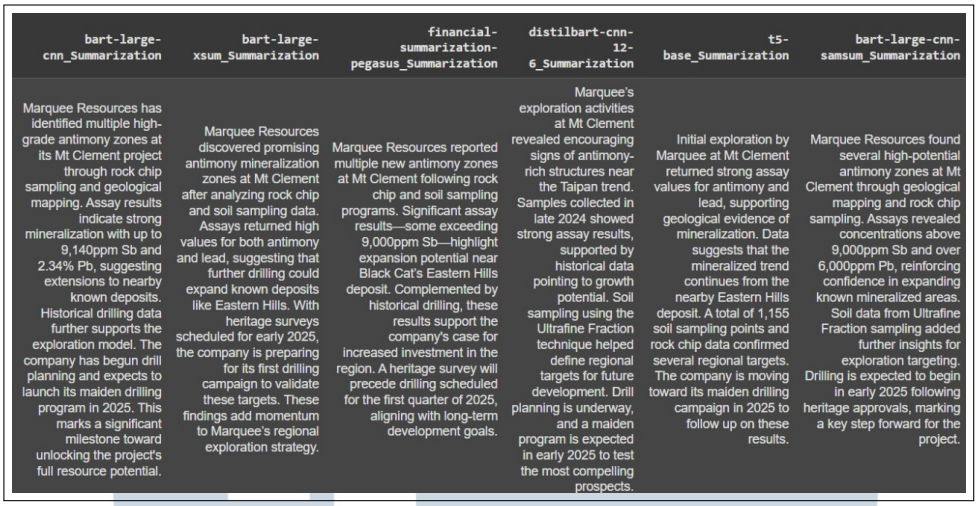
Semua model menggunakan arsitektur *encoder-decoder*, dengan *encoder* membentuk representasi kontekstual dari teks dan *decoder* menghasilkan ringkasan secara *autoregressive* (menghasilkan satu kata per langkah dengan mengandalkan kata-kata sebelumnya sebagai konteks). Beragam model ini digunakan untuk membandingkan kualitas ringkasan berdasarkan akurasi, kelengkapan informasi, dan efisiensi komputasi.

Untuk mengevaluasi performa model secara objektif, digunakan dua metrik utama.

1. ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*): metrik evaluasi ringkasan yang mengukur kemiripan antara ringkasan model dan teks referensi berdasarkan kesamaan kata atau *n-gram*. Umumnya digunakan ROUGE-1 (kata), ROUGE-2 (dua kata), dan ROUGE-L (urutan kata terpanjang).

2. BERTScore: metrik yang menilai kesamaan makna antara ringkasan dan referensi dengan membandingkan representasi vektor (*embedding*) dari setiap kata dalam kedua teks. *Embedding* ini dihasilkan oleh model BERT dan mencerminkan makna semantik dari kata-kata tersebut. Skor *precision* menunjukkan seberapa baik kata-kata dalam ringkasan model cocok secara makna dengan referensi, *recall* menunjukkan seberapa lengkap cakupan makna dari referensi, dan *F1* mencerminkan keseimbangan antara keduanya.

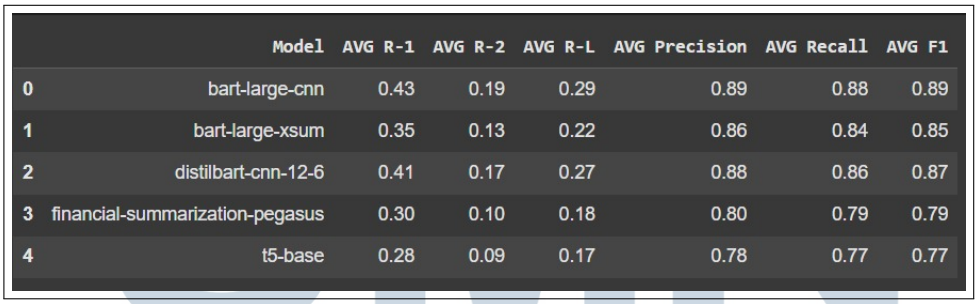
Gambar 3.5 memperlihatkan contoh perbandingan hasil ringkasan yang dihasilkan oleh masing-masing model, yang menunjukkan kelebihan model `facebook/bart-large-cnn` dalam mempertahankan informasi penting sekaligus menjaga keterbacaan ringkasan.



bart-large-cnn_Summarization	bart-large-xsum_Summarization	financial-summarization-pegasus_Summarization	distilbart-cnn-12-6_Summarization	t5-base_Summarization	bart-large-cnn-samsum_Summarization
Marquee Resources has identified multiple high-grade antimony zones at its Mt Clement project through rock chip sampling and geological mapping. Assay results indicate strong mineralization with up to 9,140ppm Sb and 2.34% Pb, suggesting extensions to nearby known deposits. Historical drilling data further supports the exploration model. The company has begun drill planning and expects to launch its maiden drilling program in 2025. This marks a significant milestone toward unlocking the project's full resource potential.	Marquee Resources discovered promising antimony mineralization zones at Mt Clement after analyzing rock chip and soil sampling data. Assays returned high values for both antimony and lead, suggesting that further drilling could expand known deposits like Eastern Hills. With heritage surveys scheduled for early 2025, the company is preparing for its first drilling campaign to validate these targets. These findings add momentum to Marquee's regional exploration strategy.	Marquee Resources reported multiple new antimony zones at Mt Clement following rock chip and soil sampling programs. Significant assay results—some exceeding 9,000ppm Sb—highlight expansion potential near Black Cat's Eastern Hills deposit. Complemented by historical drilling, these results support the company's case for increased investment in the region. A heritage survey will precede drilling scheduled for the first quarter of 2025, aligning with long-term development goals.	Marquee's exploration activities at Mt Clement revealed encouraging signs of antimony-rich structures near the Taipan trend. Samples collected in late 2024 showed strong assay results, supported by historical data pointing to growth potential. Soil sampling using the Ultrafine Fraction technique helped define regional targets for future development. Drill planning is underway, and a maiden program is expected in early 2025 to test the most compelling prospects.	Initial exploration by Marquee at Mt Clement returned strong assay values for antimony and lead, supporting geological evidence of mineralization. Data suggests that the mineralized trend continues from the nearby Eastern Hills deposit. A total of 1,155 soil sampling points and rock chip data confirmed several regional targets. The company is moving toward its maiden drilling campaign in 2025 to follow up on these results.	Marquee Resources found several high-potential antimony zones at Mt Clement through geological mapping and rock chip sampling. Assays revealed concentrations above 9,000ppm Sb and over 6,000ppm Pb, reinforcing confidence in expanding known mineralized areas. Soil data from Ultrafine Fraction sampling added further insights for exploration targeting. Drilling is expected to begin in early 2025 following heritage approvals, marking a key step forward for the project.

Gambar 3.5. Contoh perbandingan hasil *summarization* dari beberapa model pada artikel yang sama

Dengan menggunakan dua metrik yang saling melengkapi ini, yaitu ROUGE untuk mengevaluasi kesamaan tekstual secara langsung dan BERTScore untuk mengevaluasi kesamaan makna, proses evaluasi menjadi lebih menyeluruh. Hal ini sangat penting terutama dalam tugas *abstractive summarization*, di mana model sering kali menulis ulang informasi dengan gaya bahasanya sendiri. Gambar 3.6 menyajikan ringkasan hasil evaluasi kuantitatif untuk kelima model.



	Model	AVG R-1	AVG R-2	AVG R-L	AVG Precision	AVG Recall	AVG F1
0	bart-large-cnn	0.43	0.19	0.29	0.89	0.88	0.89
1	bart-large-xsum	0.35	0.13	0.22	0.86	0.84	0.85
2	distilbart-cnn-12-6	0.41	0.17	0.27	0.88	0.86	0.87
3	financial-summarization-pegasus	0.30	0.10	0.18	0.80	0.79	0.79
4	t5-base	0.28	0.09	0.17	0.78	0.77	0.77

Gambar 3.6. Ringkasan evaluasi kuantitatif menggunakan metrik ROUGE dan BERTScore

Berdasarkan hasil evaluasi kuantitatif, model `facebook/bart-large-cnn` secara konsisten memperoleh skor tertinggi pada metrik ROUGE dan BERTScore. Rata-rata skor ROUGE-1, ROUGE-2, dan ROUGE-L masing-masing mencapai 0.43, 0.19, dan 0.29, sementara skor BERTScore mencapai rata-rata nilai *precision* sebesar 0.89, *recall* sebesar 0.88, dan *F1* sebesar 0.89. Oleh karena itu, model ini dipilih sebagai model utama untuk tugas *summarization*. Selain keunggulan pada metrik evaluasi, ringkasan yang dihasilkan oleh model ini juga telah melalui proses validasi manual oleh *supervisor* perusahaan dan dinilai paling representatif terhadap isi artikel serta paling relevan untuk tujuan analisis bisnis.

B.4 *Sentiment Analysis*

Tahapan selanjutnya dalam *pipeline* adalah *sentiment analysis* terhadap isi lengkap artikel berita. Tujuannya adalah mengidentifikasi nada atau persepsi umum dari setiap artikel terkait sektor pertambangan, apakah bersifat positif, netral, atau negatif. Informasi ini digunakan sebagai referensi tambahan dalam analisis tren dan penyusunan rekomendasi strategis.

Adapun model-model sentimen yang digunakan adalah sebagai berikut.

1. `ProsusAI/finbert`: model berbasis BERT yang dikembangkan khusus untuk sektor finansial dan saham, mampu menangkap sentimen dari laporan keuangan, berita ekonomi, dan teks pasar modal.
2. `soleimanian/financial-roberta-large-sentiment`: varian RoBERTa berukuran besar yang dilatih pada korpus berita bisnis, unggul dalam menangkap konteks kalimat panjang secara mendalam.
3. `mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis`: versi ringan dari RoBERTa untuk *inference* yang lebih cepat dengan hasil yang tetap kompetitif di domain berita keuangan.
4. `ahmedrachid/FinancialBERT-Sentiment-Analysis`: model BERT yang diadaptasi untuk mengenali sentimen dalam *headline* dan konten media ekonomi.
5. `mrm8488/deberta-v3-ft-financial-news-sentiment-analysis`: model DeBERTa versi terbaru yang memiliki keunggulan dalam menangkap *linguistic dependency* dan struktur kalimat kompleks.

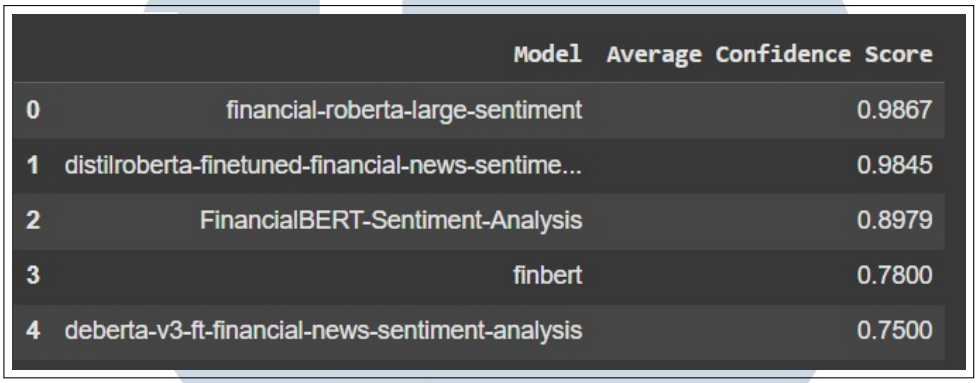
Setiap artikel berita dianalisis oleh kelima model tersebut. Masing-masing model mengembalikan hasil dalam bentuk label sentimen dan nilai *confidence score*, misalnya.

```
{"label": "negative", "score": 0.9987}
```

Untuk menentukan sentimen akhir dari sebuah artikel, dipilih prediksi dengan *confidence score* tertinggi dari seluruh hasil model. Pendekatan ini bersifat deterministik dan menghindari konflik antar model. Selama terdapat satu model dengan *confidence score* yang secara signifikan lebih tinggi dibandingkan model

lainnya, prediksinya akan digunakan. Metode ini juga memberikan fleksibilitas karena memungkinkan model yang lebih relevan secara domain untuk mendominasi keputusan akhir jika menunjukkan keyakinan yang kuat.

Kualitas masing-masing model dievaluasi menggunakan rata-rata *confidence score* hasil prediksi pada seluruh artikel. Gambar 3.7 menyajikan perbandingan hasil evaluasi untuk seluruh model yang diuji.



	Model	Average Confidence Score
0	financial-roberta-large-sentiment	0.9867
1	distilroberta-finetuned-financial-news-sentime...	0.9845
2	FinancialBERT-Sentiment-Analysis	0.8979
3	finbert	0.7800
4	deberta-v3-ft-financial-news-sentiment-analysis	0.7500

Gambar 3.7. Rata-rata *confidence score* dari seluruh model *sentiment analysis* yang diuji pada dataset artikel

Berdasarkan rata-rata *confidence score* tertinggi, model `soleimanian/financial-roberta-large-sentiment` dipilih sebagai model utama dalam *sentiment analysis*. Model ini secara konsisten menghasilkan prediksi dengan tingkat keyakinan yang tinggi terhadap setiap artikel berita keuangan, sehingga dianggap paling andal dalam konteks proyek ini.

B.5 *Risk and Opportunity Inference*

Langkah selanjutnya adalah melakukan analisis risiko dan peluang berbasis LLM. Pada tahap ini, ringkasan berita yang telah dihasilkan pada proses sebelumnya digunakan sebagai *prompt* untuk *instruction model*, yaitu jenis model *LLM* yang dirancang untuk menjawab perintah dalam bentuk bahasa alami.

Model-model yang digunakan seluruhnya diambil dari Hugging Face Model Hub, khususnya dari proyek Unsloth, yaitu sebuah *framework* yang mempermudah *fine-tuning* dan *inference* untuk LLM berskala besar dalam format efisien seperti *4-bit quantization*. Model-model berikut digunakan dalam eksperimen ini.

