Analisis Prediktif Penghasil Emisi Co2 Berbasis Industri dan Ekonomi dengan menggunakan Random Forest Regression

Studi Kasus Spanyol Periode 2010-2022

Vincentius Darren W W
Institut Bisnis dan Informatika Kesatuan

Nathaniel Steave Harjanto
Universitas Sebelas Maret

Ikhsan Ari Novianto
Universitas Sebelas Maret

Mafa Oktavia
Universitas Sebelas Maret

Dewi Nurul Istiqomah
Universitas Lampung

Ropita Yohana Situmorang
Universitas Mikroskil

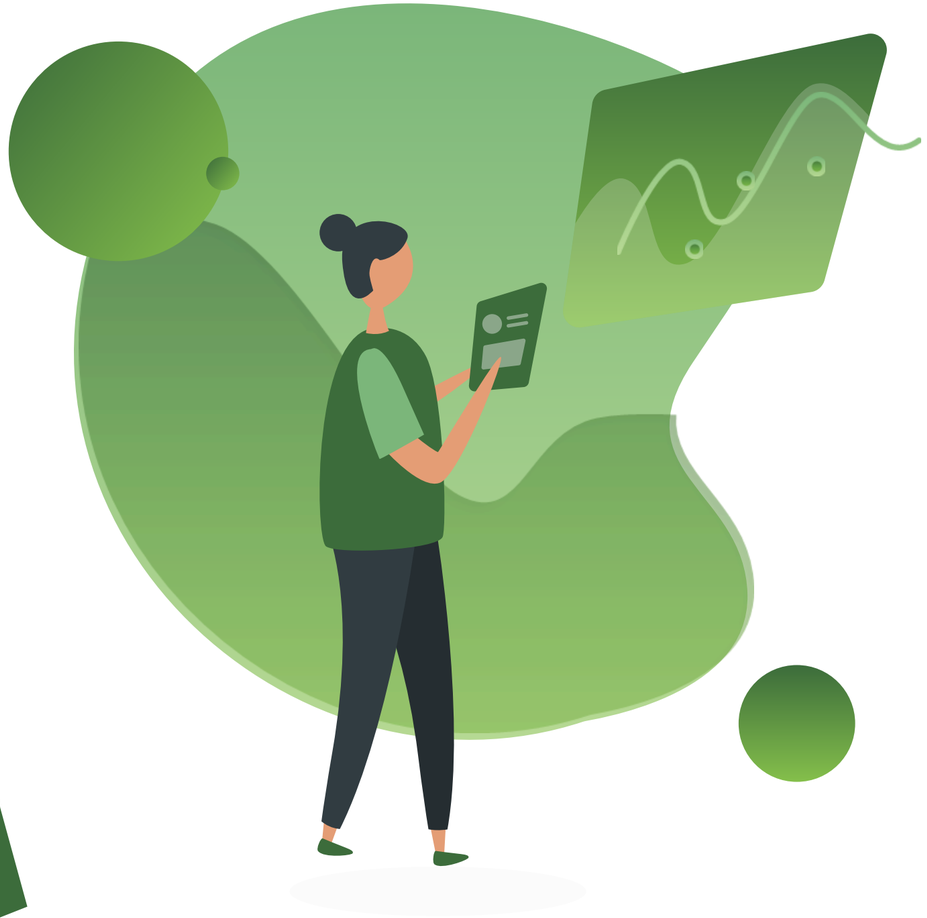
Almira Nurchawilah
Universitas Informatika dan Bisnis Indonesia

Kompi 5 | Data Rangers

Business Understanding

Latar Belakang

Menganalisis **data emisi co2** di negara **Spanyol** dapat mengedintifikasi **tren, faktor penyebab** dan **potensi solusi** untuk mengurangi **emisi co2** dalam mengatasi **perubahan iklim**.



Problem

Bagaimana perkembangan tingkat **emisi Co2** dari segi **industri , ekonomi (pajak) , energi,** serta **kepadatan penduduk** sebagai pendukung **kebijakan pengurangan emisi karbon** pada negara **Spanyol** pada tahun-tahun berikutnya?

