MODEL PEMBELAJARAN DAN LAPORAN AKHIR
PROJECT-BASED LEARNING
MATA KULIAH MACHINE LEARNING I
KELAS B

"MODEL PREDIKSI TINGKAT EMISI KARBON DIOKSIDA: PERBANDINGAN ANTARA ALGORITMA RANDOM FOREST DAN XGBOOST"

DISUSUN OLEH KELOMPOK II :

1. ANGELA LISANTHONI (21083010032) - KETUA
2. FITRI INDAH SARI (21083010025) - ANGGOTA
3. ELLEXIA LEONIE GUNAWAN (21083010027) - ANGGOTA
4. CHELSEA AYU ADHIGIADANY (21083010028) - ANGGOTA

DOSEN PENGAMPU:

DR. ENG. IR. ANGGRAINI PUSPITA SARI, ST., MT.
AVIOLLA TERSA DAMALIANA, S.SI, M.STAT

PROGRAM STUDI SAINS DATA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR
2023

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Allah SWT atas segala rahmat, hidayah, serta karunia-Nya yang telah melimpah pada penulis sehingga dapat menyelesaikan laporan tugas akhir ini. Laporan tugas akhir ini disusun sebagai salah satu syarat untuk menyelesaikan mata kuliah Machine Learning I di Universitas Pembangunan Nasional "Veteran" Jawa Timur.

Penulisan laporan tugas akhir ini tidak lepas dari bimbingan, dukungan, dan motivasi dari berbagai pihak yang dengan ikhlas memberikan kontribusi dan bantuannya. Oleh karena itu, melalui kesempatan ini, kami ingin mengucapkan terima kasih yang tak terhingga kepada:

- Ibu Dr. Eng. Ir. Anggraini Puspita Sari, ST., MT. atas bimbingan, arahan, serta pengarahannya yang telah membantu penulis dalam menyelesaikan laporan tugas akhir ini.
- Ibu Aviolla Tersa Damaliana, S.Si, M.Stat atas bimbingan, arahan, serta pengarahannya yang telah membantu penulis dalam menyelesaikan laporan tugas akhir ini.
- Orang tua dan keluarga kami atas dukungan, doa, dan semangat yang terus diberikan dalam perjalanan penulisan laporan tugas akhir ini.
- dan kepada seluruh pihak yang telah memberikan masukan, saran, serta sumbangsihnya dalam penulisan laporan tugas akhir ini.

Laporan tugas akhir ini tidak luput dari kekurangan, oleh karena itu segala saran dan kritik yang membangun sangat kami harapkan demi perbaikan di masa mendatang. Akhir kata, semoga laporan tugas akhir ini dapat memberikan manfaat serta menjadi kontribusi yang bermanfaat bagi perkembangan ilmu pengetahuan di bidang sains data.

Surabaya, 16 Juni 2023

Penulis

ABSTRACT

Global warming is a problem that needs to be aware due to its negative impacts in all areas and even affects the increase in world temperature that is currently occurring. One of the factors is the level of CO_2 emissions in the air. Therefore, it is necessary to predict accurate CO_2 emission to develop the right strategy to reduce the level of CO_2 emission. So, the aim of this research is making a comparison between the Random Forest and XGBoost algorithms which are one of the forecasting models forming algorithms. The results will be compared based on three types of evaluation parameters including the coefficient of determination (R^2), Mean Absolute Error (MAE), dan Root Mean Square Error (RMSE). Based on the research, it proves that there is no significant difference in predicting CO_2 emission between the Random Forest and XGBoost algorithms. The two algorithms being compared have an accuracy of more than 95%, however, XGBoost has the advantage that the time required for prediction is 77.46% faster than Random Forest. By considering accuracy and efficiency, XGBoost is the right choice in forming a CO_2 emission prediction model.

Keywords: Regression Supervised Learning, CO_2 Emission Prediction, Random Forest, XGBoost

ABSTRAK

Global warming menjadi masalah yang perlu diwaspadai karena berdampak negatif di segala bidang dan bahkan berdampak pada peningkatan suhu dunia yang kini sedang terjadi. Salah satu faktornya adalah tingkat emisi CO_2 di udara. Oleh sebab itu, diperlukan prediksi emisi CO_2yang akurat untuk menyusun strategi yang tepat guna mengurangi tingkat emisi CO_2. Sehingga tujuan dari penelitian ini adalah membandingkan algoritma *Random Forest* dan *XGBoost* yang merupakan salah satu algoritma pembentuk model prediksi. Hasil akan dibandingkan berdasarkan tiga jenis evaluasi parameter diantaranya koefisien determinasi (R^2), *Mean Absolute Error* (MAE), dan *Root Mean Square Error* (RMSE). Berdasarkan penelitian yang dilakukan, menunjukkan bahwa tidak ada perbedaan signifikan dalam melakukan prediksi emisi CO_2 antara algoritma *Random Forest* dan *XGBoost*. Kedua algoritma yang dibandingkan memiliki hasil akurasi lebih dari 95% namun, *XGBoost* memiliki kelebihan yakni waktu yang dibutuhkan dalam prediksi 77.46% lebih cepat dibanding *Random Forest*. Dengan mempertimbangkan akurasi dan efisiensi, algoritma *XGBoost* adalah pilihan yang tepat dalam membangun model prediksi emisi CO_2.

Kata kunci: Regresi Supervised Learning, Prediksi Emisi CO2, *Random Forest*, *XGBoost*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I: PENDAHULUAN

1.1 Latar Belakang

Global warming yang disebabkan oleh peningkatan konsentrasi gas rumah kaca di atmosfer, menjadi salah satu masalah yang perlu diwaspadai saat ini. Dampak dari global warming meliputi kenaikan suhu rata-rata bumi, perubahan pola cuaca yang ekstrem, pencairan es di kutub, kenaikan permukaan air laut, dan ancaman serius terhadap keanekaragaman hayati. Salah satu faktor utama yang menyebabkan *global warming* adalah tingkat emisi CO_2 yang tinggi di atmosfer (Turrentine, 2021).

Data dari *Our World in Data* menunjukkan bahwa pada tahun 2021, total emisi CO_2 mencapai angka mencengangkan sebesar 37,12 miliar ton. Bahkan lebih mengkhawatirkan, perkiraan menunjukkan bahwa angka ini cenderung terus meningkat seiring berjalannya waktu. Saat ini, sekitar 34 miliar ton CO_2 dilepaskan ke atmosfer setiap tahunnya (Ritchie, Roser, & Rosado, 2020).

Salah satu sektor yang menjadi penyumbang tingkat emisi CO_2 adalah sektor transportasi (Ritchie, Co2 emissions from transport, 2020). Menurut data *dari International Energy Agency* (IEA), sektor transportasi menyumbang 1/5 dari total emisi CO_2 secara global (Teter, 2022). Pada akhir tahun 2021, sekitar 37% dari total emisi CO_2 disumbangkan oleh kendaraan. Di Indonesia, menurut informasi dari *Climate Transparency*, emisi dari transportasi menyumbang sekitar 27% dari total emisi CO_2 karena sektor ini didominasi oleh bahan bakar fosil pada tahun 2019 (Tumiwa & Wijayani, 2021).

Dalam upaya untuk mengurangi emisi CO_2, perlu dilakukan prediksi yang akurat mengenai tingkat emisi CO_2 di masa depan. Dengan memahami dan memprediksi tingkat emisi CO_2 yang dihasilkan oleh sektor transportasi, dapat diambil tindakan yang tepat dalam merancang dan menerapkan kebijakan dan strategi mitigasi yang efektif. Prediksi yang akurat dapat membantu merencanakan infrastruktur transportasi yang berkelanjutan, mendorong penggunaan kendaraan ramah lingkungan, dan mempromosikan inovasi teknologi yang lebih efisien dan bersih, sehingga tingkat emisi CO_2 dari sektor transportasi juga akan berkurang (C2ES, 2008).

Dalam hal ini, teknik machine learning telah menjadi alat yang kuat dalam memprediksi emisi CO_2. Dengan menggunakan model algoritma yang efektif, seperti *Random Forest* dan *XGBoost*, dapat dilakukan prediksi yang akurat tentang tingkat

emisi CO_2 di masa depan. *Random Forest* dan *XGBoost* adalah dua model algoritma populer dalam dunia *machine learning* yang telah terbukti memberikan performa yang baik dalam berbagai aplikasi prediksi (Allwright, 2023).