1. `unsloth/mistral-7b-instruct-v0.3-bnb-4bit`: model Mistral 7B versi *instruction-tuned* yang telah dioptimalkan menggunakan teknik kuantisasi 4-bit untuk mempercepat *inference* dan menghemat memori GPU. Model ini

dilatih agar mampu memahami konteks ringkasan dan menjawab perintah eksplisit seperti "*What are the risks and opportunities?*"

2. `unsloth/llama-3-8b-Instruct-bnb-4bit`: model LLaMA 3 ukuran 8 miliar parameter yang telah dilatih dengan data instruksional dan dioptimalkan dalam format 4-bit. Model ini merupakan salah satu generasi terbaru dari keluarga LLaMA (Meta AI), dengan peningkatan arsitektur dan efisiensi konteks yang lebih luas dibanding versi sebelumnya.

3. `unsloth/mistral-7b-v0.3-bnb-4bit`: model Mistral 7B versi *non-instruct* (generatif umum), digunakan sebagai pembanding untuk mengetahui sejauh mana model yang tidak dilatih secara instruksional mampu menangkap perintah eksplisit. Meskipun tidak dirancang untuk menjawab perintah seperti "Answer:", model ini tetap memiliki kemampuan generatif yang kuat.

4. `unsloth/llama-3-8b-bnb-4bit`: model LLaMA 3 8B versi generatif standar (*non-instruct*) yang digunakan sebagai *baseline* tambahan. Evaluasi terhadap model ini memungkinkan analisis terhadap pengaruh *instruction fine-tuning* terhadap kualitas respons.

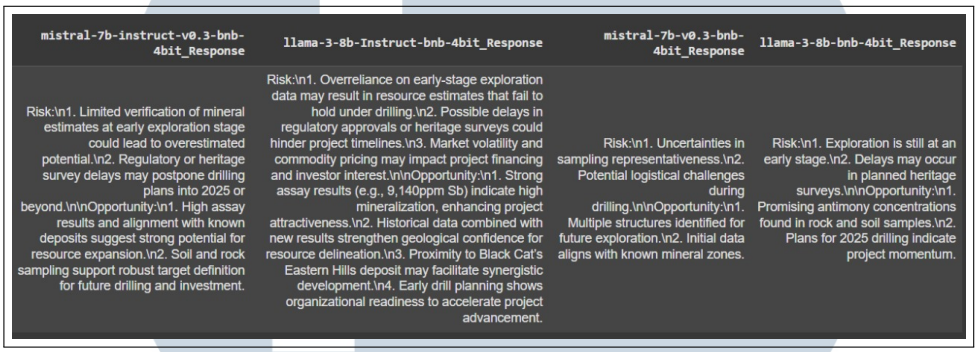
Analisis dilakukan dengan menggunakan *prompt* berikut.

```
You are a financial analyst. Based on the
following financial news summary: {data},
please let me know what the risk & opportunity
are.
```

Berbeda dengan tahap-tahap sebelumnya yang dievaluasi menggunakan metrik kuantitatif, hasil analisis risiko dan peluang dinilai secara kualitatif melalui peninjauan manual oleh *supervisor* perusahaan yang memiliki pemahaman mendalam terhadap konteks bisnis dan strategi industri. Setiap respons yang dihasilkan oleh model dibaca dan dibandingkan untuk menilai apakah jawabannya relevan terhadap isi ringkasan, apakah model mampu membedakan secara jelas antara risiko dan peluang, serta sejauh mana jawabannya bersifat informatif, spesifik, dan dapat ditindaklanjuti.

Berdasarkan penilaian tersebut, model `unsloth/llama-3-8b-Instruct-bnb-4bit` dipilih sebagai model terbaik karena secara konsisten menghasilkan respons yang koheren, relevan, dan bernuansa strategis. Model ini juga menunjukkan pemahaman yang lebih tajam

terhadap konteks industri tambang dibandingkan model lainnya, serta mampu mengidentifikasi poin-poin kunci secara efisien dan eksplisit. Gambar 3.8 memperlihatkan perbandingan hasil analisis risiko dan peluang dari beberapa model terhadap satu artikel yang sama, yang menunjukkan keunggulan model `unsloth/llama-3-8b-Instruct-bnb-4bit` dalam menyajikan respons yang lebih terstruktur dan kontekstual.



mistral-7b-instruct-v0.3-bnb-4bit_Response	llama-3-8b-Instruct-bnb-4bit_Response	mistral-7b-v0.3-bnb-4bit_Response	llama-3-8b-bnb-4bit_Response
Risk:\n1. Limited verification of mineral estimates at early exploration stage could lead to overestimated potential.\n2. Regulatory or heritage survey delays may postpone drilling plans into 2025 or beyond.\n\nOpportunity:\n1. High assay results and alignment with known deposits suggest strong potential for resource expansion.\n2. Soil and rock sampling support robust target definition for future drilling and investment.	Risk:\n1. Overreliance on early-stage exploration data may result in resource estimates that fail to hold under drilling.\n2. Possible delays in regulatory approvals or heritage surveys could hinder project timelines.\n3. Market volatility and commodity pricing may impact project financing and investor interest.\n\nOpportunity:\n1. Strong assay results (e.g., 9,140ppm Sb) indicate high mineralization, enhancing project attractiveness.\n2. Historical data combined with new results strengthen geological confidence for resource delineation.\n3. Proximity to Black Cat's Eastern Hills deposit may facilitate synergistic development.\n4. Early drill planning shows organizational readiness to accelerate project advancement.	Risk:\n1. Uncertainties in sampling representativeness.\n2. Potential logistical challenges during drilling.\n\nOpportunity:\n1. Multiple structures identified for future exploration.\n2. Initial data aligns with known mineral zones.	Risk:\n1. Exploration is still at an early stage.\n2. Delays may occur in planned heritage surveys.\n\nOpportunity:\n1. Promising antimony concentrations found in rock and soil samples.\n2. Plans for 2025 drilling indicate project momentum.

Gambar 3.8. Contoh hasil inferensi model-model LLM dalam mengidentifikasi risiko dan peluang dari satu artikel yang sama

B.6 Storage

Tahap terakhir dari *pipeline* ini adalah penyimpanan hasil akhir dari model-model terbaik yang telah dievaluasi. Hanya *output* dari model terbaik pada masing-masing tahap yang disimpan, yaitu sebagai berikut.

1. *Summarization* dari model `facebook/bart-large-cnn`.
2. *Sentiment analysis* dari model `soleimanian/financial-roberta-large-sentiment`.
3. Respons analisis risiko dan peluang dari model `unsloth/llama-3-8b-Instruct-bnb-4bit`.

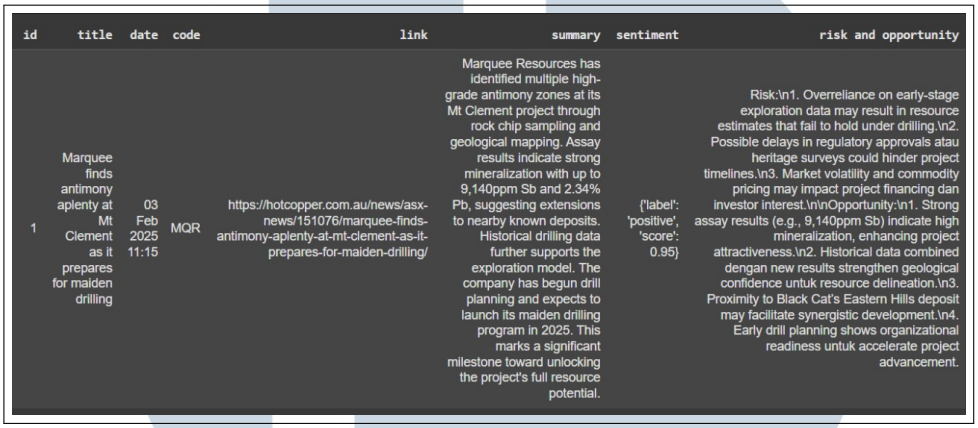
Seluruh hasil ini disimpan dalam satu file `CSV` terintegrasi, dengan struktur kolom sebagai berikut.

```
[id, title, date, code, link, summary, sentiment, risk
and opportunity]
```

File CSV tersebut kemudian diunggah ke Google Drive, sehingga dapat diakses secara mudah dan aman oleh tim strategi perusahaan. Pendekatan ini memastikan hasil analisis yang paling relevan telah terdokumentasi dan

tersimpan secara terpusat untuk digunakan kembali dalam analisis lanjutan maupun pengambilan keputusan bisnis.

Gambar 3.9 menyajikan contoh tampilan hasil penyimpanan dalam format CSV setelah seluruh proses analitik selesai dijalankan.



id	title	date	code	link	summary	sentiment	risk and opportunity
1	Marquee finds antimony aplenty at Mt Clement as it prepares for maiden drilling	03 Feb 2025 11:15	MQR	https://hotcopper.com.au/news/asx-news/151076/marquee-finds-antimony-aplenty-at-mt-clement-as-it-prepares-for-maiden-drilling/	Marquee Resources has identified multiple high-grade antimony zones at its Mt Clement project through rock chip sampling and geological mapping. Assay results indicate strong mineralization with up to 9,140ppm Sb and 2.34% Pb, suggesting extensions to nearby known deposits. Historical drilling data further supports the exploration model. The company has begun drill planning and expects to launch its maiden drilling program in 2025. This marks a significant milestone toward unlocking the project's full resource potential.	{'label': 'positive', 'score': 0.95}	Risk:\n1. Overreliance on early-stage exploration data may result in resource estimates that fail to hold under drilling.\n2. Possible delays in regulatory approvals atau heritage surveys could hinder project timelines.\n3. Market volatility and commodity pricing may impact project financing dan investor interest.\n\nOpportunity:\n1. Strong assay results (e.g., 9,140ppm Sb) indicate high mineralization, enhancing project attractiveness.\n2. Historical data combined dengan new results strengthen geological confidence untuk resource delineation.\n3. Proximity to Black Cat's Eastern Hills deposit may facilitate synergistic development.\n4. Early drill planning shows organizational readiness untuk accelerate project advancement.

Gambar 3.9. Contoh hasil akhir dalam format CSV yang berisi *summarization*, *sentiment analysis*, dan analisis risiko-peluang dari model terbaik

3.3.2 Prediksi Harga Komoditas Menggunakan Model *Machine Learning* Berbasis Data *Time Series*

A *User Requirements*

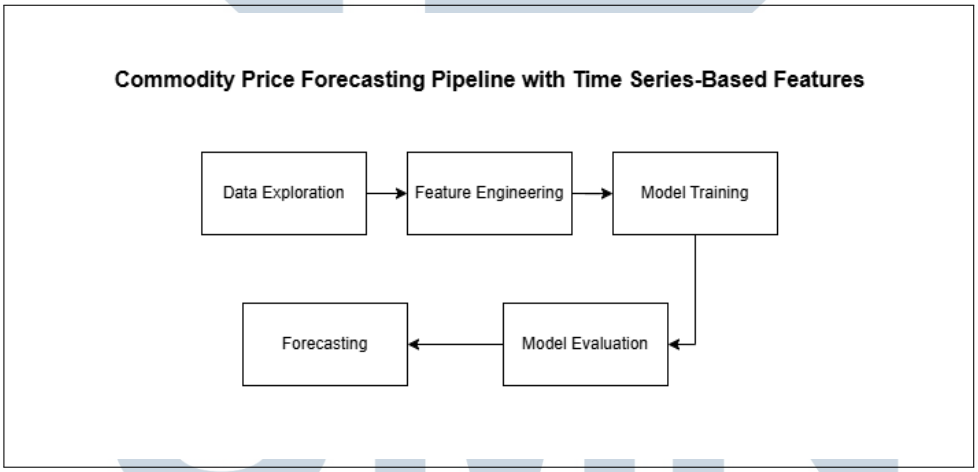
Sejalan dengan kebutuhan strategis perusahaan dalam memprediksi harga komoditas, *stakeholder* menetapkan serangkaian spesifikasi teknis yang harus dipenuhi oleh sistem yang dikembangkan. Sistem ini diharapkan mampu melakukan prediksi harga penutupan komoditas seperti tembaga dalam rentang waktu menengah hingga panjang, dengan cakupan *horizon* prediksi mulai dari 6 bulan hingga beberapa tahun ke depan. Prediksi jangka menengah didefinisikan sebagai prediksi dengan rentang waktu antara 6 hingga 12 bulan ke depan, yang penting untuk mengantisipasi perubahan harga yang mengikuti pola musiman atau tren tahunan, seperti peningkatan permintaan pada kuartal tertentu atau dampak kebijakan fiskal tahunan. Sementara itu, prediksi jangka panjang mencakup waktu lebih dari satu tahun, yang bertujuan mendukung keputusan strategis perusahaan dalam investasi maupun manajemen risiko yang bersifat jangka panjang. Model yang digunakan harus mampu mengidentifikasi pola jangka panjang, siklus musiman, serta dinamika fluktuasi mingguan yang terekam dalam data historis.

Selain ketepatan estimasi, sistem juga dituntut memiliki aspek

interpretabilitas yang tinggi. Oleh karena itu, diperlukan visualisasi eksploratif terhadap data dan hasil prediksi, serta pemetaan fitur-fitur yang paling berpengaruh terhadap *output* model. Proses pengolahan data harus mendukung *input* berbasis *time series*, serta mampu mengakomodasi berbagai pendekatan prediksi, baik yang bersifat statistik maupun berbasis *machine learning*.

Sistem juga diharapkan fleksibel dalam hal frekuensi data (harian, mingguan, atau bulanan), serta mampu melakukan *training* dan *testing* pada skenario *out-of-sample* yang realistis. Untuk mendukung pengambilan keputusan strategis, hasil akhir dari sistem perlu divisualisasikan secara kronologis dalam bentuk grafik terpadu yang menyajikan perbandingan antara harga aktual dan prediksi dari berbagai model, termasuk varian *hybrid* yang memisahkan komponen tren dan residual.

Pipeline umum dari sistem yang dikembangkan ditunjukkan pada Gambar 3.10, yang menggambarkan alur proses mulai dari eksplorasi data, *feature engineering*, *training* model, evaluasi, hingga tahap prediksi akhir.



