

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perubahan iklim telah menjadi isu global yang mendesak, di mana emisi karbon dioksida (CO_2) teridentifikasi sebagai kontributor utama pemanasan global yang memicu ketidakstabilan ekosistem (IPCC, 2023). Sebagai salah satu negara dengan populasi dan pertumbuhan ekonomi terbesar di Asia Tenggara, Indonesia memiliki peran strategis dalam upaya mitigasi perubahan iklim sesuai komitmen *Nationally Determined Contribution* (NDC) untuk mengurangi emisi karbon hingga 29% dengan usaha sendiri pada tahun 2030 (Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, 2022). Namun, pemantauan dan analisis emisi karbon menghadapi tantangan berupa keterkaitan multivariat antar indikator makroekonomi serta dinamika deret waktu jangka panjang. Data emisi karbon dipengaruhi oleh aktivitas ekonomi, konsumsi energi, dan perubahan penggunaan lahan, yang bersifat saling terkait dan non-linear (IPCC, 2023).

Pemodelan prediktif berbasis *machine learning* pada data multivariat merupakan salah satu topik yang berkembang pesat dalam penelitian ilmu komputer. Salah satu domain aplikasi yang relevan adalah pemodelan emisi karbon dioksida (CO_2) pada tingkat nasional, yang melibatkan hubungan non-linear kompleks antara sejumlah variabel makroekonomi dan energi dalam rentang waktu panjang. Kondisi inilah yang mendorong eksplorasi pendekatan *machine learning* berbasis *ensemble learning* dan seleksi fitur sebagai alternatif yang lebih sesuai.

Dari perspektif ilmu komputer, terdapat dua tantangan metodologis utama dalam membangun model prediktif pada data semacam ini. Pertama, adanya hubungan non-linear dan multikolinearitas antar variabel prediktor yang tidak dapat ditangkap secara memadai oleh model statistik linier (Chen & Guestrin, 2016). Kedua, penggunaan seluruh variabel tanpa seleksi sistematis dapat menurunkan efisiensi komputasi dan berpotensi mengurangi kemampuan generalisasi model, terutama pada *dataset* dengan jumlah observasi terbatas (Guyon & Elisseeff, 2003).

Sebagai konteks aplikasi, penelitian ini menggunakan data deret waktu tahunan emisi CO_2 per kapita Indonesia periode 1990–2023, dengan lima variabel

prediktor: Produk Domestik Bruto (PDB) per kapita, konsumsi energi primer, produksi listrik berbasis energi terbarukan, luas lahan hutan, dan jumlah populasi. Data bersumber dari World Bank Open Data, Energy Institute Statistical Review of World Energy, dan World Bank Open Data (indikator EG.ELC.RNEW.ZS). Dalam konteks nasional Indonesia, data Badan Pusat Statistik (BPS) menunjukkan bahwa sektor energi dan industri menjadi kontributor dominan emisi gas rumah kaca, yang memperkuat relevansi pemilihan variabel berbasis energi dan ekonomi dalam penelitian ini.

Penelitian terdahulu telah menunjukkan potensi *machine learning* dalam pemodelan emisi karbon. (Wang et al., 2025) menerapkan strategi seleksi fitur yang mencakup *Recursive Feature Elimination* (RFE) bersama analisis korelasi dan penilaian kepentingan fitur, diikuti pemodelan XGBoost untuk prediksi emisi karbon di sektor pembangkit listrik, dan memperoleh XGBoost sebagai model dengan akurasi tertinggi. Meski demikian, peningkatan akurasi pada studi tersebut mencakup seleksi fitur sekaligus optimasi parameter, sehingga kontribusi seleksi fitur tidak dapat dipisahkan secara terisolasi. (Begum & Mobin, 2025) menggunakan pendekatan *machine learning* multivariat dengan *feature selection* pada data emisi CO₂ sebelas negara, termasuk Indonesia, dan menunjukkan bahwa pendekatan tersebut mampu menghasilkan akurasi prediksi yang lebih baik dibandingkan model tanpa seleksi fitur. Pendekatan berbasis kepentingan fitur juga menunjukkan manfaat serupa, (Satvika et al., 2024) meningkatkan kinerja algoritma K-Nearest Neighbor menggunakan *permutation feature importance* dan memperoleh akurasi prediksi yang lebih tinggi dibandingkan model tanpa teknik tersebut. (Gurcan, 2024) dalam studi komparatif menemukan bahwa XGBoost dan metode *ensemble learning* secara konsisten mengungguli model konvensional dalam prediksi emisi CO₂ pada data tabular.