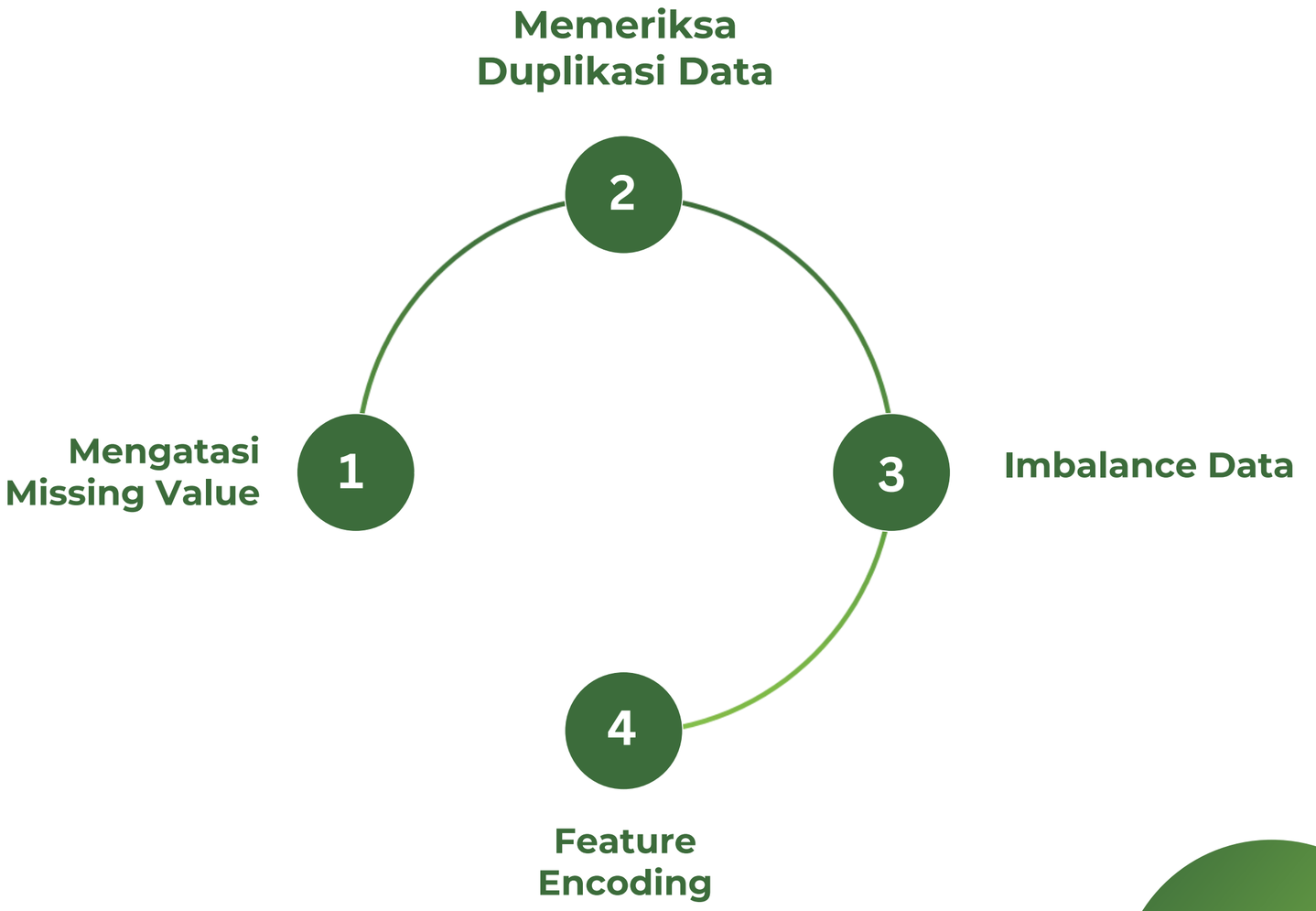
Variabel	Tipe Data	Keterangan
POPDEN	Kontinu	Kepadatan penduduk (jumlah penduduk per satuan luas)
CO2_PBEMCAP, CO2_DBEMCAP, CO2_PBEM,CO2_DBEM OWID.CB.co2 , OWID.CB.consumption_co2	Kontinu	Ukuran emisi CO2 dari 2010-2022
GDP_PC, GDP_RCAP OWID.CB.consumption_co2_per_capita	Kontinu	Produksi per kapita (produksi emisi co2 per orang)
NRG_INT	Kontinu	Intensitas energi (mengukur seberapa banyak sumber daya alam yang dikonsumsi oleh ekonomi dan populasi)
ENVTAX_VEH, NRGC, PM_MOR, LTAX, ENVTAX	Kontinu	Pajak dan biaya yang diterapkan berkaitan dengan kebijakan lingkungan dan ekonomi

Column : 16 **Row : 898**

Sumber Data : Our World in Data

Data Preparation

Proses **data preparation** dilakukan untuk memastikan **data emisi CO2** siap digunakan dalam analisis. Langkah-langkah ini mencakup **pembersihan**, **pemeriksaan**, dan **transformasi** data agar lebih berkualitas dan mendukung hasil yang akurat.