Gambar 3.10. *Pipeline* untuk memprediksi harga komoditas dengan memanfaatkan fitur berbasis *time series*

B Implementasi

B.1 *Data Exploration*

Eksplorasi dilakukan terhadap data historis harga komoditas yang berasal dari data *internal* perusahaan dan awalnya bersifat harian, kemudian diubah menjadi basis mingguan melalui proses *resampling*. Frekuensi mingguan dipilih karena mampu menangkap pola fluktuasi jangka menengah secara seimbang. Data dengan

frekuensi harian cenderung memiliki banyak fluktuasi yang bersifat acak atau tidak signifikan (*noise*), sehingga dapat mengaburkan pola tren yang sesungguhnya dan menyulitkan analisis jangka panjang. Sebaliknya, data dengan frekuensi bulanan terlalu lambat dalam menangkap perubahan harga yang dinamis, sehingga beberapa informasi penting terkait volatilitas mingguan atau respons harga terhadap suatu peristiwa dapat hilang. Oleh karena itu, frekuensi mingguan dianggap sebagai pilihan optimal yang dapat menyeimbangkan antara detail informasi dan kestabilan tren. Meski demikian, sistem ini tetap fleksibel untuk dianalisis dalam frekuensi harian maupun bulanan tergantung pada kebutuhan analisis.

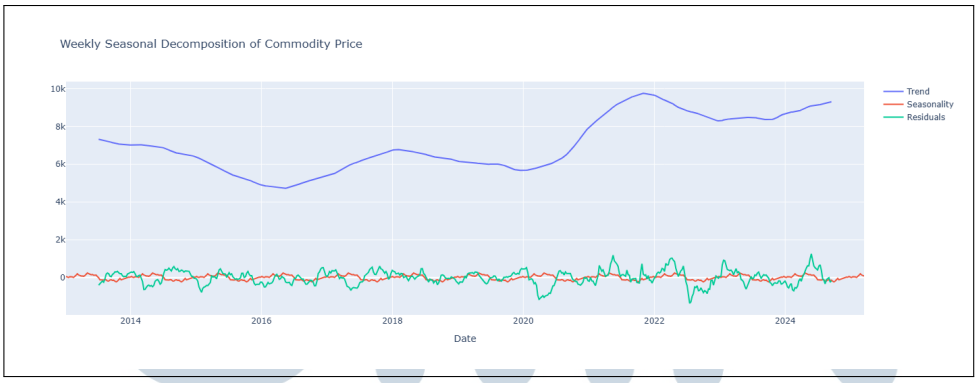
Visualisasi awal dilakukan untuk memahami dinamika harga dari waktu ke waktu. Gambar 3.11 memperlihatkan tren harga mingguan komoditas yang fluktuatif dengan beberapa fase naik-turun yang signifikan, terutama lonjakan harga setelah tahun 2020. Plot ini berguna untuk mengidentifikasi periode-periode penting dalam pergerakan harga, serta mengamati adanya pola atau anomali yang mencolok.

Gambar 3.11. Pergerakan harga komoditas mingguan dari waktu ke waktu

Gambar 3.12 menunjukkan distribusi harga mingguan komoditas dalam bentuk histogram. Distribusinya tampak tidak simetris dan cenderung *right-skewed*, yang mengindikasikan bahwa harga lebih sering berada pada rentang rendah namun tetap memiliki peluang untuk mengalami lonjakan ekstrem. Visualisasi ini membantu dalam mengidentifikasi kemungkinan *outlier* serta memperkirakan rentang harga yang paling umum muncul selama periode historis.

Gambar 3.12. Distribusi harga mingguan komoditas

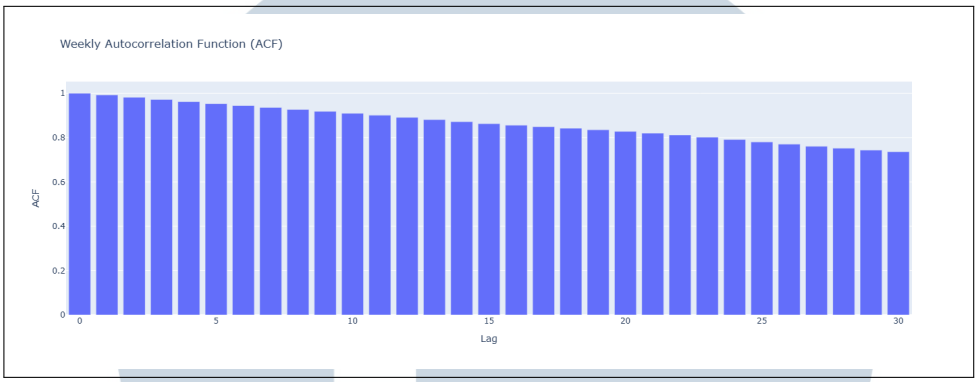
Untuk menggali struktur *internal* dari data harga, dilakukan dekomposisi musiman terhadap sinyal harga komoditas. Proses ini memisahkan tiga komponen utama: tren jangka panjang, pola musiman, dan fluktuasi residual. Gambar 3.13 menunjukkan bahwa tren harga cenderung naik setelah tahun 2020, sedangkan pola musiman tetap relatif konsisten dan fluktuasi residual mengandung variasi yang tidak dijelaskan oleh dua komponen sebelumnya. Informasi ini menjadi dasar dalam pemodelan *hybrid* yang memisahkan sinyal deterministik dan stokastik secara eksplisit.



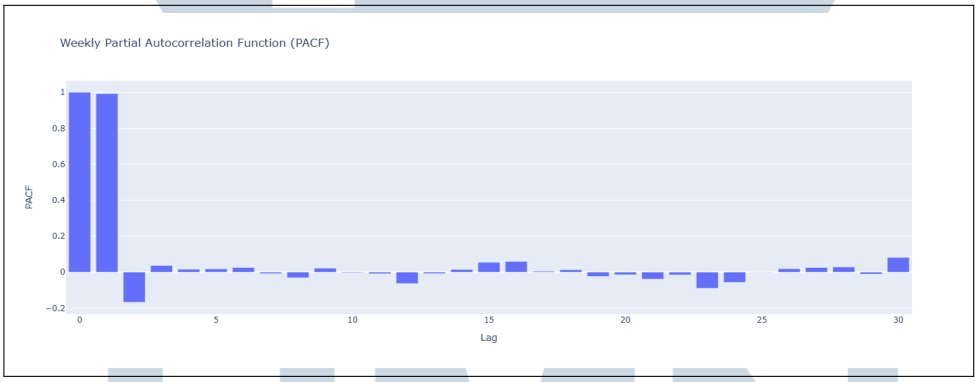
Gambar 3.13. Dekomposisi musiman terhadap harga mingguan komoditas

Selain itu, untuk mengevaluasi ketergantungan temporal dalam data, digunakan fungsi Autocorrelation (ACF) dan Partial Autocorrelation (PACF). ACF mengukur sejauh mana harga saat ini berkorelasi dengan nilai pada waktu sebelumnya dalam berbagai *lag*, sementara PACF memperlihatkan korelasi langsung antara titik saat ini dan titik di masa lalu dengan mengeliminasi pengaruh *intermediate lag*. Gambar 3.14 menunjukkan bahwa harga memiliki korelasi jangka panjang yang cukup kuat dan menurun secara perlahan, menunjukkan adanya struktur serial yang stabil. Sedangkan pada Gambar 3.15, terlihat bahwa *lag* 1 dan 2

memberikan kontribusi korelasi langsung yang paling signifikan terhadap harga saat ini. Hasil ini penting dalam perancangan fitur *lag* yang digunakan dalam pemodelan prediktif.



Gambar 3.14. Plot ACF terhadap harga mingguan komoditas



Gambar 3.15. Plot PACF terhadap harga mingguan komoditas

Keseluruhan eksplorasi ini memperlihatkan bahwa harga komoditas tidak hanya dipengaruhi oleh faktor acak, tetapi juga memiliki struktur musiman, tren jangka panjang, dan keterkaitan temporal yang kuat antar periode waktu. Temuan ini menjadi dasar penting dalam pengembangan model prediktif berbasis *time series* dan penentuan strategi fitur yang lebih informatif.

B.2 *Feature Engineering*

Proses *feature engineering* dilakukan secara menyeluruh untuk mengubah sinyal harga komoditas mentah menjadi representasi numerik yang kaya informasi dan siap digunakan dalam model prediktif berbasis *time series*. Teknik-teknik yang digunakan mencakup transformasi statistikal, indikator teknikal, analisis musiman, serta ekstraksi pola nonlinier.

Langkah awal dilakukan dengan menambahkan fitur dasar seperti *price difference*, *percentage return*, dan *log return* yang merepresentasikan perubahan harga absolut dan relatif dari satu periode ke periode berikutnya, berfungsi sebagai indikator dasar untuk mendeteksi momentum jangka pendek. Selanjutnya diterapkan *lag features*, yaitu nilai harga dan *return* pada periode sebelumnya, yang penting untuk menangkap dependensi temporal dalam data. Beberapa *lag* digunakan, seperti *lag* 1, *lag* 2, dan seterusnya, disesuaikan dengan hasil analisis PACF sebelumnya.

Berbagai indikator teknikal ditambahkan untuk memperkaya representasi data, antara lain *Average True Range* (ATR) sebagai ukuran volatilitas harga, *Bollinger Bands* untuk mengukur deviasi harga dari rata-rata bergeraknya, *Relative Strength Index* (RSI) untuk mendeteksi kondisi jenuh beli atau jenuh jual, serta *Moving Average Convergence Divergence* (MACD) dan komponennya seperti *signal*, histogram, dan *divergence* untuk mengidentifikasi momentum dan potensi perubahan arah tren.

Transformasi lanjutan dilakukan melalui fitur statistik *rolling* seperti *mean*, *standard deviation*, *minimum*, *maximum*, *median*, *skewness*, dan *kurtosis* dalam jendela waktu tertentu, yang membantu menangkap karakter distribusi lokal harga. Fitur lain seperti *price acceleration*, *rate of change* (ROC), *Average Directional Index* (ADX), *price-to-SMA ratio*, dan *volatility ratio* ditambahkan untuk mengukur kekuatan dan percepatan tren harga.

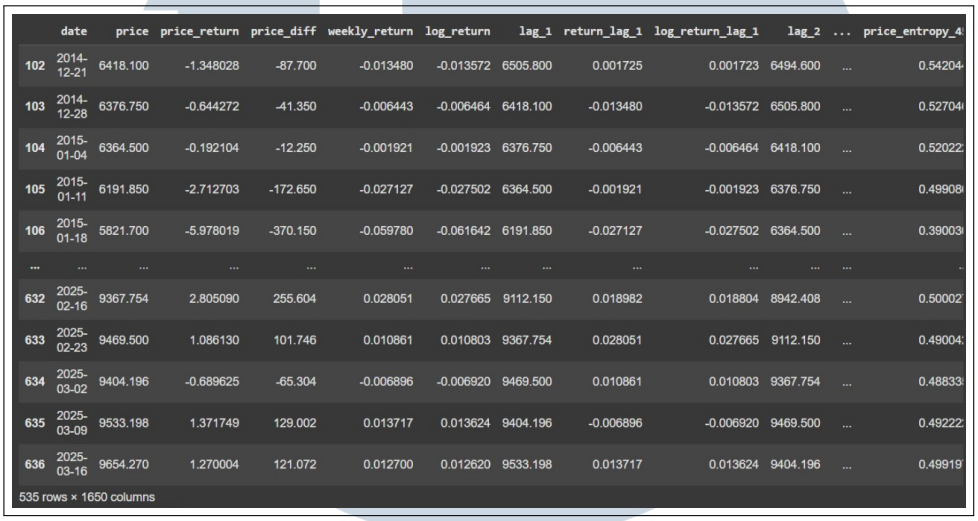
Metode *Exponential Weighted Moving Average* (EWMA) juga diterapkan sebagai alternatif dari *moving average* biasa, disertai fitur turunan seperti *price–EWMA distance*, EWMA *volatility*, dan sinyal *crossover* antar EWMA untuk mendeteksi potensi titik balik tren.

Deteksi pola lokal dilakukan dengan mengidentifikasi *local minima* dan *maxima* dalam jendela tertentu serta pola *zigzag* berdasarkan persentase pembalikan harga, yang berguna dalam mengenali pola teknikal umum. Selain itu, fitur berbasis kalender seperti *month*, *quarter*, *week of year*, serta indikator akhir bulan, kuartal, dan tahun turut ditambahkan. Untuk memperkuat representasi musiman, digunakan transformasi siklikal sin dan cos dari bulan dan minggu agar bersifat kontinu.

Sinyal harga kemudian didekomposisi menjadi tiga komponen musiman: tren, musiman, dan residual. Dari hasil dekomposisi ini, diturunkan fitur tambahan seperti *trend strength*, *trend change*, *seasonality strength*, *residual volatility*, dan *trend acceleration*, yang mampu menangkap dinamika tersembunyi dari harga mentah. Sebagai pelengkap, ditambahkan fitur berbasis *entropy* untuk mengukur

kompleksitas dan keteraturan lokal sinyal harga, serta interaksi antar komponen tren dan volatilitas.