Meskipun demikian, terdapat celah penelitian yang belum terjawab. Studi-studi yang menerapkan RFE dengan *estimator* Random Forest pada pemodelan emisi CO₂ berfokus pada data operasional dengan dimensi fitur besar, seperti (Tu et al., 2024) yang mereduksi 46 fitur pada pembangkit listrik tenaga batu bara. Studi berbasis XGBoost lainnya juga bertumpu pada konteks operasional atau sektoral: (Wang et al., 2025) menggunakan 18 parameter pada satu pembangkit listrik

dengan data harian, (Cai et al., 2024) terbatas pada sektor kelistrikan, dan (Gurcan, 2024) pada emisi kendaraan bermotor. Sementara itu, studi yang mencakup Indonesia pada level nasional bersifat lintas negara, seperti (Begum & Mobin, 2025), atau tidak menerapkan seleksi fitur eksplisit, seperti (Pandi & Kharisudin, 2025). Akibatnya, perhatian terhadap evaluasi dampak seleksi fitur pada data makroekonomi nasional dengan jumlah observasi dan prediktor yang terbatas masih sangat sedikit. Pada konteks Indonesia, (Elmi et al., 2025) memprediksi emisi CO₂ nasional menggunakan pendekatan hybrid ARIMA-LSTM dengan pembagian data kronologis 80:20, namun berfokus pada akurasi peramalan deret waktu dan tidak menerapkan seleksi fitur maupun analisis kontribusi prediktif variabel. Hal ini menegaskan bahwa pendekatan yang mengombinasikan seleksi fitur RFE dan *ensemble learning* untuk menganalisis kontribusi prediktif variabel emisi karbon nasional Indonesia masih belum dieksplorasi. Penelitian ini mengusulkan alur penelitian yang mengintegrasikan RFE dengan *estimator* Random Forest sebagai tahap seleksi fitur dan XGBoost sebagai model prediksi, untuk mengevaluasi apakah seleksi fitur RFE dapat meningkatkan kinerja prediksi emisi karbon Indonesia dan mengidentifikasi variabel dengan kontribusi prediktif tertinggi.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang, permasalahan penelitian dapat diidentifikasi sebagai berikut:

1. Emisi karbon di Indonesia dipengaruhi oleh berbagai faktor, seperti aktivitas ekonomi, konsumsi energi, produksi listrik, jumlah penduduk, dan kondisi lingkungan, sehingga diperlukan pendekatan yang mampu menganalisis hubungan antar variabel tersebut.
2. Hubungan antara faktor-faktor yang mempengaruhi emisi karbon tidak selalu bersifat sederhana, sehingga diperlukan model prediksi yang mampu menangani pola data yang kompleks.
3. Penggunaan seluruh variabel prediktor tanpa proses seleksi fitur dapat menyulitkan interpretasi model, terutama dalam menentukan variabel mana yang paling dominan terhadap emisi karbon.

4. Penelitian yang secara khusus membandingkan performa XGBoost tanpa seleksi fitur dan XGBoost dengan seleksi fitur menggunakan RFE pada konteks prediksi emisi karbon di Indonesia masih terbatas.

1.3 Ruang Lingkup dan Batasan Penelitian

Penelitian ini dibatasi pada ruang lingkup dan aspek-aspek berikut untuk menjaga fokus, ketercapaian tujuan, dan kejelasan kontribusi ilmiah yang dihasilkan.