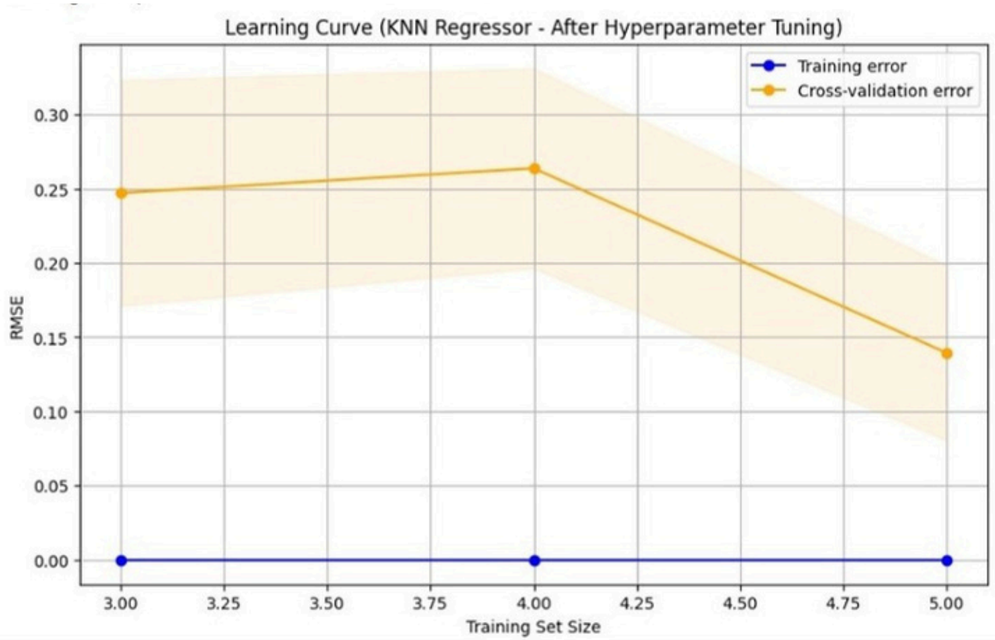
Model

Terdapat dua model yang akan dipilih, yakni KNN (K-Nearest Neighbors) dan Random Forest Regression.

Berikut penjelasan mengenai setiap model

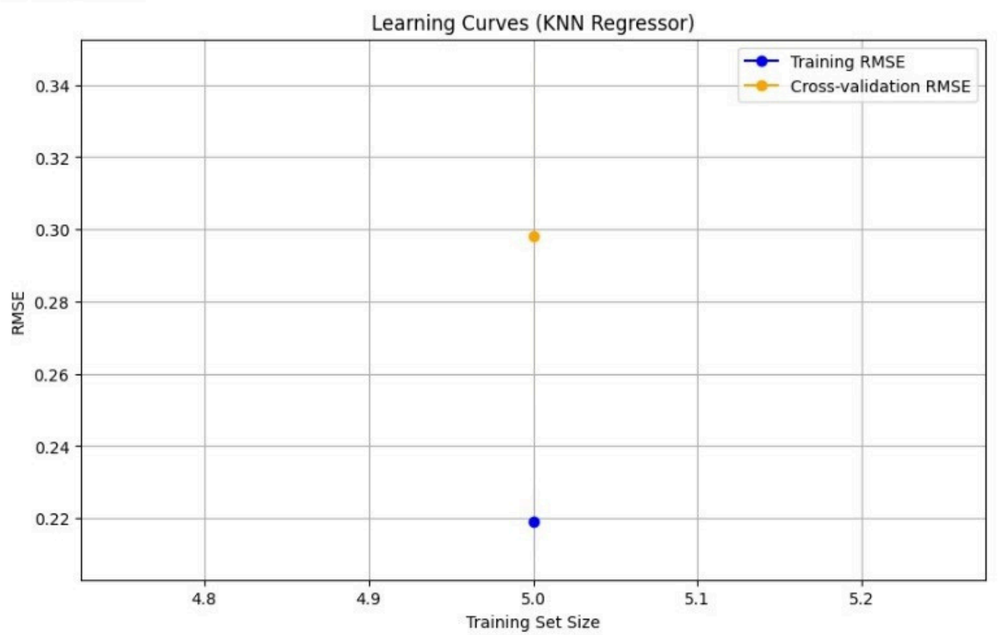
	KNN	Random Forest Regression
Prinsip	Mengestimasi nilai target berdasarkan rata-rata nilai dari KNN di ruang fitur.	Kombinasi prediksi dari beberapa pohon keputusan (ensemble) dengan rata-rata hasil.
Kelebihan	Sederhana, intuitif	Akurat, tahan terhadap overfitting, mampu menangani fitur yang tidak relevan.
Kekurangan	Rentan terhadap dimensi tinggi, lambat saat prediksi, rentan terhadap outlier.	Relatif kompleks, membutuhkan lebih banyak waktu untuk pelatihan, dan konsumsi memori lebih besar.