Gambar 3.16 menampilkan cuplikan *dataframe* hasil akhir setelah seluruh tahapan *feature engineering*, di mana setiap baris merepresentasikan satu observasi mingguan yang telah dilengkapi ribuan fitur turunan dan siap digunakan dalam proses *training* model.



	date	price	price_return	price_diff	weekly_return	log_return	lag_1	return_lag_1	log_return_lag_1	lag_2	...	price_entropy_4
102	2014-12-21	6418.100	-1.348028	-87.700	-0.013480	-0.013572	6505.800	0.001725	0.001723	6494.600	...	0.54204
103	2014-12-28	6376.750	-0.644272	-41.350	-0.006443	-0.006464	6418.100	-0.013480	-0.013572	6505.800	...	0.52704
104	2015-01-04	6364.500	-0.192104	-12.250	-0.001921	-0.001923	6376.750	-0.006443	-0.006464	6418.100	...	0.52022
105	2015-01-11	6191.850	-2.712703	-172.650	-0.027127	-0.027502	6364.500	-0.001921	-0.001923	6376.750	...	0.49908
106	2015-01-18	5821.700	-5.978019	-370.150	-0.059780	-0.061642	6191.850	-0.027127	-0.027502	6364.500	...	0.39003
...	...	...	...	...	...	...	...	...	...	...	...	...
632	2025-02-16	9367.754	2.805090	255.604	0.028051	0.027665	9112.150	0.018982	0.018804	8942.408	...	0.50002
633	2025-02-23	9469.500	1.086130	101.746	0.010861	0.010803	9367.754	0.028051	0.027665	9112.150	...	0.49004
634	2025-03-02	9404.196	-0.689625	-65.304	-0.006896	-0.006920	9469.500	0.010861	0.010803	9367.754	...	0.48833
635	2025-03-09	9533.198	1.371749	129.002	0.013717	0.013624	9404.196	-0.006896	-0.006920	9469.500	...	0.49222
636	2025-03-16	9654.270	1.270004	121.072	0.012700	0.012620	9533.198	0.013717	0.013624	9404.196	...	0.49919

535 rows × 1650 columns

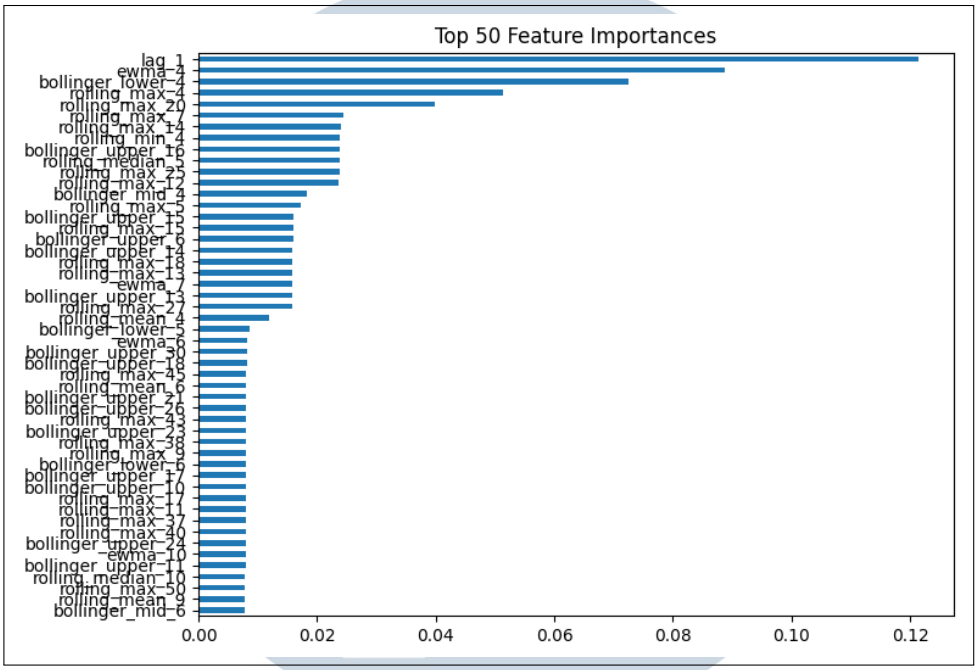
Gambar 3.16. Hasil *dataframe* setelah proses *feature engineering*

Dengan konstruksi fitur yang komprehensif dan sistematis seperti ini, model *machine learning* dapat menangkap berbagai aspek penting dalam sinyal harga, mulai dari dinamika jangka pendek, fluktuasi musiman, pola tren, hingga sinyal pasar yang lebih kompleks secara lebih akurat dan informatif.

B.3 *Model Training*

Setelah seluruh fitur hasil *feature engineering* dimasukkan ke dalam *dataset*, langkah berikutnya adalah melakukan proses seleksi fitur dengan pendekatan *feature importance* berbasis Random Forest. Model Random Forest dilatih pada seluruh fitur, kemudian *feature importance* tiap fitur dihitung untuk mengidentifikasi variabel-variabel yang paling berpengaruh terhadap prediksi harga. Dari ribuan fitur yang dihasilkan, hanya sekitar 50 fitur teratas dengan tingkat kepentingan tertinggi yang dipilih sebagai *input* utama bagi seluruh model prediktif. Langkah ini bertujuan untuk mengurangi kompleksitas model, mempercepat proses *training*, serta meminimalisasi risiko *overfitting*. Visualisasi hasil seleksi fitur

dapat dilihat pada Gambar 3.35, yang menunjukkan *ranking* fitur berdasarkan nilai kepentingan relatifnya terhadap target.



Gambar 3.17. Visualisasi *feature importance* berdasarkan hasil Random Forest

Model-model yang digunakan terdiri dari berbagai pendekatan *machine learning*, *deep learning*, serta model *hybrid* yang mengombinasikan beberapa teknik secara sistematis sesuai kebutuhan aplikasi di industri. Untuk model *machine learning* berbasis pohon keputusan, digunakan XGBoost dan LightGBM. Kedua model ini memanfaatkan teknik *boosting*, yaitu membangun *ensemble* pohon-pohon keputusan secara berurutan, di mana tiap pohon baru memperbaiki kelemahan dari pohon-pohon sebelumnya. XGBoost dikenal sangat efisien secara komputasi dan mampu menangani data dengan *missing values*, sementara LightGBM menawarkan *training* yang lebih cepat serta penggunaan memori yang lebih hemat melalui teknik *histogram-based learning*. Optimasi *hyperparameter* dilakukan dengan *grid search*, sehingga kombinasi parameter yang digunakan dapat menghasilkan performa prediksi yang optimal.

Pada sisi *deep learning*, *pipeline* ini menerapkan beberapa model yang dapat menangkap pola temporal dan dinamis harga komoditas. Model LSTM (Long Short-Term Memory) digunakan untuk mengatasi masalah *vanishing gradient* pada jaringan saraf konvensional sehingga dapat mengingat informasi penting dalam rentang waktu panjang, cocok untuk tren jangka panjang maupun pola musiman. Sementara GRU (Gated Recurrent Unit) adalah varian yang lebih

sederhana dan komputasi lebih ringan, namun tetap efektif dalam *modeling* data sekuensial. Selain itu, diterapkan juga model N-BEATS (Neural Basis Expansion Analysis for Time Series) yang mengusung arsitektur interpretable dan mampu memisahkan komponen *trend* serta *seasonality* secara eksplisit, sehingga prediksi yang dihasilkan tidak hanya akurat tetapi juga lebih mudah diinterpretasikan.

Sebagai pelengkap, *pipeline* prediksi ini juga mengimplementasikan berbagai skema *hybrid* model yang menggabungkan kekuatan beberapa pendekatan secara bersamaan. *Hybrid* model dirancang untuk memaksimalkan akurasi dan stabilitas prediksi dengan memanfaatkan keunggulan model statistik, *machine learning*, dan *deep learning* sekaligus. Berikut beberapa variasi *hybrid* model yang diterapkan.

1. *Hybrid* Linear Trend + XGBoost: Data harga komoditas dipisahkan menjadi komponen tren linier dan residual. Tren diekstraksi secara eksplisit, sementara fluktuasi jangka pendek dan anomali dipelajari menggunakan XGBoost.

2. *Hybrid* STL + LSTM: Data harga di-*decompose* menggunakan STL (*Seasonal-Trend decomposition using Loess*) untuk memisahkan komponen tren dan musiman. Komponen tren diprediksi secara terpisah, sedangkan residualnya dipelajari menggunakan LSTM.

3. *Hybrid* Seasonal Decomposition + XGBoost / LightGBM: *Dekomposisi musiman* menghasilkan komponen tren dan residual, di mana residual selanjutnya dipelajari menggunakan XGBoost atau LightGBM.

4. *Hybrid* Seasonal Decomposition + LSTM: Mirip seperti sebelumnya, namun residual yang dihasilkan dari *dekomposisi musiman* diprediksi menggunakan LSTM.

5. *Hybrid* Linear Regression + XGBoost: Komponen tren dipetakan dengan regresi linier berbasis waktu, sementara residualnya dipelajari lebih lanjut menggunakan XGBoost.

Keunggulan *hybrid* model terletak pada kemampuannya mengintegrasikan prediksi tren makro dengan pola musiman atau fluktuasi harga yang seringkali sulit ditangkap oleh satu model saja. Pendekatan ini secara empiris mampu menghasilkan performa prediksi yang lebih stabil dan akurat, khususnya pada periode terjadi perubahan pola harga atau anomali pasar.

B.4 *Model Evaluation*

Evaluasi performa seluruh model dilakukan secara menyeluruh dengan membandingkan hasil prediksi terhadap harga aktual pada data uji menggunakan tiga metrik utama, yaitu MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*), dan MAPE (*Mean Absolute Percentage Error*). Hasil evaluasi yang ditampilkan pada Gambar 3.18 menunjukkan bahwa model GRU memberikan performa terbaik, dengan MAE sebesar 145,93, RMSE sebesar 224,42, dan MAPE sebesar 1,56%. Model LightGBM dan XGBoost juga menghasilkan performa yang sangat kompetitif, masing-masing dengan MAE sebesar 150,01 dan 162,21, RMSE sebesar 173,16 dan 191,47, serta MAPE sebesar 1,62% dan 1,77%. Model LSTM memperoleh MAE sebesar 274,19, RMSE sebesar 314,50, dan MAPE sebesar 3,01%, sedangkan N-BEATS menunjukkan error yang jauh lebih tinggi dengan MAE sebesar 1121,82, RMSE sebesar 1146,91, dan MAPE sebesar 12,14%.

Model	MAE	RMSE	MAPE
GRU	145.934763	224.419707	1.561528
LightGBM	150.014227	173.162183	1.624606
XGBoost	162.208306	191.468675	1.767210
LSTM	274.185185	314.499706	3.007831
Hybrid_Seasonal_LGB	283.334329	360.685327	3.100667
Hybrid_Seasonal_NN	285.389428	345.945772	3.141222
Hybrid_STL_LSTM	318.972292	370.447644	3.512581
Hybrid_Seasonal_XGB	395.244319	479.661891	4.326974
N-BEATS	1121.821753	1146.913513	12.143053
Hybrid_LR_XGB	1414.134991	1482.112224	15.475297
Hybrid_Linear_XGB	1414.134991	1482.112224	15.475297

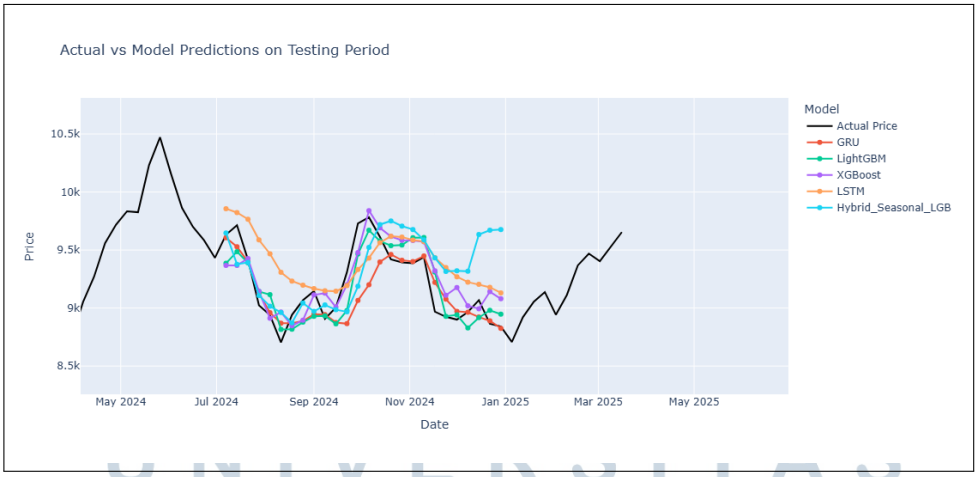
Gambar 3.18. Tabel hasil evaluasi metrik MAE, RMSE, dan MAPE dari seluruh model

Temuan ini mengindikasikan bahwa model-model berbasis GRU, LightGBM, dan XGBoost secara konsisten mampu mengatasi karakteristik data harga komoditas yang bersifat dinamis dan penuh fluktuasi, sehingga menghasilkan prediksi yang lebih presisi. Keunggulan GRU terletak pada kemampuannya dalam menangkap pola sekuensial jangka panjang tanpa kompleksitas berlebih, sementara LightGBM dan XGBoost terbukti efektif dalam memanfaatkan relasi non-linear antar fitur hasil *feature engineering*. Sebaliknya, performa LSTM sedikit tertinggal, kemungkinan akibat pengaturan *hyperparameter* yang belum optimal untuk data ini, sedangkan N-BEATS yang secara arsitektural lebih kompleks justru kurang mampu menyesuaikan diri dengan pola historis harga komoditas.

Hasil ini menegaskan pentingnya pemilihan model yang sesuai dengan karakteristik data, serta menunjukkan bahwa model dengan struktur yang lebih sederhana namun tepat guna dapat memberikan hasil prediksi yang lebih baik dibandingkan model yang lebih kompleks. Nilai-nilai metrik error tersebut menjadi dasar utama dalam proses seleksi model untuk tahap *forecasting* berikutnya.

Selain evaluasi kuantitatif melalui tabel metrik, visualisasi juga digunakan untuk memperjelas perbandingan antara hasil prediksi model dengan harga aktual pada periode *testing*. Visualisasi ini secara khusus menampilkan lima model teratas yang terpilih berdasarkan hasil evaluasi pada metrik MAE, RMSE, dan MAPE. Model-model dengan performa terbaik, yaitu GRU, LightGBM, XGBoost, LSTM, dan Hybrid_Seasonal_LGB, dipilih agar interpretasi plot menjadi lebih jelas dan tidak terlalu padat, meskipun seluruh model yang dikembangkan telah diuji pada data *testing* yang sama.

Pada Gambar 3.19, ditampilkan kurva harga aktual beserta hasil prediksi dari lima model terbaik selama periode *testing* (Juli hingga Desember 2024). Visualisasi ini memberikan gambaran yang lebih intuitif mengenai kinerja setiap model dalam mengikuti pola pergerakan harga riil komoditas pada rentang waktu yang benar-benar berada di luar data *training*.



Gambar 3.19. Kurva perbandingan harga aktual dan prediksi lima model terbaik pada periode *testing* (Juli–Desember 2024)

Dari hasil visualisasi tersebut, terlihat bahwa model GRU, LightGBM, dan XGBoost secara konsisten mampu mengikuti fluktuasi harga aktual dengan deviasi yang relatif kecil. Kurva prediksi yang dihasilkan oleh model-model ini berada cukup dekat dengan garis harga aktual, memperlihatkan kemampuan adaptasi model terhadap pola dinamis yang terjadi pada data riil. Sementara itu, model

LSTM dan Hybrid_Seasonal_LGB juga menunjukkan performa yang kompetitif, meskipun pada beberapa titik terdapat deviasi yang lebih besar dibandingkan model-model terbaik lainnya.