Dari sisi objek dan data, penelitian berfokus pada pemodelan prediktif emisi CO₂ per kapita Indonesia menggunakan data sekunder deret waktu tahunan periode 1990–2023 yang bersumber dari World Bank Open Data, Energy Institute Statistical Review of World Energy, dan World Bank Open Data (indikator EG.ELC.RNEW.ZS untuk EBT). Pemilihan periode ini didasarkan pada ketersediaan data yang konsisten untuk seluruh variabel prediktor. Sebanyak 34 observasi tahunan merupakan konsekuensi dari sifat data agregat nasional yang umumnya hanya tersedia dalam resolusi tahunan. Ukuran ini membatasi kompleksitas model yang dapat diterapkan dan menjadi pertimbangan utama dalam keputusan menggunakan konfigurasi parameter standar.

Dari sisi variabel, penelitian menggunakan lima fitur prediktor yang dipilih berdasarkan relevansi teoritis dan ketersediaan data konsisten: PDB per kapita (USD konstan 2015), konsumsi energi primer (TWh), produksi listrik berbasis EBT (%), luas lahan hutan (Km²), dan jumlah populasi (jiwa). Variabel prediktor dibatasi pada kelima indikator ini, sehingga faktor-faktor lain yang berpotensi relevan tidak dipertimbangkan dalam model.

Dari sisi metode dan lingkup komputasi, penelitian menggunakan RFE dengan *estimator* Random Forest sebagai metode seleksi fitur dan XGBoost dengan konfigurasi parameter standar sebagai model prediksi, dievaluasi menggunakan RMSE dan R² pada data uji. Model *Multiple Linear Regression* (MLR) disertakan sebagai pembanding sekunder, untuk menguji apakah hubungan antar variabel pada data ini cukup dijelaskan oleh model linier. Fokus utama penelitian tetap pada evaluasi dampak seleksi fitur RFE terhadap kinerja XGBoost.

Penelitian ini diposisikan sebagai *exploratory predictive modeling* dan analisis kontribusi prediktif fitur bukan pembuktian kausalitas, *deep sequential forecasting*, maupun generalisasi ke negara atau periode lain. Interpretasi *feature importance* bersifat prediktif dan asosiatif: fitur dengan kontribusi prediktif tinggi tidak dapat disimpulkan sebagai penyebab perubahan emisi karbon

1.4 Rumusan Masalah

Rumusan masalah penelitian ini adalah:

1. Variabel prediktor mana yang memiliki kontribusi prediktif tertinggi terhadap emisi karbon Indonesia berdasarkan hasil seleksi RFE dengan *estimator* Random Forest?
2. Bagaimana kinerja model prediksi emisi karbon XGBoost dengan fitur hasil seleksi RFE, diukur menggunakan RMSE dan R^2 pada data uji?
3. Apakah model XGBoost yang menggunakan fitur hasil seleksi RFE menghasilkan nilai RMSE yang lebih rendah dibandingkan model XGBoost *baseline* yang menggunakan seluruh fitur?

1.5 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengidentifikasi variabel prediktor yang memiliki kontribusi prediktif tertinggi terhadap emisi karbon Indonesia melalui proses seleksi fitur menggunakan RFE dengan *estimator* Random Forest.
2. Membangun dan mengevaluasi model prediksi emisi karbon berbasis XGBoost menggunakan fitur hasil seleksi RFE, dengan kinerja diukur melalui RMSE dan R^2 pada data uji.
3. Mengevaluasi apakah penerapan seleksi fitur RFE dapat meningkatkan kinerja model XGBoost dibandingkan model *baseline*, berdasarkan perbandingan nilai RMSE pada data uji.

1.6 Manfaat Penelitian

Manfaat Teoretis: Memberikan kontribusi metodologis pada penerapan RFE dan *ensemble learning* untuk pemodelan data multivariat berdimensi terbatas;

menjadi referensi bagi penelitian *machine learning* terapan pada domain lingkungan.

Manfaat Praktis: Menyediakan model prediktif berbasis data sebagai alat analitik untuk memahami kontribusi prediktif variabel makroekonomi dan energi terhadap emisi karbon Indonesia, yang dapat mendukung pengambilan keputusan berbasis data.