1.2 Permasalahan

- Minimnya tingkat kesadaran masyarakat dalam mengurangi penggunaan kendaraan pribadi sebagai pemicu emisi CO_2
- Belum tersedianya model prediksi yang tepat dalam pengurangan emisi CO_2 dari sektor transportasi

1.3 Tujuan

- Untuk meningkatkan tingkat kesadaran masyarakat terhadap bahayanya emisi CO_2 yang berasal dari polusi kendaraan
- Tersedianya model prediksi yang tepat dalam pengurangan emisi CO_2 dari sektor transportasi

1.4 Manfaat

- Manfaat bagi pemerintah dan lembaga terkait

  Penelitian ini dapat memberikan data dan informasi yang dapat digunakan oleh pemerintah dan lembaga terkait untuk mengembangkan kebijakan dan strategi pengurangan emisi CO2 dalam sektor transportasi. Hal ini dapat membantu dalam merumuskan langkah-langkah konkret untuk mengurangi dampak negatif terhadap lingkungan dan kesehatan masyarakat.
- Manfaat bagi masyarakat umum

  Informasi dari penelitian ini dapat meningkatkan kesadaran masyarakat tentang dampak penggunaan kendaraan pribadi terhadap emisi CO2 dan perubahan iklim. Hal ini dapat mendorong masyarakat untuk mengadopsi kebiasaan transportasi yang lebih berkelanjutan, seperti menggunakan transportasi umum, bersepeda, atau berjalan kaki.
- Manfaat bagi peneliti

  Penelitian ini dapat menjadi kontribusi terhadap pengetahuan dan pemahaman ilmiah tentang hubungan antara penggunaan kendaraan pribadi dan emisi CO2.

Hasil penelitian ini juga dapat menjadi landasan untuk penelitian lanjutan dalam bidang transportasi berkelanjutan dan mitigasi perubahan iklim.

- Manfaat bagi organisasi lingkungan

  Hasil penelitian ini dapat menjadi dasar untuk advokasi dan kampanye organisasi lingkungan dalam mendorong pengurangan penggunaan kendaraan pribadi dan promosi alternatif transportasi yang ramah lingkungan.

- Manfaat bagi industri transportasi

  Penelitian ini dapat memberikan wawasan bagi perusahaan dan industri transportasi tentang pentingnya mengurangi emisi CO2 dalam operasional kendaraan. Hal ini dapat mendorong pengembangan dan implementasi solusi transportasi yang lebih efisien dan ramah lingkungan.

BAB II: TINJAUAN PUSTAKA

2.1 Teori Penunjang

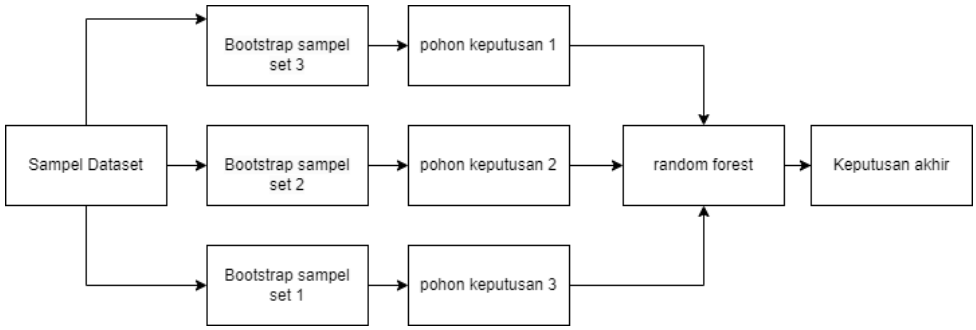
2.1.1 Emisi Karbon Dioksida (CO_2)

Emisi karbon dioksida (CO_2) adalah emisi gas rumah kaca yang menjadi faktor utama penyebab fenomena pemanasan global (Labiba & Pradoto, 2018). Emisi CO_2 dapat berasal dari berbagai sumber, termasuk pembakaran bahan bakar fosil (seperti batu bara, minyak, dan gas alam) dalam industri, transportasi, dan pembangkit listrik. Sumber lainnya termasuk deforestasi, perubahan penggunaan lahan, proses industri, dan kegiatan pertanian. Emisi CO_2 yang tinggi berkontribusi terhadap pemanasan global dan dampak negatifnya, termasuk perubahan suhu ekstrem, kenaikan permukaan laut, perubahan pola curah hujan, kekeringan, banjir, dan ancaman terhadap keanekaragaman hayati.

Hal ini juga dapat mempengaruhi sektor pertanian, ketersediaan air, dan kesehatan manusia. Emisi CO_2 yang tinggi berkontribusi terhadap pemanasan global dan dampak negatifnya, termasuk perubahan suhu ekstrem, kenaikan permukaan laut, perubahan pola curah hujan, kekeringan, banjir, dan ancaman terhadap keanekaragaman hayati. Hal ini juga dapat mempengaruhi sektor pertanian, ketersediaan air, dan kesehatan manusia.

2.1.2 Algoritma *Random Forest*

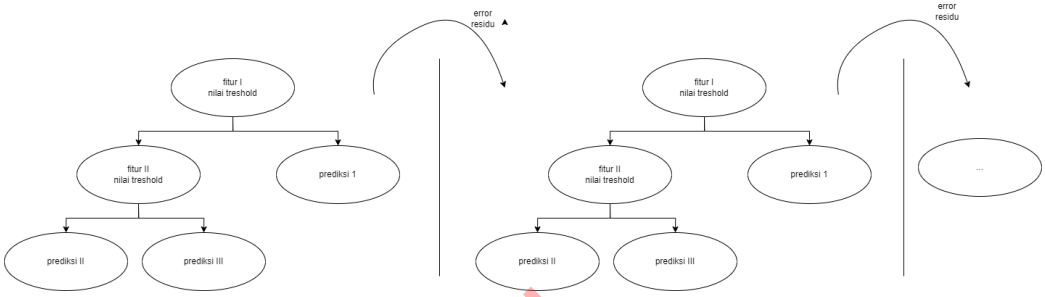
Algoritma *Random Forest* dapat digunakan untuk melakukan regresi maupun klasifikasi tergantung tujuan analisis yang diinginkan. Jika digunakan untuk melakukan regresi, maka hasil yang ditampilkan adalah rata - rata dari pohon yang berbeda (Alam, Farid, & Rossetti, 2018).



Gambar 1 *Proses algoritma Random Forest*

Gambar 1 menunjukkan ilustrasi algoritma *Random Forest* berjalan. Data akan diproses melalui metode *bootstrap* untuk menemukan k sampel data yang berbeda. Di setiap *bootstrap*, terdapat pohon keputusan yang dilakukan secara berulang untuk memecah data menjadi dua kelompok dengan kriteria tertentu. Setiap pohon akan memberikan keputusan hingga didapatkan k keputusan yang berbeda. Kemudian, diambil satu keputusan sebagai keputusan akhir berdasarkan suara *voting* terbanyak (Wu & Liu, 2017).

2.1.3 *Algoritma eXtreme Gradient Boosting (XGBoost)*



Gambar 2 *Proses algoritma XGBoost*

Algoritma *XGBoost* mengimplementasikan pengembangan dari pohon keputusan dan menerapkan teknik *ansambel* artinya model akan terus diperbarui untuk memperbaiki kesalahan pada model sebelumnya. Algoritma ini dibuat dalam rangka meningkatkan efisien waktu pemroses termasuk automasi menangani data hilang, melakukan pohon keputusan secara paralel, serta dapat melakukan training data terus - menerus untuk meningkatkan hasil akurasi (Brownlee, 2018). Ilustrasi proses algoritma *XGBoost* ditampilkan pada gambar 2 (Osman, Ahmed, Chow, Huang, & El-Shafie, 2021). Algoritma ini dirumuskan sebagai berikut:

$$obj(\theta) = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad f_k \epsilon F \tag{1}$$

L menyatakan perbedaan standar deviasi antara nilai prediksi $\hat{y}_i$ dan nilai asli y_i. Ω menyatakan fungsi regularisasi kompleksitas model untuk menghindari overfitting. f menyatakan fungsi dalam ruang fungsional F dan F menyatakan himpunan semua pohon yang dibuat (Luo, Zhang, Fu, & Rao, 2021).