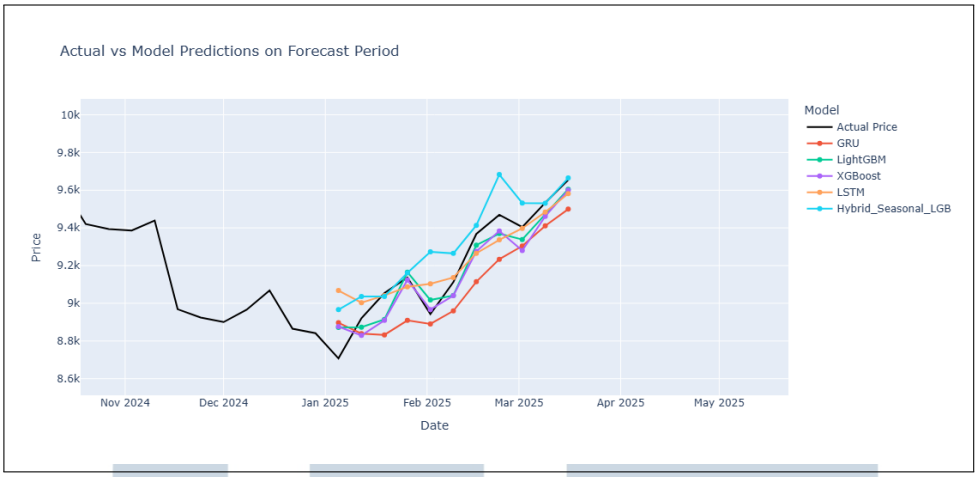
Plot ini tidak hanya menampilkan tingkat kedekatan antara hasil prediksi dengan harga aktual, tetapi juga memperlihatkan bagaimana respons setiap model terhadap perubahan tren mendadak dan lonjakan harga pada periode *testing*. Hal ini sangat penting dalam konteks *forecasting* harga komoditas, karena ketepatan model dalam mengantisipasi perubahan tren dan volatilitas akan sangat menentukan kualitas prediksi di masa mendatang. Dengan demikian, visualisasi ini berperan sebagai alat bantu penting dalam mengevaluasi *robustness* dan *reliability* model ketika diterapkan pada data yang benar-benar baru.

Kombinasi antara evaluasi kuantitatif melalui metrik error dan evaluasi visual pada plot hasil prediksi ini semakin memperkuat temuan bahwa model GRU, LightGBM, dan XGBoost layak dipilih sebagai kandidat utama untuk tahap *forecasting* lanjutan, karena mampu memberikan prediksi yang presisi dan stabil pada data yang belum pernah dilihat sebelumnya.

B.5 *Forecasting*

Setelah tahap evaluasi model pada data *testing*, langkah selanjutnya adalah melakukan *forecasting* harga komoditas untuk periode mendatang menggunakan model-model terbaik yang telah teridentifikasi sebelumnya. Pada proses ini, model GRU, LightGBM, XGBoost, LSTM, dan Hybrid_Seasonal_LGB dilatih ulang (*retrain*) dengan data historis hingga akhir Desember 2024, kemudian digunakan untuk menghasilkan prediksi harga pada enam bulan ke depan, yaitu dari Januari hingga April 2025.

Seluruh model terbaik yang telah dievaluasi pada tahap sebelumnya diikutsertakan dalam proses *forecasting*, namun visualisasi hasil prediksi pada Gambar 3.20 hanya menampilkan lima model teratas agar interpretasi plot lebih jelas dan mudah diamati.



Gambar 3.20. Kurva perbandingan harga aktual dan prediksi lima model terbaik pada periode *forecasting* (Januari–April 2025)

Berdasarkan visualisasi pada Gambar 3.20, dapat diamati bahwa seluruh model utama mampu mengikuti arah tren harga aktual pada awal periode *forecasting*. Model GRU, LightGBM, dan XGBoost menghasilkan kurva prediksi yang cenderung lebih dekat dengan harga aktual, terutama pada fase tren naik di bulan Februari hingga Maret 2025. Model LSTM dan Hybrid_Seasonal_LGB juga memberikan hasil yang kompetitif, meskipun pada beberapa titik menunjukkan deviasi yang sedikit lebih besar terhadap harga riil.

Berdasarkan hasil evaluasi metrik pada tahap sebelumnya maupun performa visualisasi pada periode *forecasting*, model GRU ditetapkan sebagai model terbaik karena secara konsisten menghasilkan prediksi paling presisi dengan deviasi yang paling kecil terhadap harga aktual. Model LightGBM dan XGBoost dapat digunakan sebagai model pembanding (*benchmark*) karena juga menunjukkan performa yang sangat kompetitif.

3.3.3 Membangun Model *Machine Learning* untuk Mengelompokkan Data *Logging* Geofisika dalam Rangka Identifikasi Zona Batubara

A *User Requirements*

Sistem yang dibangun harus mampu mengelompokkan data *logging* geofisika, khususnya *Gamma Ray* (GR), *Caliper* (CL), *Density* (LD), dan *Sonic* (SD), secara otomatis untuk mengidentifikasi zona batubara di berbagai sumur eksplorasi. Setiap fitur *log* memiliki peran penting dalam proses identifikasi. *Gamma Ray* digunakan untuk mengukur radiasi alami batuan. Nilai GR yang

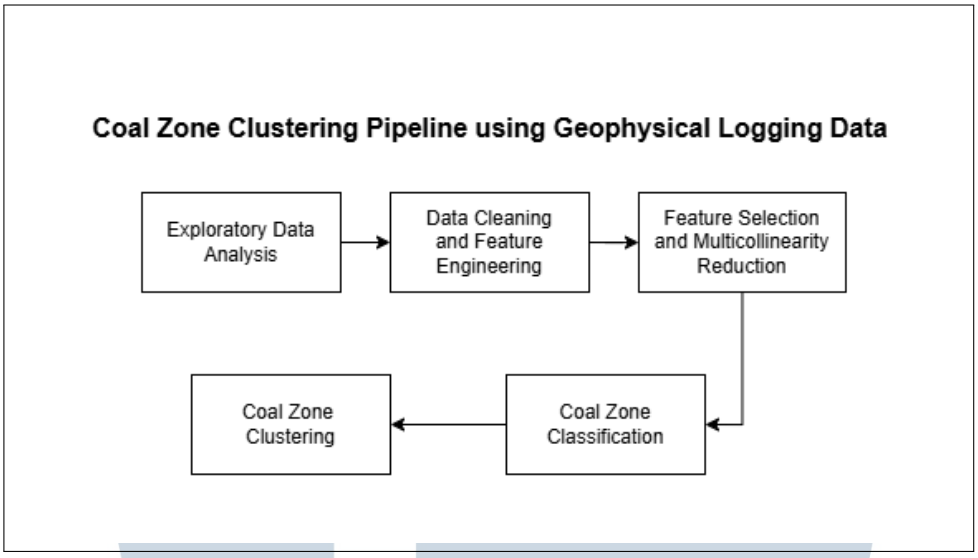
rendah biasanya menandakan keberadaan batubara karena batubara memiliki kandungan radioaktif yang rendah. *Caliper* berfungsi untuk mengukur diameter lubang bor dan sangat bermanfaat untuk mendeteksi *washout*, *collapse*, atau bentuk lubang yang tidak sempurna. *Density* (LD atau *Long Spaced Density*) mengukur densitas batuan, di mana batubara selalu menunjukkan densitas lebih rendah dibandingkan batuan sekitarnya. Sementara *Sonic* (SD), yang kadang disebut juga *Short Spaced Density* atau *Sonic/Neutron Log* tergantung alat yang digunakan, biasanya mengukur kecepatan rambat gelombang suara atau densitas dengan konfigurasi tertentu.

Karakteristik utama zona batubara pada data *log* ditandai oleh kombinasi nilai GR yang rendah, LD yang sangat rendah, SD yang sangat tinggi, serta CL yang sering menunjukkan indikasi washout atau pembesaran lubang. Untuk mendukung identifikasi yang akurat, *pipeline* yang dirancang harus mampu melakukan pembersihan data secara menyeluruh, mulai dari penanganan *missing values*, deteksi dan penghapusan *outlier* geologi, hingga *filtering* multivariat agar validitas hasil tetap terjaga.

Proses *feature engineering* perlu mengakomodasi pembuatan fitur *rolling statistics*, rasio antar *log*, komposit indeks seperti *Coal Index*, serta deteksi *boundary* atau lapisan, agar pola zona batubara dapat terpetakan dengan konsisten dan sesuai prinsip geologi. Seleksi fitur juga harus mengutamakan interpretabilitas serta menghindari *multicollinearity*, salah satunya melalui *threshold Variance Inflation Factor* (VIF).

Algoritma *supervised learning* hanya digunakan untuk memetakan probabilitas *Coal Index* berdasarkan label litologi asli dari geolog. Sementara itu, proses identifikasi zona dilakukan secara *unsupervised* melalui *clustering* sehingga pola alami segmentasi lapisan bisa ditangkap tanpa bergantung pada label. Semua hasil analisis wajib divisualisasikan secara eksploratif, sehingga tim geologi dapat melakukan interpretasi zona batubara dengan lebih efisien pada sumur baru ataupun kasus eksplorasi lainnya.

Pipeline pengelompokan zona batubara ditunjukkan pada Gambar 3.21, dimulai dari analisis awal hingga klasifikasi dan segmentasi zona berdasarkan data *geophysical logging*.



Gambar 3.21. *Pipeline* untuk mengelompokkan zona batubara berdasarkan data *geophysical logging*

B *Implementasi*

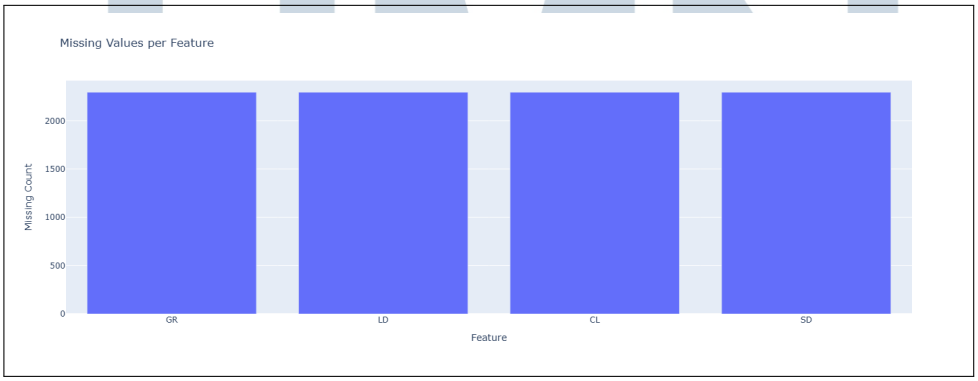
B.1 *Exploratory Data Analysis*

Tahap eksplorasi data awal dilakukan untuk memperoleh pemahaman mendalam terkait karakteristik data *logging* geofisika yang menjadi dasar pembangunan model. Pada tahap ini, analisis fokus pada empat fitur utama yaitu GR, CL, LD, dan SD, karena fitur-fitur ini secara geologi paling relevan untuk membedakan zona batubara dari litologi lain. Dengan memahami pola distribusi, jumlah *missing values*, dan korelasi antar fitur, proses pengolahan dan analisis lanjutan dapat dilakukan dengan lebih terarah serta mengantisipasi potensi kendala data yang mungkin mempengaruhi kinerja model.

Gambar 3.22 menunjukkan distribusi dari empat fitur utama data geofisika (GR, LD, CL, dan SD) setelah dilakukan transformasi logaritmik serta pemangkasan nilai ekstrem pada kuantil ke-1 hingga ke-99. Transformasi *log* digunakan untuk mengurangi *skewness* dan menyeimbangkan skala nilai, sedangkan *clipping* mencegah *outlier* ekstrem mendominasi visualisasi. Hasilnya, distribusi GR dan CL menjadi lebih simetris menyerupai distribusi normal, sedangkan LD dan SD masih menunjukkan puncak tajam, mengindikasikan dominasi nilai-nilai tertentu yang kemungkinan merefleksikan litologi seragam di area pengukuran.

Gambar 3.22. Distribusi masing-masing fitur utama (GR, LD, CL, SD) pada data geofisika

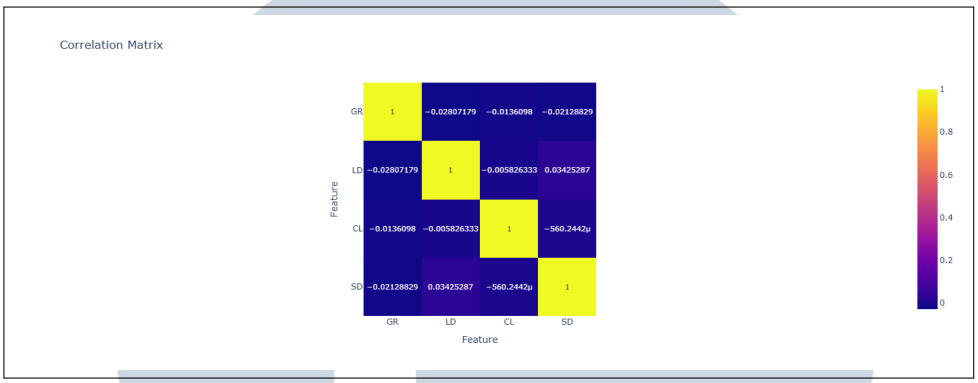
Selain distribusi, salah satu aspek penting dalam eksplorasi awal adalah mengetahui sejauh mana data yang tersedia memiliki *missing values*, karena *missing values* yang terlalu banyak dapat mengganggu proses pelatihan model. Gambar 3.23 menunjukkan jumlah *missing values* pada masing-masing fitur utama. Dapat diamati bahwa seluruh fitur utama memiliki jumlah *missing values* yang hampir seragam, sehingga strategi penanganan *missing values* perlu dirancang secara konsisten untuk seluruh fitur tersebut. Kondisi ini mengingatkan pentingnya proses imputasi dan *filtering* data pada tahap *preprocessing*, agar hasil analisis tidak bias akibat data yang tidak lengkap.