1.7 Kebaruan Penelitian (Novelty)

Penelitian prediksi emisi karbon berbasis *machine learning* telah berkembang pesat dalam beberapa tahun terakhir. Beberapa studi terdahulu yang paling relevan adalah (Begum & Mobin, 2025; Gurcan, 2024; Pandi & Kharisudin, 2025; Wang et al., 2025) dan (Cai et al., 2024). Masing-masing studi memberikan kontribusi metodologis yang berarti, namun sekaligus meninggalkan celah yang menjadi justifikasi penelitian ini.

(Pandi & Kharisudin, 2025) mengembangkan model CEEMDAN-LSTM dan CEEMDAN-GRU untuk peramalan emisi CO₂ Indonesia dan menunjukkan keefektifan model hybrid dekomposisi pada data deret waktu Indonesia. Meski relevan dari sisi konteks geografis, penelitian tersebut tidak menerapkan seleksi fitur eksplisit dan tidak bertujuan menganalisis kontribusi prediktif variabel predictor, dua aspek yang justru menjadi fokus utama penelitian ini.

(Wang et al., 2025) merupakan studi dengan kemiripan metodologis paling dekat dengan penelitian ini. Studi tersebut menerapkan strategi seleksi fitur hybrid yang menggabungkan analisis korelasi, penilaian kepentingan fitur berbasis model, dan RFE, diikuti pemodelan XGBoost untuk prediksi emisi karbon. Hasilnya menunjukkan bahwa seleksi fitur yang memuat RFE, dikombinasikan dengan XGBoost yang dioptimasi, menghasilkan akurasi prediksi yang tinggi. Meski demikian, terdapat perbedaan mendasar yang menjadikan penelitian ini memiliki kontribusi tersendiri. (Wang et al., 2025) menggunakan data operasional teknikal dari perusahaan pembangkit listrik, seperti beban listrik, efisiensi turbin, dan parameter operasional, pada skala mikro industri. Penelitian ini menerapkan kerangka serupa pada data makroekonomi dan energi nasional Indonesia dalam deret waktu tahunan jangka panjang. Perbedaan skala, jenis data, dan tujuan analisis

tersebut melahirkan pertanyaan penelitian yang berbeda: bukan “bagaimana memprediksi emisi operasional secara akurat”, melainkan “variabel makroekonomi dan energi mana yang memiliki kontribusi prediktif tertinggi terhadap tren emisi karbon nasional Indonesia”. Penelitian ini dengan demikian mengadopsi metode yang telah terbukti valid untuk menjawab pertanyaan yang belum terjawab pada konteks yang berbeda secara substansial.

(Begum & Mobin, 2025) menggunakan pendekatan *machine learning* multivariat dengan *feature selection* pada data emisi sebelas negara penghasil emisi terbesar, termasuk Indonesia, dan menunjukkan bahwa pendekatan *machine learning* dengan seleksi fitur mampu menghasilkan akurasi yang lebih baik dibandingkan model tanpa seleksi fitur. Meski mencakup Indonesia, studi tersebut bersifat lintas negara dan tidak dirancang untuk menganalisis kontribusi prediktif fitur secara spesifik pada konteks nasional Indonesia.

(Gurcan, 2024) melakukan studi komparatif komprehensif menggunakan 18 algoritma termasuk XGBoost pada data emisi CO₂ kendaraan bermotor dan membuktikan keunggulan *ensemble learning*, namun tidak menerapkan seleksi fitur sistematis dan berfokus pada domain yang berbeda dari emisi nasional.

(Cai et al., 2024) mengembangkan model hybrid Optuna-LightGBM-XGBoost untuk estimasi emisi karbon berbasis data ketenagalistrikan pada 25 perusahaan, namun terbatas pada sektor kelistrikan dan tidak mengidentifikasi kontribusi prediktif variabel makroekonomi.

Berdasarkan kajian tersebut, terdapat tiga celah metodologis yang belum terjawab secara bersamaan:

- Penerapan RFE dengan *estimator* Random Forest secara eksplisit pada data makroekonomi dan energi nasional;
- Perbandingan langsung kinerja model XGBoost dengan dan tanpa seleksi fitur pada konteks data nasional Indonesia;
- Analisis kontribusi prediktif fitur yang mudah diinterpretasikan untuk keperluan analitik. Penelitian ini dirancang untuk mengisi ketiga celah tersebut.