2.1.4 Evaluasi Parameter

Untuk melakukan evaluasi terhadap model, terdapat tiga parameter yang akan dibandingkan yaitu koefisien determinasi R^2 yang digunakan untuk menafsir kelayakan model dengan rentang nilai $0 \leq R^2 \leq 1$. *Root Mean Square Error* (RMSE) yang merupakan perbedaan nilai prediksi dengan nilai sebenarnya dan akar dari *Mean Square Error* (MSE), serta *Mean Absolute Error* (MAE) yang merupakan rata - rata kesalahan mutlak antara nilai prediksi dengan nilai yang sebenarnya. Ketiga parameter dirumuskan sebagai berikut (Luo, Zhang, Fu, & Rao, 2021) (Putra & Juarna, 2021):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(\hat{y}_1))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (4)$$

2.1.5 *Python*

Python adalah bahasa pemrograman yang sangat populer dan kuat yang digunakan secara luas dalam pengembangan aplikasi dan analisis data, termasuk dalam bidang machine learning. Hal ini dikarenakan *Python* menyediakan banyak *library* dan alat untuk manipulasi, visualisasi, dan *preprocessing* data (Manalu & Gunadi, 2022). *Pandas*, *NumPy*, dan *Matplotlib* adalah beberapa contoh *library* yang sering digunakan dalam pengolahan data untuk *machine learning*. Mereka menyediakan fungsi dan metode yang efisien untuk membaca, mengubah, dan menganalisis data sebelum membangun model.

2.1.6 *Open Canada*

Open Canada adalah sebuah inisiatif pemerintah Kanada untuk mempromosikan transparansi dan akuntabilitas pemerintah. Salah satu program yang diluncurkan oleh *Open Canada* adalah *Open Data*. *Open Data* adalah sebuah portal yang menyediakan akses terbuka ke data-data pemerintah Kanada yang relevan bagi masyarakat Kanada. Di dalam portal ini, masyarakat dapat mencari data yang mereka butuhkan, belajar cara bekerja dengan dataset, dan melihat apa yang telah dilakukan oleh orang-orang dengan data terbuka di seluruh negeri (Canada.ca, 2023).

2.2 Penelitian Terkait

2.2.1 *Forecasting of CO2 Emissions in Iran Based On Time Series and Regression Analysis* (Hosseini, Saifoddin, Shirmohammadi, & Aslani, 2019)

Penelitian ini memprediksi emisi CO_2 di Iran pada tahun 2030 dengan asumsi dua skenario, yaitu *business as usual* (BAU) dan Rencana Pembangunan Keenam (SDP), dengan menggunakan analisis regresi linier berganda (MLR) dan regresi polinomial berganda (MPR). Dari hasil penelitian didapatkan bahwa skenario BAU akan menghasilkan peningkatan emisi CO_2 Iran sebesar 30% pada tahun 2030, sementara implementasi SDP dapat menghasilkan pengurangan sekitar 60%. Temuan menunjukkan bahwa Iran kemungkinan besar tidak akan memenuhi komitmennya terhadap Perjanjian Paris berdasarkan asumsi BAU namun, implementasi penuh dari SDP yang dibentuk secara ambisius dapat memenuhi target pada akhir 2018.

2.2.2 *Knowledge-Based Machine Learning Techniques For Accurate Prediction of CO2 Storage Performance in Underground Saline Aquifers* (Thanh, Yasin, Al-Mudhafar, & Lee, 2022)

Penelitian ini membahas tentang penyimpanan karbon dioksida dalam akuifer garam bawah tanah yang dianggap sebagai teknik yang menjanjikan untuk mengurangi emisi CO_2 di atmosfer. Dalam penelitian ini, efisiensi penangkapan sekuestrasi CO_2 dalam formasi garam diprediksi dengan mengembangkan tiga model berbasis *Machine Learning* (ML) yang diawasi: *Random Forest* (RF), *eXtreme Gradient Boosting* (XGBoost), dan *Support Vector Regression* (SVR). Dari hasil prediksi, model ML yang diusulkan diberi peringkat berdasarkan akurasinya: *XGBoost* > RF > SVR. Model prediktif berbasis *XGBoost* mencapai kesalahan kuadrat rata-rata akar yang sangat rendah (RMSE = 0,0041) dan faktor korelasi tinggi (R2 = 0,9993) untuk efisiensi penjebakan residu dan kelarutan. Namun, RF dan SVR menunjukkan nilai RMSE 0,0243 dan 0,074 dan R^2 masing-masing 0,9781 dan 0,9284. Selain itu, penerapan model *XGBoost* telah divalidasi dan hanya 15 titik data yang diduga terdeteksi di seluruh database. Oleh karena itu, model *XGBoost* dapat digunakan sebagai alat penyaringan dan perencanaan proses yang kuat untuk penilaian ketidakpastian proyek penyimpanan karbon.

2.2.3 *Predicting The Carbon Dioxide Emission Caused by Road Transport Using A Random Forest (RF) Model Combined by Meta-Heuristic Algorithms* (Khajavi & Rastgoo, 2023)

Penelitian ini memprediksi emisi CO_2 dari 30 kota besar di Cina dengan metode *Random Forest*, *Support Vector Regression*, dan *Response Surface*. Akurasi metode tersebut dibandingkan melalui *Standard Error* (SE), Root *Mean Square Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), Mean *Absolute Error* (MAE), *Relative Absolute Error* (RAE), dan koefisien determinasi (R^2). Hasil yang diperoleh menunjukkan bahwa regresi vektor dukungan dengan pengoptimal Harris Hawk memiliki akurasi terbaik dalam proses pelatihan dengan nilai R^2 sebesar 0,9999, dan *Random Forest* dengan Algoritma *Slime Mould* dengan nilai R^2 sebesar 0,9641 memiliki akurasi terbaik dalam pengujian proses. Oleh karena itu, *Random Forest with Slime Mold Algorithm* (RF-SMA) merupakan metode terbaik untuk memprediksi emisi CO_2.

BAB III: METODOLOGI PENELITIAN

Gambar 3 *Alur penelitian*

3.1 Data

Pada proyek ini, kami menggunakan data *CO2 Emission by Vehicle* yang diambil dari situs data terbuka resmi Pemerintah Kanada. Dataset ini mencatat detail tentang bagaimana emisi CO_2 oleh sebuah kendaraan dapat bervariasi dengan fitur-fitur yang berbeda selama 7 tahun. Dataset ini memiliki 12 kolom dan juga 7385 baris. Berikut adalah kolom-kolom yang terdapat pada dataset:

- *CO2 Emissions yaitu emisi CO2 (g/km)* untuk mengemudi gabungan di kota dan jalan raya
- *Make* yaitu merek atau *brand* dari kendaraan
- *Vehicle Class* yaitu kategori kendaraan berdasarkan ukuran, jenis, dan tujuan penggunaan
- *Engine Size (L)* yaitu ukuran mesin yang merujuk pada volume mesin kendaraan dalam liter (L)
- *Cylinders* yaitu jumlah silinder dalam mesin kendaraan
- *Model* yaitu model-model/ jenis-jenis kendaraan. Ada beberapa jenis kendaraan sebagai berikut:
  - 4WD/4X4 (*Four-wheel drive*) = kendaraan dengan sistem penggerak 4 roda.
  - AWD (*All-wheel drive*) = kendaraan dengan sistem penggerak roda empat secara otomatis yang dapat mendistribusikan daya secara otomatis di antara roda-roda yang dibutuhkan.
  - FFV (*Flexible-fuel vehicle*) = kendaraan yang dapat menggunakan lebih dari satu jenis bahan bakar.
  - SWB (*Short wheelbase*) = kendaraan dengan jarak antara roda depan dan roda belakang yang pendek.
  - LWB (*Long wheelbase*) = kendaraan dengan jarak antara roda depan dan roda belakang yang lebih panjang.
  - EWB (*Extended wheelbase*) = kendaraan dengan jarak antara roda depan dan roda belakang yang lebih panjang dari *Long Wheelbase*.