Gambar 3.23. Jumlah *missing values* pada masing-masing fitur utama

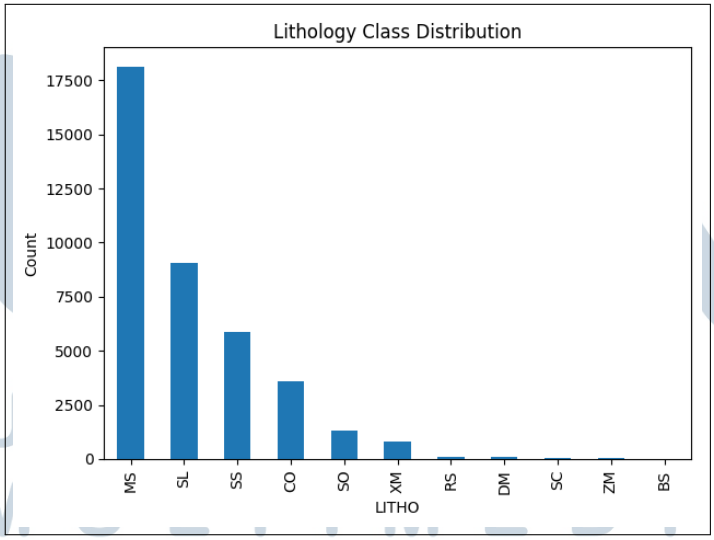
Selanjutnya, untuk memastikan tidak ada fitur yang redundant atau saling berkorelasi sangat kuat (multikolinearitas tinggi), dilakukan analisis korelasi antar fitur utama. Hasil pada Gambar 3.24 memperlihatkan bahwa korelasi antar fitur utama relatif rendah, tidak ditemukan hubungan linier yang signifikan di antara GR, CL, LD, dan SD. Hal ini menandakan bahwa setiap fitur menyumbang

informasi unik dan dapat dipertahankan dalam model tanpa risiko redundansi yang berlebihan. Korelasi rendah juga memberikan fleksibilitas lebih dalam membangun fitur turunan serta mendorong model untuk belajar pola yang lebih kaya dari data.



Gambar 3.24. Matriks korelasi antar fitur utama pada data geofisika

Selain fitur *log*, distribusi kelas litologi juga menjadi faktor penting untuk dianalisis sejak awal, khususnya untuk mengantisipasi potensi ketidakseimbangan kelas pada data target. Gambar 3.25 menampilkan distribusi jumlah data pada masing-masing kelas litologi. Tampak jelas bahwa kelas MS, SL, dan SS mendominasi jumlah data, sementara kelas CO (*coal*) memiliki proporsi yang jauh lebih sedikit dibandingkan beberapa kelas lain.



Gambar 3.25. Distribusi jumlah data pada masing-masing kelas litologi

B.2 *Data Cleaning and Feature Engineering*

Setelah memahami karakteristik awal data, tahap selanjutnya adalah melakukan pembersihan data dan *feature engineering*. Tahap ini penting untuk

memastikan kualitas data yang masuk ke model benar-benar optimal, serta menghasilkan fitur-fitur baru yang lebih informatif dan relevan untuk proses identifikasi zona *coal*.

Pada proses pembersihan data, langkah pertama yang dilakukan adalah menghilangkan data yang teridentifikasi sebagai *geological outlier* berdasarkan *domain knowledge*, seperti nilai GR di luar rentang karakteristik *coal* ataupun nilai-nilai ekstrem pada LD dan SD. Data yang memiliki *missing values* berlebihan atau terdeteksi sebagai *outlier* multivariat juga dihapus agar tidak mengganggu proses pembelajaran model. Selain itu, sisa nilai kosong pada fitur utama diimputasi menggunakan nilai rata-rata, sehingga *dataset* menjadi lebih konsisten dan siap untuk tahap pengolahan berikutnya.

Selanjutnya, dilakukan *feature engineering* dengan membangun berbagai fitur turunan yang secara geologi relevan untuk membedakan zona *coal* dan non-*coal*. Contoh fitur yang dikembangkan antara lain rasio GR terhadap LD (GR/LD), skor litologi gabungan (*coal lithology score*), rasio Caliper terhadap Density untuk mendeteksi *washout*, serta *Coal Index* yang mengintegrasikan berbagai parameter *log* untuk memberikan estimasi probabilitas keberadaan *coal*. Proses rekayasa fitur ini juga mencakup normalisasi, perhitungan gradien (*derivative*) antar *log*, *rolling statistics* untuk menangkap pola lokal, serta *flagging* terhadap interval yang mengalami *washout* atau ketidakstabilan lubang bor.

Rangkaian proses *cleaning* dan *feature engineering* menghasilkan *dataset* dengan fitur-fitur baru yang lebih kaya informasi serta telah lolos dari potensi anomali data. Jumlah data yang semula terdiri dari 39.046 baris dan 6 kolom, setelah proses ini bertransformasi menjadi 35.256 baris dengan 65 kolom, seperti terlihat pada Gambar 3.26. Hal ini menandakan adanya pengurangan data yang tidak relevan serta penambahan fitur-fitur hasil *feature engineering* yang siap digunakan pada proses pemodelan.

	HOLE_ID	DEPT	GR	CL	LD	SD	GR_grad	GR_cum_abs_grad	CL_grad	CL_cum_abs_grad	...	CL_norm_contrast	LD_contrast	LD_norm_contrast	SD_contrast	SD_
0	HOLE_001	1.75	42.01	5.86	1382.497294	9722.787366	0.0	0.000000	0.0	0.000000	...	1.133893	0.0000	0.000000	0.0000	
1	HOLE_001	1.80	44.83	5.82	1382.497294	9722.787366	56.4	0.043687	-0.8	0.000620	...	-0.095059	0.0000	0.000000	0.0000	
2	HOLE_001	1.85	48.20	5.81	1382.497294	9722.787366	67.4	0.095895	-0.2	0.000775	...	-0.321288	0.0000	0.000000	0.0000	
3	HOLE_001	1.90	45.13	5.80	1382.497294	9722.787366	-61.4	0.143455	-0.2	0.000930	...	-0.670820	0.0000	0.000000	0.0000	
4	HOLE_001	1.95	46.28	5.80	1382.497294	9722.787366	23.0	0.161270	0.0	0.000930	...	-0.447214	0.0000	0.000000	0.0000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
35251	HOLE_039	147.40	12.45	2.85	511.940000	6901.400000	-84.4	158.194670	-0.4	0.764489	...	0.249136	-0.7980	-0.036781	-34.2220	
35252	HOLE_039	147.45	15.58	2.81	489.460000	6933.420000	62.6	158.216175	-0.8	0.764764	...	-0.041523	-21.3860	-1.013409	64.5380	
35253	HOLE_039	147.50	14.68	2.77	544.130000	6914.020000	-18.0	158.222359	-0.8	0.765039	...	-0.508001	31.9040	1.586160	27.8820	
35254	HOLE_039	147.55	17.24	2.76	512.430000	6810.700000	51.2	158.239947	-0.2	0.765108	...	-0.630126	0.1325	0.005705	-71.6225	
35255	HOLE_039	147.60	117.37	2.76	503.170000	6871.150000	2002.6	158.927889	0.0	0.765108	...	-0.577350	-16.7400	-0.779327	5.8600	

35256 rows × 65 columns

Gambar 3.26. Contoh tampilan data hasil pembersihan data dan *feature engineering*

B.3 *Feature Selection and Multicollinearity Reduction*

Tahap berikutnya setelah rekayasa fitur adalah melakukan seleksi fitur untuk memastikan hanya fitur-fitur paling relevan dan bebas dari multikolinearitas yang digunakan dalam pemodelan. Seleksi fitur ini dilakukan dengan memanfaatkan uji VIF, yaitu metode statistik untuk mendeteksi adanya korelasi linear berlebih antar fitur numerik. Setiap fitur yang memiliki nilai VIF di atas ambang batas tertentu secara bertahap dihilangkan hingga tersisa *subset* fitur yang saling bebas secara statistik.

Gambar 3.27 menampilkan daftar fitur hasil seleksi setelah proses VIF filtering. Terlihat bahwa fitur yang dipertahankan memiliki nilai VIF di bawah 10, sehingga potensi multikolinearitas sudah ditekan seminimal mungkin. Beberapa fitur utama yang lolos seleksi antara lain rasio GR terhadap LD, *rolling statistics*, *flag washout*, hingga parameter *boundary strength*. Dengan komposisi fitur yang sudah tervalidasi secara statistik ini, model dapat dilatih dengan lebih *robust* dan risiko bias akibat korelasi *internal* antar fitur dapat diminimalisasi.

```
Selected features after VIF filtering (with VIF values):
GR_LD_product                VIF = 9.07
transition_zone              VIF = 7.57
GR_contrast                  VIF = 6.28
GR_roll_cv                   VIF = 6.20
GR_norm_contrast             VIF = 5.92
GR_to_LD                     VIF = 5.67
SD_grad                      VIF = 5.01
CL_cum_abs_grad              VIF = 4.62
SD_contrast                  VIF = 4.11
LD_grad                      VIF = 3.72
rel_pos_in_layer             VIF = 3.50
Coal_Boundary                VIF = 3.48
layer_thickness              VIF = 3.48
LD_roll_range                VIF = 3.47
high_ash_flag                VIF = 3.01
Coal_Boundary_Strength       VIF = 2.92
SD_grad2                     VIF = 2.88
CL_excursion                 VIF = 2.66
SD_roll_cv                   VIF = 2.64
param_trend_consistency      VIF = 2.60
LD_contrast                  VIF = 2.49
CL_grad                      VIF = 2.48
CL_washout_flag              VIF = 2.38
CL_contrast                  VIF = 2.10
LD_grad2                     VIF = 2.00
SD_norm_contrast             VIF = 1.97
LD_norm_contrast             VIF = 1.87
CL_grad2                     VIF = 1.85
CL_roll_range                VIF = 1.84
CL_norm_contrast             VIF = 1.35
GR_grad2                     VIF = 1.26
```

Gambar 3.27. Daftar fitur hasil seleksi setelah proses VIF *filtering*

B.4 *Coal Zone Classification*

Setelah proses pembersihan data dan rekayasa fitur, tahap selanjutnya adalah klasifikasi zona batubara menggunakan model *machine learning*. Pada proses

ini, model utama yang digunakan adalah Random Forest dan XGBoost. Evaluasi dilakukan baik pada data *training* (*in-sample*) maupun data *testing* (*hold-out*) untuk memastikan model tidak hanya menghafal data *training*, tetapi juga mampu melakukan generalisasi ke data baru yang belum pernah dilihat.

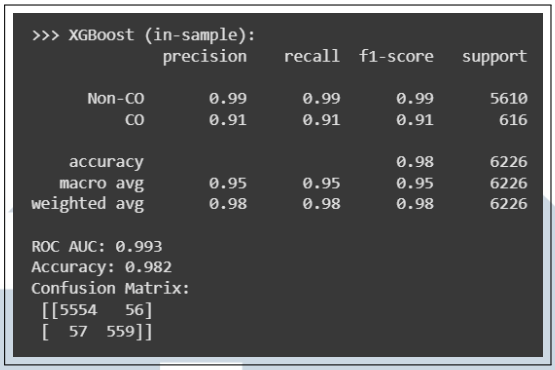
Penilaian kinerja dilakukan menggunakan metrik klasifikasi seperti *precision*, *recall*, *F1-score*, akurasi, ROC AUC, serta *confusion matrix*. Hasil evaluasi pada model Random Forest dan XGBoost diperlihatkan secara komprehensif pada Gambar 3.28 sampai Gambar 3.31.

Gambar 3.28 memperlihatkan performa Random Forest pada data *training* (*in-sample*), dengan akurasi mencapai 97,8% dan ROC AUC sebesar 0,981. Seluruh metrik utama menunjukkan bahwa model mampu mendeteksi zona batubara (CO) dan non-batubara (Non-CO) dengan sangat baik, meskipun nilai *recall* pada kelas batubara sedikit lebih rendah dibandingkan kelas lainnya.

```
>>> Random Forest (in-sample):
              precision    recall  f1-score   support

      Non-CO       0.99      0.99      0.99      5610
          CO       0.89      0.88      0.89       616

    accuracy                           0.98      6226
   macro avg       0.94      0.94      0.94      6226
weighted avg       0.98      0.98      0.98      6226

ROC AUC: 0.981
Accuracy: 0.978
Confusion Matrix:
 [[5543   67]
 [  71  545]]
```

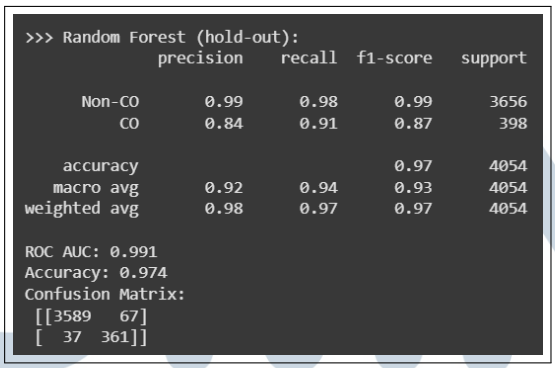
Gambar 3.28. Evaluasi Random Forest pada data *training* (*in-sample*)

Selanjutnya, performa XGBoost pada data *training* ditunjukkan pada Gambar 3.29, dengan akurasi mencapai 98,2% dan ROC AUC sebesar 0,993. Nilai *precision*, *recall*, dan *F1-score* pada kelas batubara sedikit lebih tinggi dibandingkan Random Forest, yang mengindikasikan bahwa model XGBoost sedikit lebih optimal dalam mendeteksi zona batubara pada data *training*.