- *Transmission* yaitu jenis transmisi yang digunakan kendaraan. Ada beberapa jenis transmisi sebagai berikut:
  - “A" (*Automatic*) : Transmisi otomatis, yang mengatur perpindahan gigi secara otomatis tanpa intervensi pengemudi.
  - "M" (*Manual*) : Transmisi manual, yang mengharuskan pengemudi mengoperasikan kopling dan mengubah gigi secara manual dengan tuas gigi atau paddle-shift di kemudi.
  - "AV" (*Continuously variable*) : Transmisi variabel kontinu, yang mengatur perpindahan gigi dengan cara yang lebih halus dan tanpa adanya langkah yang terlihat secara jelas.
  - "AM" (*Automated manual*) : Transmisi manual otomatis, yang mengubah gigi secara otomatis, tetapi masih mengharuskan pengemudi mengoperasikan kopling secara manual.
  - "AS" (*Automatic with select shift*) : Transmisi otomatis yang memungkinkan pengemudi memilih perpindahan gigi secara manual dengan tuas atau *paddle-shift* di kemudi.
  - Angka 3 hingga 10 : Jumlah gigi pada transmisi manual atau otomatis, yang menentukan rasio perpindahan gigi yang berbeda untuk kecepatan yang berbeda.
- *Fuel Type* yaitu jenis bahan bakar yang digunakan kendaraan. Ada beberapa jenis bahan bakar sebagai berikut:
  - "X" (*Regular gasoline*) : Bensin biasa atau tanpa timbal (*unleaded gasoline*).
  - "Z" (*Premium gasoline*) : Bensin berkualitas lebih tinggi yang umumnya memiliki oktan yang lebih tinggi dari bensin biasa.
  - "D" (*Diesel*) : Bahan bakar minyak yang digunakan pada mesin diesel, yang biasanya lebih efisien daripada bensin tetapi menghasilkan lebih banyak emisi NOx dan partikulat.
  - "E" (*Ethanol E85*) : Bahan bakar campuran yang terdiri dari 85% etanol dan 15% bensin, yang biasanya digunakan pada kendaraan fleksibel bahan bakar (FFV).
  - "N" (*Natural gas*) : Bahan bakar alternatif yang terdiri dari gas alam terkompresi (CNG) atau gas alam terkondensasi (LNG), yang umumnya lebih bersih dan efisien daripada bahan bakar fosil.

- *Fuel Consumption City (L/100KM)* yaitu konsumsi bahan bakar di kota dalam liter per 100 kilometer
- *Fuel Consumption HWY (L/100KM)* yaitu konsumsi bahan bakar di jalan raya dalam liter per 100 kilometer
- *Fuel Consumption COMB (L/100KM)* yaitu kombinasi nilai konsumsi bahan bakar di kota (55%) dan di jalan raya (45%) dalam liter per 100 kilometer
- *Fuel Consumption COMB (mpg)* yaitu kombinasi nilai konsumsi bahan bakar di kota (55%) dan di jalan raya (45%) dalam mil per gallon

3.2 Pre-Processing Data

3.2.1 Duplikasi Data

Data duplikasi dan data rangkap merupakan kondisi dimana satu atribut memiliki dua atau lebih nilai yang sama, sehingga perlu dihapus agar akurasi dari model meningkat. Oleh sebab itu, pada tahap ini kami ingin mengecek dan menghapus duplikasi yang ada pada data CO_2 *Emissions*. Ditemukan 1103 data duplikat yang kemudian dihapus dan didapatkan data 6282 baris dan 12 kolom.

3.2.2 *Missing Value*

Missing Value adalah hilangnya beberapa data yang telah diperoleh. Penyebab umum missing value antara lain kesalahan pengisian formulir, kesalahan pengukuran, atau sifat alami data yang tidak diketahui atau tidak dapat diukur. *Missing value* dapat memiliki dampak yang signifikan dalam analisis data karena dapat mengganggu keakuratan dan validitas hasil yang dihasilkan. Oleh karena itu, pada proyek ini kami ingin memastikan dan menghapus baris dan kolom yang terdapat *missing value*. Hasil tidak menunjukkan adanya missing value, sehingga tidak ada perubahan dalam data.

3.2.3 *Feature Selection*

Feature selection adalah proses pemilihan subset relevan dari fitur atau variabel yang ada dalam kumpulan data yang akan digunakan dalam proses analisis atau pemodelan. Tujuan utama dari *feature selection* adalah mengidentifikasi dan memilih fitur yang paling informatif, relevan, dan memiliki dampak signifikan terhadap prediksi atau pemahaman yang diinginkan, serta

mengurangi kompleksitas dan menghindari *overfitting* dalam model. Pada studi kasus ini, jumlah kolom yang semula 12 dikurangi menjadi 7 kolom dengan 6282 baris.

3.2.4 Data *Outlier*

Data *Outlier* yaitu nilai observasi yang secara signifikan berbeda dari pola umum atau sebagian besar data dalam sebuah kumpulan data. *Outlier* dapat menjadi nilai yang jauh lebih tinggi atau lebih rendah dari nilai-nilai lain dalam distribusi data. *Outlier* memiliki potensi untuk mempengaruhi analisis dan model statistik dengan mempengaruhi estimasi dan kesimpulan yang diambil dari data sehingga perlu dilakukan modifikasi atau penghapusan.

Salah satu cara untuk mengidentifikasi data *outlier*, dapat dilakukan dengan metode statistik IQR (*Interquartile Range*). Ada beberapa cara untuk menangani data *outlier* seperti, menghapus outlier dari analisis, mengubah nilai *outlier* dengan nilai yang lebih tepat, atau menggunakan teknik *robust* yang kurang sensitif terhadap *outlier* dalam analisis. Dalam studi kasus ini, nilai *outlier* akan dicek untuk seluruh kolom dan nilai yang teridentifikasi sebagai *outlier* akan dihapus. Sehingga data berkurang menjadi 5816 baris dan 7 kolom.

3.3 Korelasi dan Visualisasi Data

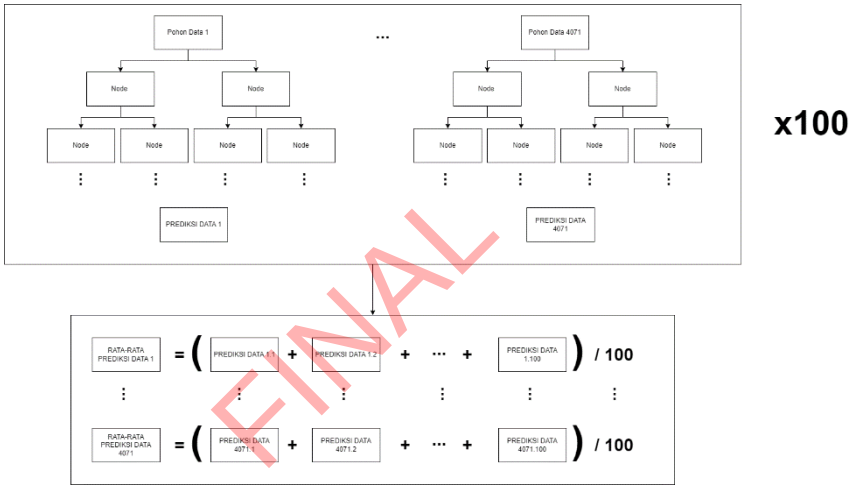
Setelah data melalui tahap pre-processing, peneliti akan melihat korelasi atau seberapa keterkaitan variabel fitur dengan variabel target. Korelasi akan dilihat menggunakan visualisasi heatmap dengan memanfaatkan library seaborn. Jika tingkat korelasi positif maupun negatif tinggi, maka terbukti bahwa variabel fitur yang digunakan berkaitan erat dengan variabel target.

Selanjutnya, untuk mengecek kenormalitasan data, peneliti menggunakan visualisasi histogram dari *Python*. Dari hasil visualisasi tersebut, kita dapat melihat apakah persebaran data berdistribusi normal atau tidak. Jika ternyata data tidak berdistribusi normal, maka kita perlu memproses ulang data. Namun jika data berdistribusi normal, maka kita dapat melanjutkan ke tahap berikutnya.

3.4 *Splitting*, *Modelling*, dan *Evaluasi*

Sebelum dilakukan proses *splitting*, variabel fitur terlebih dahulu akan distandarisasi menggunakan metode z-score sehingga *range* dari variabel fitur sama. *Splitting* data yaitu membagi data menjadi data latih dan data uji. Pada proyek ini, peneliti menggunakan rasio 70:30, artinya data secara acak akan dibagi 70% untuk data latih dan 30% untuk data uji. Sehingga ada 4071 baris dan 6 kolom sebagai data latih sedangkan ada 1745 baris dan 6 kolom sebagai data training. Data dibagi secara acak untuk menghindari bias pada data. Setelah melakukan *splitting* data, maka kita dapat melatih data sesuai dengan algoritma yang ingin digunakan. Pada proyek ini, peneliti ingin melatih data dengan dua algoritma diantaranya:

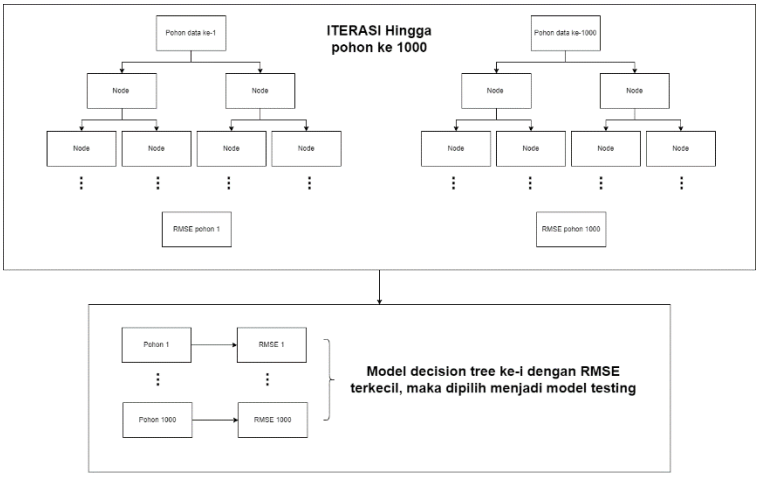
3.4.1 *Random Forest*



Gambar 4 *Proses Random Forest*

Gambar 4 menunjukkan bagaimana algoritma *Random Forest* akan diimplementasikan dalam dataset yang digunakan dalam studi kasus ini. *Random Forest* adalah algoritma yang berbasis *decision tree* oleh sebab itu, saat melatih model dengan 4071 data *training*, akan ada total 4071 x 100 (nilai parameter yang diinputkan) pohon keputusan yang dibuat. Sebagai gambarannya, pada data pertama, terdapat 100 pohon keputusan yang menghasilkan 100 hasil regresi yang berbeda. Dari 100 hasil tersebut akan dihitung rata – ratanya sebagai keputusan hasil regresinya. Hal ini juga berlaku hingga 4071 data.

3.4.2 *XGBoost*



***Gambar 5** Proses XGBoost*

Gambar 5 menunjukkan bagaimana algoritma *XGBoost* akan diimplementasikan dalam dataset yang digunakan dalam studi kasus ini. XGBoost adalah algoritma yang menggunakan Teknik ansamble sehingga akan ada iterasi pohon keputusan sebanyak 1000x (berdasarkan nilai parameter yang diinputkan). Dalam pemilihan model pohon keputusan terbaik, akan diambil pohon keputusan yang memiliki total nilai RMSE terkecil dan model tersebut yang akan digunakan dalam data testing. Jika dalam iterasi ke-10, tidak ada peningkatan dalam RMSEnya maka proses iterasi akan dihentikan.

3.5 Membandingkan Hasil Algoritma

Peneliti membandingkan hasil algoritma dengan mengamati beberapa evaluasi parameter, yaitu koefisien determinasi, RMSE, dan MAE, serta berapa lama waktu yang dibutuhkan dalam pemrosesan berjalan. Dengan membandingkan beberapa nilai tersebut, kami berharap dapat memilih algoritma yang terbaik untuk proyek ini. Selain membandingkan beberapa nilai tersebut, kami juga akan memvisualisasikan secara grafik antara data prediksi dan data aktual. Sehingga dapat lebih mudah dilihat keakuratan yang dihasilkan dari setiap algoritma.

BAB IV: HASIL DAN PEMBAHASAN

4.1 *Pre-processing*

4.1.1 Duplikasi Data

```
df.duplicated().sum()
df.drop_duplicates(inplace=True)
```

Gambar 6 *Potongan script untuk mengecek dan menghapus duplikasi data*

Gambar 6 menunjukkan code yang digunakan untuk mengecek dan menghapus duplikasi data. Dalam dataset, ditemukan terdapat 1103 duplikasi data yang kemudian dihapus untuk meningkatkan akurasinya. Sehingga, data sekarang tersisa 6282 data.

4.1.2 Mengecek *Missing Value*

```
#Cek nilai Null
print(df.isna().sum())
```

Gambar 7 *Potongan script untuk mengecek nilai null*


```
Make                              0
Model                             0
Vehicle Class                     0
Engine Size(L)                    0
Cylinders                         0
Transmission                      0
Fuel Type                         0
Fuel Consumption City (L/100 km)  0
Fuel Consumption Hwy (L/100 km)   0
Fuel Consumption Comb (L/100 km)  0
Fuel Consumption Comb (mpg)       0
CO2 Emissions(g/km)               0
dtype: int64
```

Gambar 8 *Hasil pengecekan nilai null*

Pada setiap kolom dataset, tidak ditemukan nilai null yang dibuktikan dengan angka 0. Sehingga, tidak perlu dilakukan proses penghapusan nilai null.

4.1.3 Feature Selection

```
df = df.drop(['Make', 'Model', 'Vehicle Class',
'Transmission','Fuel Type'], axis = 1)
```

Gambar 9 *Potongan script untuk menghapus kolom*

Peneliti menduga bahwa kolom 'Make', 'Model', 'Vehicle Class', 'Transmission','Fuel Type' tidak memiliki dampak atau pengaruh yang cukup

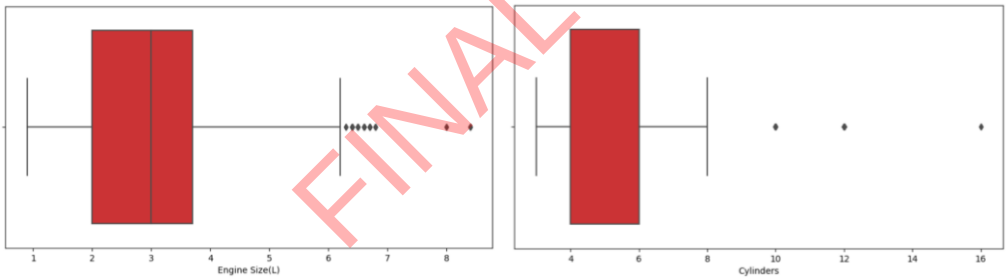
signifikan terhadap kolom target. Oleh sebab itu, kolom 'Make', 'Model', 'Vehicle Class', 'Transmission','Fuel Type' dihapus. Dataset sekarang memiliki 6 kolom fitur dan 1 kolom target.

4.1.4 Menghapus *Outlier*

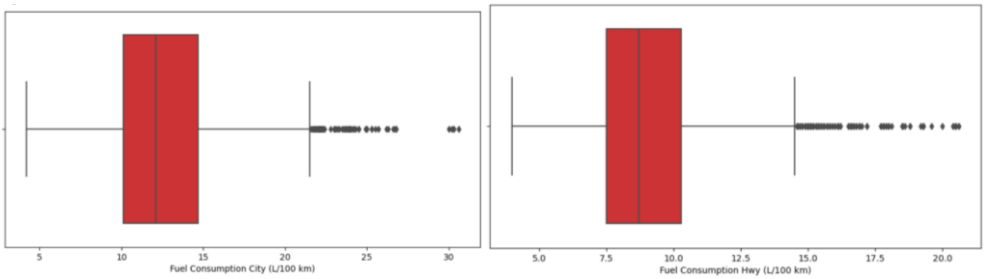
```
#Outliers check
num_feat = ['Engine Size(L)', 'Cylinders', 'Fuel
Consumption City (L/100 km)', 'Fuel Consumption Hwy
(L/100 km)', 'Fuel Consumption Comb (L/100 km)',
'Fuel Consumption Comb (mpg)', 'CO2
Emissions(g/km)']
cat_feat = ['Transmission', 'Fuel Type']

for num in num_feat :
  plt.figure(figsize = (10,5))
  sns.boxplot(data = df, x = num, palette = 'Set1')
  plt.figure()
```