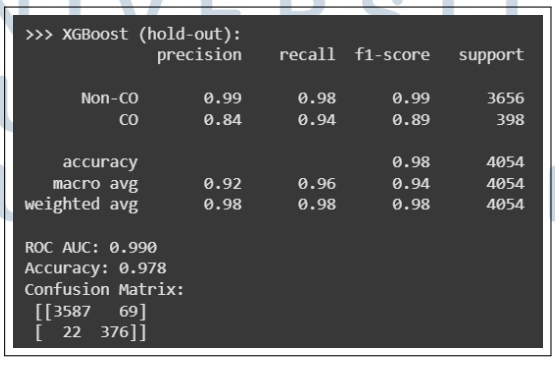


Gambar 3.29. Evaluasi XGBoost pada data *training* (*in-sample*)

Pada data *testing hold-out*, performa model Random Forest ditunjukkan pada Gambar 3.30, sedangkan XGBoost pada Gambar 3.31, keduanya menunjukkan akurasi yang tetap konsisten di atas 97%. Hal ini mengindikasikan bahwa model tidak mengalami *overfitting*. Evaluasi pada data *testing* yang benar-benar baru memberikan gambaran objektif mengenai kemampuan generalisasi model. Performa XGBoost pada data *hold-out* sedikit lebih unggul dibandingkan Random Forest, terutama pada nilai *recall* dan *F1-score* untuk kelas batubara, yang sangat krusial dalam konteks aplikasi geologi agar zona batubara seminimal mungkin terlewatkan.



Gambar 3.30. Evaluasi Random Forest pada data *testing* (*hold-out*)



Gambar 3.31. Evaluasi XGBoost pada data *testing* (*hold-out*)

Selanjutnya, untuk validasi model terhadap label batubara orisinal dari geologi menggunakan keseluruhan data, diperoleh akurasi yang sangat tinggi sebagaimana ditunjukkan pada Gambar 3.32 dan Gambar 3.33. Random Forest mampu menandai sekitar 3486 dari 3582 interval batubara aktual, sementara XGBoost sedikit lebih baik dengan mendeteksi 3491 interval. Hasil ini menunjukkan bahwa kedua model dapat diandalkan dalam mendeteksi batubara, dengan XGBoost sedikit unggul dalam presisi dan *recall* pada label aktual.

```
=== Random Forest vs Original Coal Labels ===
Accuracy:  97.79%
Precision: 88.73%
Recall:    88.93%
F1-score:  88.83%
Confusion Matrix:
 [[31309   393]
 [  385  3093]]

Classification Report for Random Forest:
              precision    recall  f1-score   support

      Non-CO       0.99      0.99      0.99     31702
          CO       0.89      0.89      0.89      3478

    accuracy                           0.98     35180
   macro avg       0.94      0.94      0.94     35180
weighted avg       0.98      0.98      0.98     35180
```

Gambar 3.32. Evaluasi Random Forest pada label batubara orisinal (*full-data*)

```
=== XGBoost vs Original Coal Labels ===
Accuracy:  98.80%
Precision: 93.78%
Recall:    94.13%
F1-score:  93.96%
Confusion Matrix:
 [[31485   217]
 [  204  3274]]

Classification Report for XGBoost:
              precision    recall  f1-score   support

      Non-CO       0.99      0.99      0.99     31702
          CO       0.94      0.94      0.94      3478

    accuracy                           0.99     35180
   macro avg       0.97      0.97      0.97     35180
weighted avg       0.99      0.99      0.99     35180
```

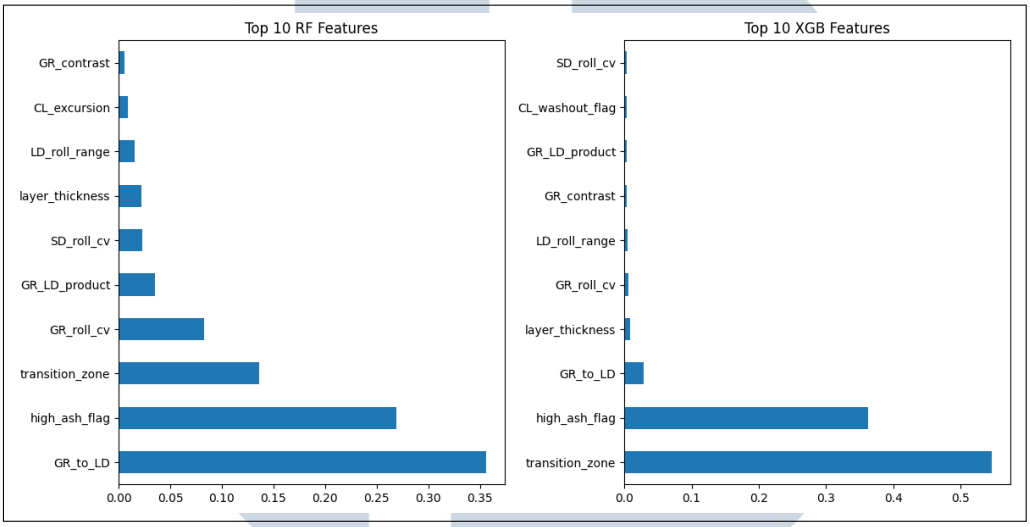
Gambar 3.33. Evaluasi XGBoost pada label batubara orisinal (*full-data*)

```
Random Forest flagged 3486 out of 3582 actual coal intervals
XGBoost flagged 3491 out of 3582 actual coal intervals
```

Gambar 3.34. Jumlah interval batubara aktual yang berhasil dideteksi masing-masing model

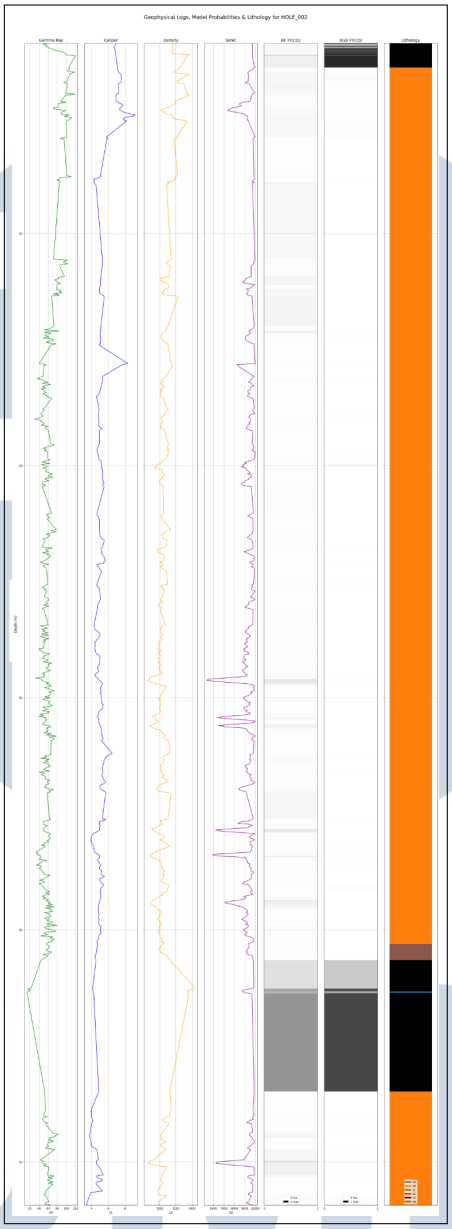
Selain evaluasi metrik numerik, analisis lebih lanjut dilakukan untuk mengetahui fitur-fitur paling penting dalam membedakan zona batubara. Gambar 3.35 menunjukkan sepuluh fitur teratas menurut model Random Forest

dan XGBoost. Terlihat bahwa fitur seperti GR_to_LD, high_ash_flag, dan transition_zone secara konsisten muncul sebagai penentu utama dalam deteksi batubara. Hal ini menunjukkan bahwa kombinasi *log Gamma Ray*, *Density*, dan indikator kualitas batubara sangat berperan penting dalam pemodelan ini.



Gambar 3.35. *Feature importance* pada model Random Forest dan XGBoost (*Top 10*)

Gambar 3.36 menampilkan visualisasi profil *log* sumur yang memperlihatkan data log utama, probabilitas prediksi batubara dari model, serta label litologi aktual untuk salah satu sumur. Visualisasi ini sangat penting untuk mengevaluasi apakah prediksi model telah sesuai dengan pola *log* yang diharapkan oleh domain geologi.

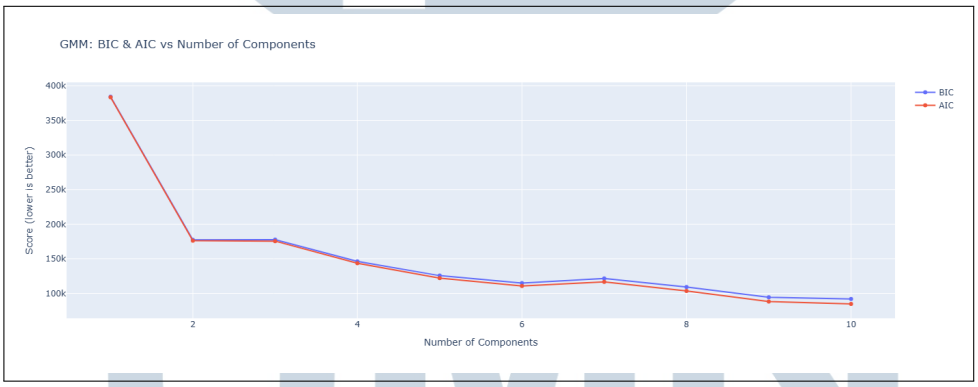


Gambar 3.36. Visualisasi *log* geofisika, probabilitas model, dan label litologi pada satu sumur

Berdasarkan seluruh hasil evaluasi, baik Random Forest maupun XGBoost menunjukkan performa yang sangat baik dengan akurasi tinggi dan kemampuan deteksi batubara yang konsisten, baik pada data *training*, data *testing*, maupun validasi terhadap label geologi aktual. Namun, XGBoost secara konsisten sedikit lebih unggul pada aspek *recall* dan *precision* khusus untuk kelas batubara (CO), yang menjadi fokus utama pada aplikasi eksplorasi batubara. Dengan demikian, XGBoost menjadi model utama sebagai *baseline* deteksi zona batubara untuk dataset ini.

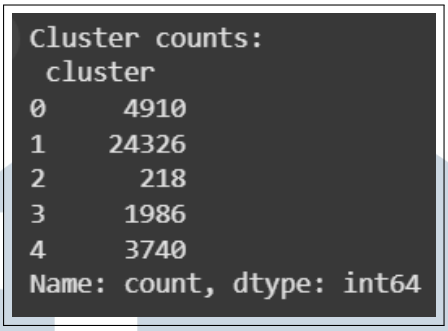
B.5 *Coal Zone Clustering*

Tahapan *clustering* diterapkan setelah klasifikasi untuk mendapatkan segmentasi zona batubara berdasarkan kemiripan karakteristik hasil *feature engineering*. Pada tahap ini, Gaussian Mixture Model (GMM) dipilih karena kemampuannya memodelkan distribusi data yang kompleks dan menangkap variasi antar kelompok secara lebih fleksibel. Sebelum memutuskan jumlah *cluster*, dilakukan evaluasi model menggunakan Bayesian Information Criterion (BIC) dan Akaike Information Criterion (AIC) untuk menghindari *overfitting* sekaligus memastikan tiap *cluster* benar-benar merepresentasikan pola nyata di data, bukan *noise* acak. Hasilnya dapat dilihat pada visualisasi tren BIC dan AIC di Gambar 3.37, yang menunjukkan penurunan tajam hingga lima *cluster* pertama lalu relatif datar. Hal ini menandakan bahwa penambahan *cluster* di atas lima tidak memberikan informasi tambahan yang signifikan, sehingga lima *cluster* dianggap paling optimal untuk membedakan zona batubara berdasarkan fitur geofisika yang ada.



Gambar 3.37. Perbandingan nilai BIC dan AIC untuk berbagai jumlah *cluster* pada GMM

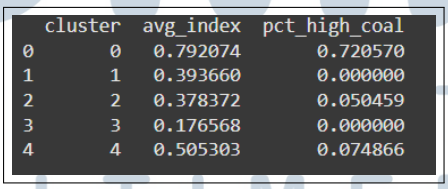
Setelah proses *clustering* dengan lima *cluster*, distribusi anggota pada masing-masing *cluster* dapat dilihat pada Gambar 3.38. Terlihat bahwa sebagian besar data terkonsentrasi pada *Cluster* 1, sementara beberapa *cluster* lain, seperti *Cluster* 0, memiliki jumlah anggota yang jauh lebih sedikit. Ketimpangan distribusi ini wajar dalam konteks eksplorasi batubara, karena secara alami zona-zona dengan karakteristik tertentu (misalnya zona batubara yang sesungguhnya) memang jauh lebih sedikit dibanding zona non-batubara. Analisis distribusi ini penting untuk memastikan bahwa hasil *clustering* tidak didominasi satu kelompok saja, sehingga seluruh variasi karakteristik di data tetap bisa terwakili dan zona-zona dengan fitur unik tetap bisa teridentifikasi secara mandiri.



Gambar 3.38. Distribusi jumlah data pada masing-masing *cluster*

Pada analisis statistik per *cluster* berikutnya, *Coal Index* yang digunakan merupakan hasil prediksi probabilitas zona batubara dari model XGBoost pada tahap klasifikasi sebelumnya. Nilai ini memberikan estimasi tingkat keyakinan terhadap keberadaan batubara di setiap interval berdasarkan pola *log* geofisika.