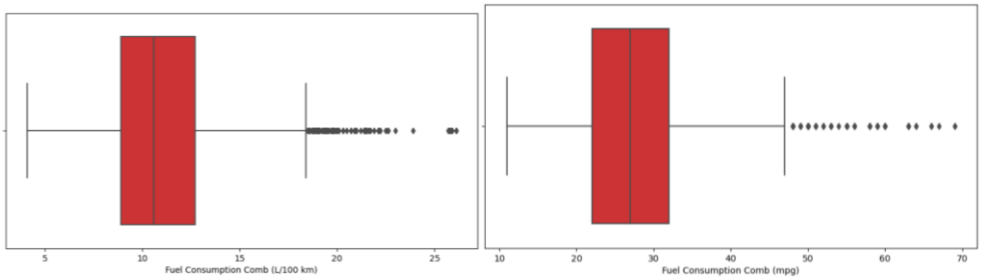
Gambar 10 *Potongan script untuk membuat boxplot tiap kolom*



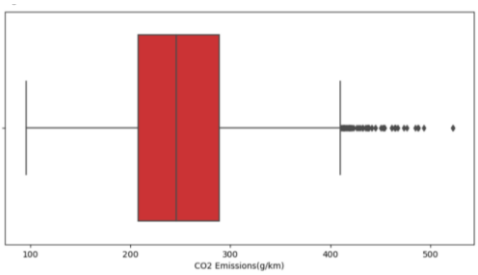
(a) (b)



(c) (d)



(e) (f)



(g)

Gambar 11 *Boxplot data kolom (a) Engine Size(L), (b) Cylinders, (c) Fuel Consumption City (L/100 km), (d) Fuel Consumption Hwy (L/100 km), (e) Fuel Consumption (L/100 km), (f) Fuel Consumption Comb (mpg), (g) CO2 Emissions(g/km)*

Dari gambar boxplot diatas, terlihat bahwa banyak data yang diluar range sehingga menjadi outlier. Data ini perlu dihapus atau diganti agar tidak mengurangi tingkat akurasi. Namun, pada studi kasus ini, data outlier akan dihapus karena data yang sudah dimiliki cukup.

```
#IQR Method | Remove Outliers
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)

IQR = Q3 - Q1

df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5
* IQR))).any(axis=1)]
```


Gambar 12 *Potongan script untuk menghapus data outlier*

Untuk menghapus *outlier*, dapat menggunakan metode IQR. Hal pertama yang dilakukan adalah menghitung quartil 1 dan quartil 3, kemudian hitung IQR = Q3 - Q1. Kemudian hapus data yang dibawah nilai Q1 - 1.5IQR atau data yang diatas nilai Q3 + 1.5IQR. Hasil data sekarang adalah 5816 baris dan 7 kolom yang digunakan dalam proses *modelling*.

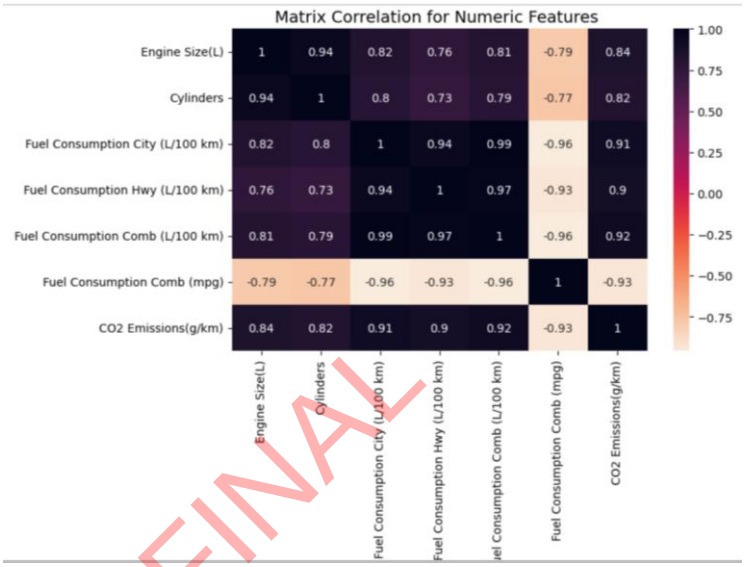
4.2 Visualisasi Korelasi dan Penyebaran Data

4.2.1 Visualisasi Korelasi

```
plt.figure(figsize = (8,5))
correlation_matrix = df.corr().round(2)

sns.heatmap(data = correlation_matrix, annot =
True, cmap = 'rocket_r')
plt.title("Matrix Correlation for Numeric Features
", size = 14)
```

Gambar 13 *Potongan script untuk mencari korelasi dan membuat visualisasi heatmap*



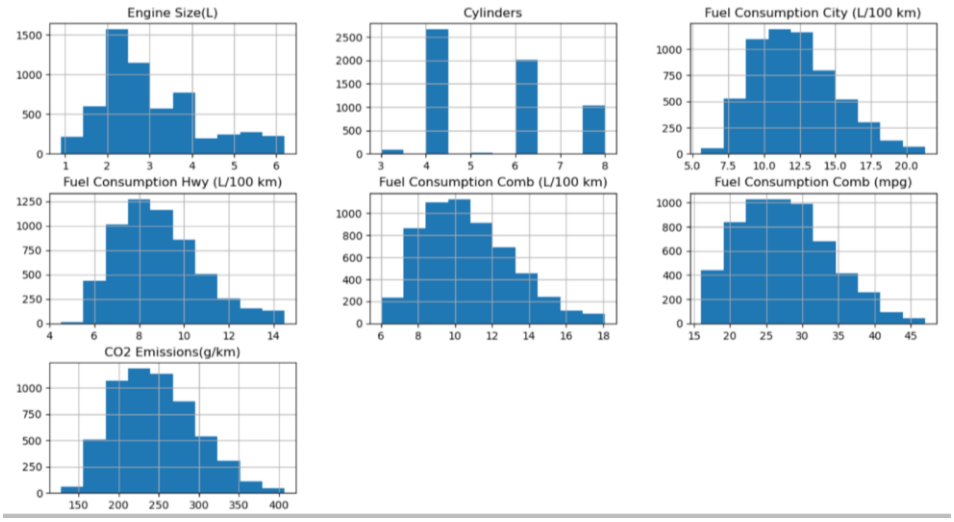
Gambar 14 *Output visualisasi heatmap*

Terlihat dari nilai korelasinya yang cukup tinggi terhadap kolom target maka, proses seleksi fitur menunjukkan hasil yang bagus. Sehingga dengan adanya korelasi yang tinggi, diharapkan hasil prediksi model dapat lebih mendekati data aktual.

4.2.2 Visualisasi Penyebaran Data

```
df.hist(figsize=(15, 8))
```

Gambar 15 *Potongan script untuk melihat penyebaran data*



Gambar 16 *Output visualisasi penyebaran data*

Dari penyebaran datanya, terlihat bahwa data cukup normal namun, secara kategorikan untuk *cylinders* memang cukup tidak rata. Namun, itu bukan menjadi masalah dikarenakan data lainnya berdistribusi normal.

4.3 *Splitting*, *Modelling* dan Evaluasi

4.3.1 Proses *Splitting*

Dalam studi kasus ini, data akan dibagi menjadi data *training* dan data *testing* dengan perbandingan rasio 70:30. Proses Splitting dibagi menjadi 3 seperti yang terlihat pada gambar 17.


```
# Memisahkan variabel target dan fitur
fitur = df.drop('CO2 Emissions(g/km)',1)
target = df['CO2 Emissions(g/km)']

# proses standarisasi
from sklearn import preprocessing
X = fitur.values
X = preprocessing.scale(X)
y = target.values

# Membagi data latih dan data uji dengan
perbandingan 70:30
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =
train_test_split(X,y,test_size=0.3,random_state=1)
```

Gambar 17 *Potongan script untuk proses splitting*

Proses pertama yaitu memisahkan antara variabel target dan fitur dimana variabel targetnya yaitu kolom ‘CO2 Emissions(g/km)’. Kemudian dilakukan proses standarisasi untuk variabel fitur (X) agar skala setiap fitur setara. Dilakukan proses splitting dari *library* sklearn untuk membagi data menjadi data *training* dan data *testing* dengan rasio 70:30 dan didapatkan data training sebanyak 4071 *records* dan data testing sebanyak 1745 *records*.

4.3.2 *Random Forest*

```
import time
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error,
mean_squared_error
start_time = time.time()
reg = RandomForestRegressor()
reg = reg.fit(X_train,y_train)
predict = reg.predict(X_test)
accuracy = reg.score(X_test, y_test)
mae = mean_absolute_error(y_test, predict)
end_time = time.time()
lama = end_time - start_time

print("Random Forest Regressor")
print("="*35)
print(f"R2 Score: {round(accuracy * 100, 2)}")
print(f'RMSE:
{round(np.sqrt(mean_squared_error(y_pred=predict,y_t
rue=y_test)),2)}’)
print(f"Mean Absolute Error: {round(mae, 2)}")
print("Lama pemrosesan: ", lama, "detik")
```


Gambar 18 *Potongan script untuk proses algoritma Random Forest*

Berikut adalah *code* untuk melakukan regresi menggunakan algoritma *Random Forest* dengan bantuan *library sklearn*. Hal pertama yang dilakukan adalah mendefinisikan class *RandomForestRegressor* yang kemudian dilatih dengan data *training* yaitu X_train dan y_train. Kemudian model akan diujikan dengan data *testing* yaitu X_test dan y_test. Evaluasi yang digunakan terdapat akurasi (koefisien determinasi), RMSE, dan MAE.

4.3.3 *XGBoost*

```
import xgboost as xgb
start_time = time.time()
xgb_model =
xgb.XGBRegressor(objective="reg:squarederror",
n_estimators=1000, random_state=42,n_jobs=-1)
xgb_model.fit(X_train,y_train,early_stopping_rounds=
10,eval_set=[(X_test, y_test)],verbose = 0)
y_pred = xgb_model.predict(X_test)

xgb_mae = mean_absolute_error(y_test, y_pred)
end_time = time.time()
lama = end_time - start_time
print("XGBOOST")
print("="*35)
print(f'R2 Score:
{round(xgb_model.score(X_test,y_test)*100,2)}')
print(f'RMSE:
{round(np.sqrt(mean_squared_error(y_pred=y_pred,y_tr
ue=y_test)),2)}')
print(f"Mean Absolute Error: {round(xgb_mae, 2)}")
print("Lama pemrosesan: ", lama, "detik")
```

Gambar 19 *Potongan script untuk proses algoritma XGBoost*

Berikut adalah *code* untuk melakukan regresi menggunakan algoritma *XGBoost* dengan bantuan *library xgboost*. Hal pertama yang dilakukan adalah mendefinisikan *class XGBRegressor* yang kemudian dilatih dengan data *training* yaitu X_train dan y_train. Kemudian model akan diujikan dengan data *testing* yaitu X_test dan y_test. Evaluasi yang digunakan terdapat akurasi (koefisien determinasi), RMSE, dan MAE.

4.4 Perbandingan Hasil Algoritma

4.4.1 Koefisien Determinasi, RMSE, MAE, Waktu

Tabel 1 *Perbandingan Hasil Koefisien Determinasi, RMSE, MAE, Waktu*

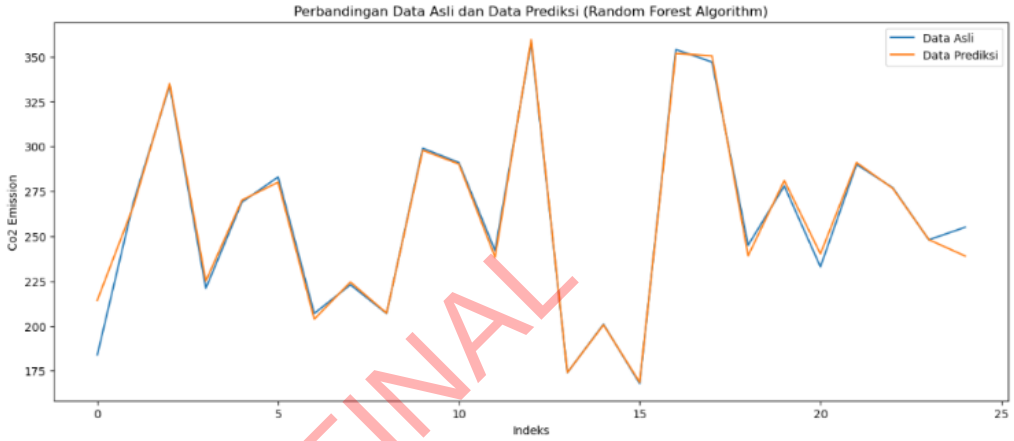
Algoritma	Koefisien Determinasi	RMSE	MAE	Waktu Pemrosesan
Random Forest	97.56%	7.77	3.22	0.538 s
XGBoost	98.03%	6.98	3.18	0.119 s

Dari Hasil Perbandingan yang ditunjukkan pada tabel 1, *XGBoost* lebih unggul dibandingkan *Random Forest*. Hal ini dikarenakan akurasi serta

kelayakan model yang ditunjukkan pada koefisien determinasi, nilai *XGBoost* lebih tinggi. Pada perbandingan selisih kesalahan prediksi yang ditunjukkan pada RMSE dan MAE, nilai *XGBoost* lebih kecil. Sedangkan, pada waktu pemrosesan, kecepatan menggunakan algoritma *XGBoost* 5x lebih cepat.

4.4.2 Perbandingan Data Prediksi dengan Data Aktual

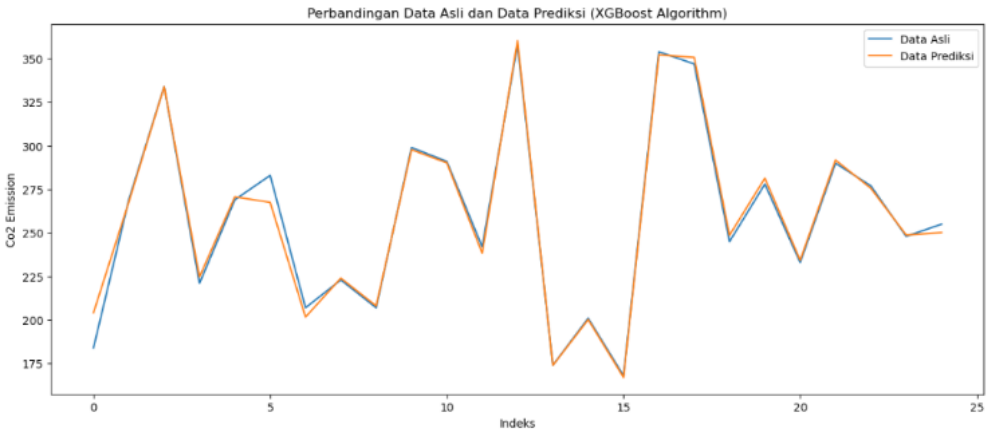
Untuk melihat perbandingan secara lebih jelas, maka akan dibandingkan data hasil prediksi dengan data aktual untuk algoritma *Random Forest* yang tampak pada gambar 20 dan tabel 2 sedangkan, untuk algoritma *XGBoost* tampak pada gambar 21 dan tabel 3.



Gambar 20 *25 Pertama Data Prediksi dengan Data Aktual Menggunakan Random Forest*

Tabel 2 *25 Pertama Data Prediksi dengan Data Aktual Menggunakan Random Forest*

No	Data Asli	Data Prediksi	No	Data Asli	Data Prediksi
1	184	214.413	14	174	174.090
2	269	266.831	15	201	200.660
3	334	335.044	16	168	168.700
4	221	225.052	17	354	351.775
5	269	270.100	18	347	350.465
6	283	280.090	19	245	239.132
7	207	203.818	20	278	281.050
8	223	224.502	21	233	240.200
9	207	207.185	22	290	291.023
10	299	297.884	23	277	276.665
11	291	290.192	24	248	248.090
12	242	238.358	25	255	239.010
13	358	359.518			



Gambar 21 *25 Pertama Data Prediksi dengan Data Aktual Menggunakan XGBoost*

Tabel 3 *25 Pertama Data Prediksi dengan Data Aktual Menggunakan XGBoost*

No	Data Asli	Data Prediksi	No	Data Asli	Data Prediksi
1	184	204.133	14	174	173.916
2	269	267.726	15	201	200.217
3	334	333.931	16	168	167.011
4	221	224.686	17	354	352.232
5	269	270.707	18	347	350.757
6	283	267.519	19	245	248.664
7	207	201.661	20	278	281.390
8	223	223.998	21	233	234.286
9	207	208.073	22	290	291.803
10	299	297.754	23	277	275.575
11	291	290.201	24	248	248.767
12	242	238.357	25	255	250.178
13	358	360.270			

Berdasarkan selisih dan hasil visualisasi, terlihat bahwa *XGBoost* memiliki selisih hasil prediksi dengan data aktual lebih kecil dibandingkan algoritma *Random Forest*. Sehingga, *XGBoost* memiliki keunggulan dalam studi kasus ini.