Statistik per *cluster* kemudian dianalisis untuk memperjelas karakteristik masing-masing kelompok. Gambar 3.39 menyajikan ringkasan statistik berupa rata-rata nilai *Coal Index* dan persentase interval dengan *Coal Index* tinggi di tiap *cluster*. *Cluster* 0 secara jelas menonjol sebagai zona yang paling prospektif dengan rata-rata nilai *Coal Index* tinggi (0,79) dan persentase zona batubara berkualitas tinggi mencapai 72,06%. Sebaliknya, *Cluster* 1 dan 3 menunjukkan karakteristik *non-coal*, dengan rata-rata nilai *Coal Index* yang rendah (masing-masing 0,39 dan 0,17) serta tidak terdapat zona batubara berkualitas tinggi. *Cluster* 2 memiliki rata-rata nilai *Coal Index* rendah (0,38) dengan sedikit persentase interval batubara tinggi (5,05%), sehingga hanya dianggap memiliki potensi batubara marginal. Terakhir, *Cluster* 4 berada pada posisi menengah dengan rata-rata nilai *Coal Index* sebesar 0,50 dan persentase zona batubara tinggi sebesar 7,49%, menunjukkan potensi batubara moderat namun prospeknya lebih terbatas dibandingkan *Cluster* 0.



	cluster	avg_index	pct_high_coal
0	0	0.792074	0.720570
1	1	0.393660	0.000000
2	2	0.378372	0.050459
3	3	0.176568	0.000000
4	4	0.505303	0.074866

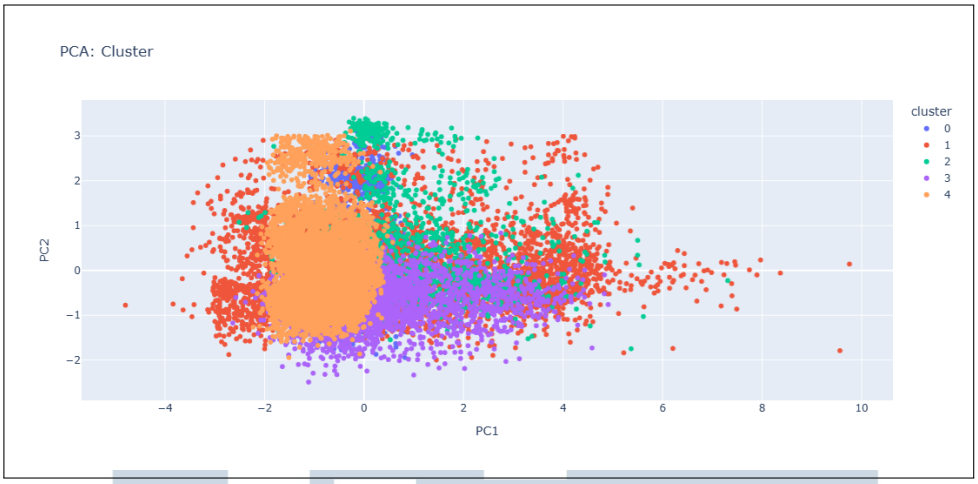
Gambar 3.39. Rata-rata *Coal Index* dan proporsi *high-coal* pada tiap *cluster*

Distribusi zona batubara terhadap kedalaman diperjelas lagi pada plot Gambar 3.40. Sumbu Y memperlihatkan rentang kedalaman sumur, sementara warna titik mengindikasikan *cluster*. Terlihat pola bahwa *Cluster 0* cenderung mengelompok di interval kedalaman tertentu, dan zona ini didominasi nilai *Coal Index* tinggi. Sebaliknya, *cluster* dengan nilai indeks lebih rendah tersebar secara

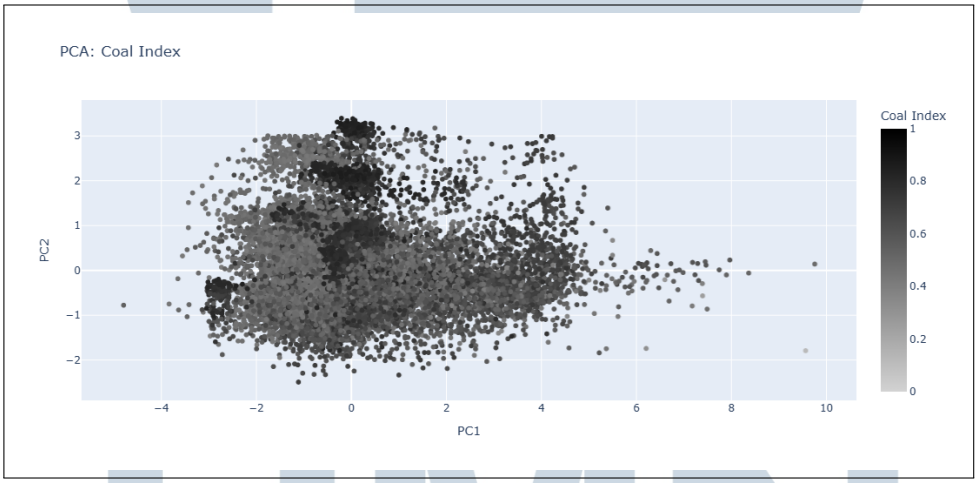
acak pada kedalaman beragam. Visualisasi ini penting untuk keperluan eksplorasi geologi karena bisa membantu memprioritaskan interval target pengeboran pada kedalaman-kedalaman yang memang kaya batubara, sekaligus meminimalisasi pengeboran di zona yang kurang prospektif.

Gambar 3.40. Distribusi *Coal Index* terhadap kedalaman pada masing-masing *cluster*

Analisis visual berikutnya menggunakan PCA (Principal Component Analysis) untuk memproyeksikan data hasil *clustering* ke ruang dua dimensi berdasarkan komponen utama. Gambar 3.41 menampilkan sebaran data pada dua komponen utama PCA dengan pewarnaan berdasarkan *cluster*. Terlihat sebagian besar *cluster* memiliki kecenderungan mengelompok, meskipun ada area *overlap*. Hal ini wajar karena batasan antara zona batubara dan non-batubara tidak selalu tegas secara statistik. Namun, pemisahan *cluster* tetap terlihat dan memperkuat bahwa fitur hasil *engineering* berhasil menangkap struktur laten dari data. Selanjutnya, saat pewarnaan menggunakan *Coal Index* pada Gambar 3.42, distribusi ini makin tegas memperlihatkan konsentrasi zona *Coal Index* tinggi di area yang sama dengan *Cluster 0*, memperkuat validitas metode *clustering* yang digunakan.

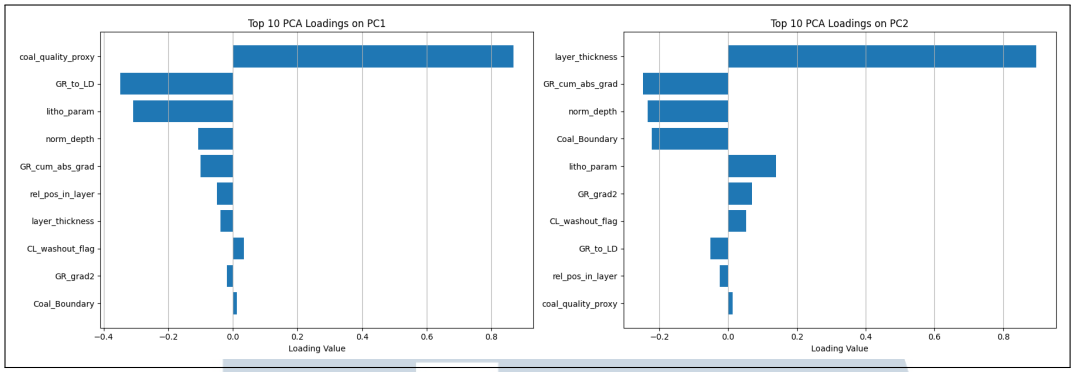


Gambar 3.41. Distribusi *cluster* pada ruang dua komponen utama PCA



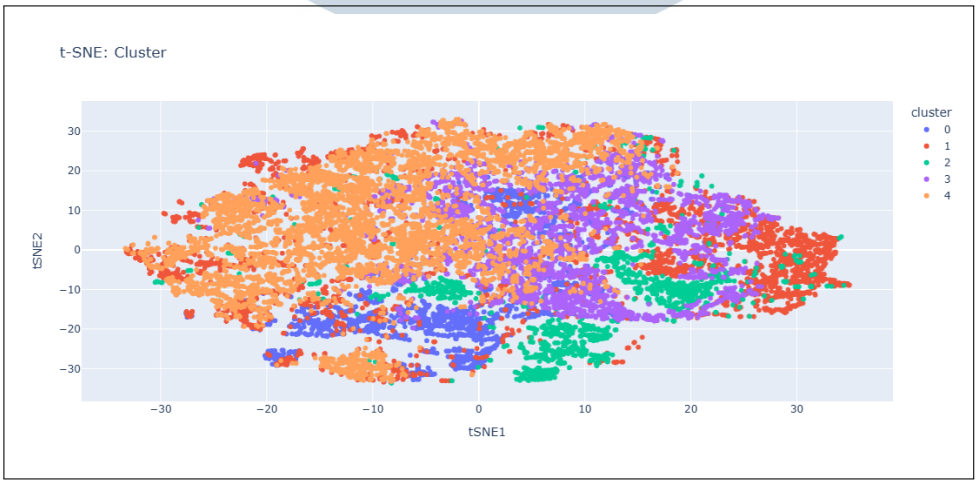
Gambar 3.42. Distribusi *Coal Index* pada ruang dua komponen utama PCA

Kontribusi fitur utama pada dua komponen PCA dapat dilihat pada Gambar 3.43. Fitur seperti *coal_quality_proxy*, *GR_to_LD*, *litho_param*, dan *layer_thickness* merupakan penyumbang *loading* terbesar, yang berarti fitur-fitur ini sangat menentukan posisi masing-masing data dalam ruang PCA. Fakta bahwa fitur-fitur tersebut berhubungan erat dengan sifat fisik batubara (kualitas, densitas, dan ketebalan lapisan) menegaskan bahwa model tidak hanya bekerja secara statistik, namun juga tetap relevan secara geologi.



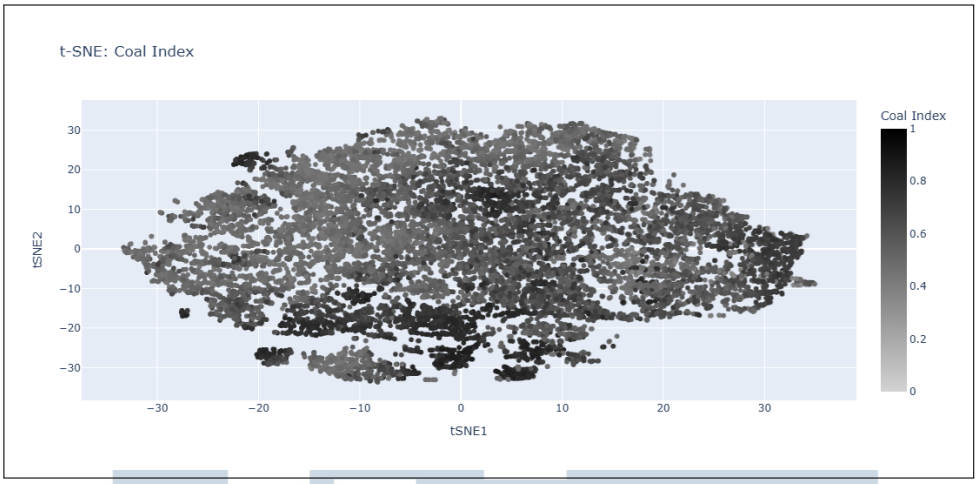
Gambar 3.43. *Top 10* PCA *loadings* untuk dua komponen utama (PC1 dan PC2)

Proyeksi *clustering* ke ruang non-linear divisualisasikan dengan t-SNE (t-Distributed Stochastic Neighbor Embedding) pada Gambar 3.44. Hasil t-SNE memperlihatkan bahwa meskipun ada *overlap*, pola spasial antar *cluster* tetap bisa dipetakan dan terdapat kelompok besar yang mewakili *cluster* mayoritas, serta beberapa area padat yang didominasi *Cluster 0* (batubara). Ini menegaskan konsistensi hasil *clustering* baik pada ruang linear (PCA) maupun non-linear (t-SNE).



Gambar 3.44. Visualisasi *cluster* pada ruang t-SNE

Terakhir, Gambar 3.45 menunjukkan t-SNE yang diwarnai menggunakan *Coal Index*, sehingga pola distribusi zona dengan nilai indeks tinggi makin terlihat menonjol. Area konsentrasi *Coal Index* tinggi tetap terjaga pada *Cluster 0*, selaras dengan hasil *clustering* dan analisis pada ruang PCA. Visualisasi ini memberikan justifikasi tambahan bahwa zona batubara memang membentuk kelompok unik dan konsisten secara statistik, spasial, dan geologi.



Gambar 3.45. Visualisasi *Coal Index* pada ruang t-SNE

Secara keseluruhan, seluruh rangkaian analisis *clustering* menunjukkan bahwa zona batubara dengan *Coal Index* tinggi berhasil dipisahkan dan diidentifikasi dengan baik menggunakan kombinasi GMM, PCA, dan t-SNE. *Cluster 0* terbukti paling prospektif untuk eksplorasi lebih lanjut, karena konsisten memiliki nilai *Coal Index* tinggi pada seluruh visualisasi dan statistik, sehingga bisa menjadi acuan utama dalam pengambilan keputusan eksplorasi batubara berikutnya.

3.4 Kendala yang Ditemukan

Selama pelaksanaan kerja magang di PT Berau Coal Energy, terdapat beberapa kendala teknis yang dihadapi dalam pengerjaan proyek-proyek yang terkait dengan analisis data dan pemodelan *machine learning*, yaitu sebagai berikut.

1. Terdapat kendala komputasi dalam pemilihan dan evaluasi model *summarization*, *sentiment analysis*, dan identifikasi risiko dan peluang akibat ukuran model berbasis *transformer* yang besar dan keterbatasan sumber daya komputasi (GPU dan memori) pada lingkungan kerja Google Colab.
2. Beberapa hasil analitik otomatis, baik pada prediksi harga maupun deteksi zona batubara memerlukan validasi domain oleh ahli geologi atau praktisi bisnis. Namun, keterbatasan waktu dan tenaga ahli menyebabkan proses validasi kadang harus ditunda, sehingga *feedback* untuk model *improvement* menjadi kurang cepat.

3.5 Solusi atas Kendala yang Ditemukan

Berdasarkan berbagai kendala yang ditemukan selama pelaksanaan kerja magang, telah dirancang dan diterapkan beberapa solusi strategis untuk mengatasi dan meminimalisasi dampak negatif dari kendala tersebut, yaitu sebagai berikut.

1. Menggunakan model *transformer* yang telah dioptimasi dengan teknik seperti *quantization* (4-bit), serta memilih model yang lebih ringan namun tetap kompetitif seperti DistilBART, DistilRoBERTa, dan model khusus seperti FinancialBERT. Strategi ini efektif untuk mengatasi keterbatasan sumber daya komputasi di Google Colab tanpa menurunkan kualitas analisis secara signifikan.
2. Untuk mengatasi keterbatasan validasi domain, dilakukan sesi diskusi rutin dengan *supervisor* yang memiliki latar belakang pendidikan geologi untuk memverifikasi hasil dari model utama. Seluruh masukan yang diberikan didokumentasikan agar dapat diintegrasikan ke dalam pengembangan model selanjutnya.