BAB V: PENUTUP

5.1 Kesimpulan

Karbon dioksida menyebabkan banyak dampak negatif diantaranya adalah *global warming*. Salah satu penyebabnya adalah kendaraan sehingga pentingnya adanya prediksi jumlah karbon dioksida yang dihasilkan. Berdasarkan penelitian yang telah dilakukan dengan membandingkan algoritma *Random Forest* dan *XGBoost* tidak menunjukkan perbedaan secara signifikan untuk melakukan prediksi terhadap jumlah karbon dioksida yang dikeluarkan oleh kendaraan.

Hasil algoritma *Random Forest* yang memiliki koefisien determinasi sebesar 97.56%, nilai RMSE sebesar 7.77, dan nilai MAE sebesar 3.22 sedangkan, algoritma *XGBoost* memiliki koefisien determinasi sebesar 98.03%, nilai RMSE sebesar 6.98, dan nilai MAE sebesar 3.18 yang artinya berdasarkan kelayakan model serta selisih nilai prediksi, *XGBoost* adalah algoritma yang lebih tepat untuk digunakan. Secara perbedaan waktu pemrosesan, *XGBoost* hanya membutuhkan 0.119 detik sedangkan, *Random forest* membutuhkan 0.538 detik sehingga *XGBoost* lebih cepat lima kali lipat. Dari hasil perbandingan, algoritma *XGBoost* adalah pilihan yang tepat dalam membangun model prediksi karbon dioksida yang dihasilkan kendaraan karena memiliki akurasi tinggi dan efisiensi dalam waktu.

5.2 Saran

Untuk penelitian selanjutnya, dapat dipertimbangkan untuk mengembangkan algoritma *sensitive* biaya berdasarkan undersampling untuk meningkatkan performa metode *XGBoost* pada kumpulan data tingkat emisi karbon dioksida di udara.

DAFTAR PUSTAKA

Alam, I., Farid, D. M., & Rossetti, R. J. (2018). The Prediction of Traffic Flow with Regression Analysis. *Crossmark*, 661 - 671.

Allwright, S. (2023, Maret 6). *XGBoost vs Random Forest*. Retrieved from stephenallwright.com: https://stephenallwright.com/xgboost-vs-random-forest/

Brownlee, J. (2018). *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn.* Machine Learning Mastery.

C2ES. (2008). *Policies to Reduce Emissions from the Transportation Sector.* Washington, DC: Center For Climate And Energy Solutions.

Canada.ca. (2023, Mei 26). *About Canada.ca*. Retrieved from Open.canada.ca: https://www.canada.ca/en/government/about.html

Hosseini, S. M., Saifoddin, A., Shirmohammadi, R., & Aslani, A. (2019). Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports*, 619 - 631.

Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. *Sustainable Cities and Society*.

Labiba, D., & Pradoto, W. (2018). SEBARAN EMISI CO2 DAN IMPLIKASINYA TERHADAP PENATAAN RUANG AREA INDUSTRI DI KENDAL. *Jurnal Pengembangan Kota*, 164 - 173.

Luo, J., Zhang, Z., Fu, Y., & Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics 27*.

Manalu, D. A., & Gunadi, G. (2022). IMPLEMENTASI METODE DATA MINING K-MEANS CLUSTERING TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON PADA CV DIGITAL DIMENSI. *INFOTECH: JOURNAL OF TECHNOLOGY INFORMATION*, 45 - 54.

Osman, A. I., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, 1545-1556.

Putra, A. E., & Juarna, A. (2021). Prediksi Produksi Daging Sapi Nasional dengan Metode Regresi Linier dan Regresi Polinomial. *Jurnal Ilmiah KOMPUTASI*, 209 - 215.

Ritchie, H. (2020, Oktober 6). *Co2 emissions from transport*. Retrieved from OurWorldInData.org: https://ourworldindata.org/co2-emissions-from-transport

Ritchie, H., Roser, M., & Rosado, P. (2020). *CO_2 and Greenhouse Gas Emissions*. Retrieved from OurWorldInData.org: https://ourworldindata.org/co2-emissions

Teter, J. (2022). *Transport.* iea.org.

Thanh, H. V., Yasin, Q., Al-Mudhafar, W. J., & Lee, K.-K. (2022). Knowledge-based machine learning techniques for accurate prediction of CO2 storage performance in underground saline aquifers. *Applied Energy*.

Tumiwa, F., & Wijayani, L. (2021). *Climate Transparency Report 2021*. Jakarta: Institute for Essential Services Reform.

Turrentine, J. (2021, April 7). *Global Warming 101*. Retrieved from NRDC.org: https://www.nrdc.org/stories/global-warming-101#warming

Wu, H., & Liu, Y. (2017). Prediction of Road Traffic Congestion Based on Random Forest. *10th International Symposium on Computational Intelligence and Design* (pp. 361 - 364). IEEE.

LAMPIRAN

Lampiran 1: Penggalan dataset awal yang digunakan

	Make	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0	ACURA	ILX	COMPACT	2.0	4	AS5	Z	9.9	6.7	8.5	33	196
1	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
2	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6.0	5.8	5.9	48	136
3	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
4	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
...	...	...	...	...	...	...	...	...	...	...	...	...
7380	VOLVO	XC40 T5 AWD	SUV - SMALL	2.0	4	AS8	Z	10.7	7.7	9.4	30	219
7381	VOLVO	XC60 T5 AWD	SUV - SMALL	2.0	4	AS8	Z	11.2	8.3	9.9	29	232
7382	VOLVO	XC60 T6 AWD	SUV - SMALL	2.0	4	AS8	Z	11.7	8.6	10.3	27	240
7383	VOLVO	XC90 T5 AWD	SUV - STANDARD	2.0	4	AS8	Z	11.2	8.3	9.9	29	232
7384	VOLVO	XC90 T6 AWD	SUV - STANDARD	2.0	4	AS8	Z	12.2	8.7	10.7	26	248

7385 rows × 12 columns

Lampiran 2: Penggalan dataset setelah proses pre-processing yang digunakan

	Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
0	2.0	4	9.9	6.7	8.5	33	196
1	2.4	4	11.2	7.7	9.6	29	221
2	3.5	6	12.7	9.1	11.1	25	255
3	3.5	6	12.1	8.7	10.6	27	244
4	3.5	6	11.9	7.7	10.0	28	230
...	...	...	...	...	...	...	...
5811	2.0	4	10.7	7.7	9.4	30	219
5812	2.0	4	11.2	8.3	9.9	29	232
5813	2.0	4	11.7	8.6	10.3	27	240
5814	2.0	4	11.2	8.3	9.9	29	232
5815	2.0	4	12.2	8.7	10.7	26	248

5816 rows × 7 columns

Lampiran 3: Penggalan hasil standarisasi variabel fitur

	Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)
0	-0.834465	-0.917270	-0.816221	-1.143000	-0.926063	0.882886
1	-0.505370	-0.917270	-0.361453	-0.606714	-0.463656	0.220963
2	0.399642	0.400951	0.163278	0.144085	0.166897	-0.440959
3	0.399642	0.400951	-0.046615	-0.070429	-0.043287	-0.109998
4	0.399642	0.400951	-0.116579	-0.606714	-0.295509	0.055483
...	...	...	...	...	...	...
5811	-0.834465	-0.917270	-0.536364	-0.606714	-0.547730	0.386444
5812	-0.834465	-0.917270	-0.361453	-0.284943	-0.337546	0.220963
5813	-0.834465	-0.917270	-0.186543	-0.124057	-0.169398	-0.109998
5814	-0.834465	-0.917270	-0.361453	-0.284943	-0.337546	0.220963
5815	-0.834465	-0.917270	-0.011633	-0.070429	-0.001250	-0.275478

5816 rows × 6 columns