Komparasi



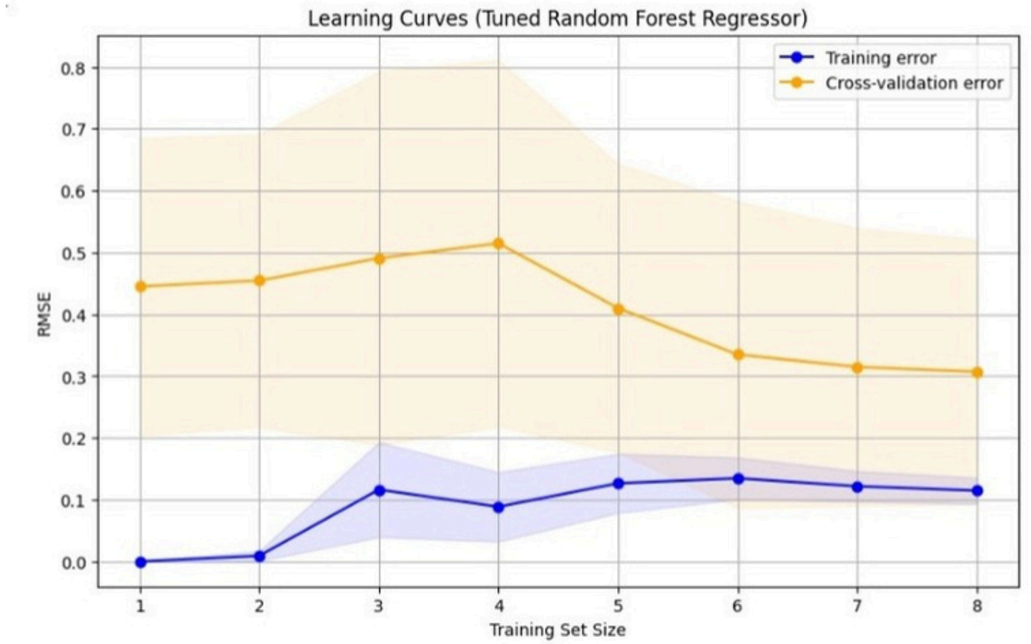
- **Cross-validation error** lebih tinggi dari training error, menunjukkan adanya gap performa antara data pelatihan dan validasi.
- Error menurun seiring bertambahnya **data pelatihan**, mengindikasikan kebutuhan data lebih besar untuk **generalisasi.**
- Model **KNN** menunjukkan **overfitting**, dengan **training error** yang jauh lebih rendah dibandingkan **cross-validation error.**

Hanya ada beberapa **titik** yang terlihat pada grafik, yang mengindikasikan bahwa proses perhitungan l**earning curves** pada model **KNN** tidak menghasilkan variasi data yang jelas (hanya satu titik per **ukuran training set**).



Komparasi

Garis menunjukkan **variasi (standard deviation)** pada error tersebut. Plot ini membantu mengidentifikasi masalah seperti **overfitting** atau **underfitting** dan memberikan wawasan tentang **kemampuan generalisasi model.**



ALASAN

Pemilihan Random Forest model karena kemampuannya:

- **Mengatasi interaksi fitur**
- **Toleran terhadap noise**
- **Mengatasi skala variabel yang berbeda**
- **Feature Importance**
- **Pola Non-Linear**

Metrik Evaluasi Model

MAE

10.7%

Mengukur rata-rata besarnya kesalahan prediksi dalam skala data asli.

MSE

0.17%

Mengukur rata-rata kuadrat kesalahan prediksi.

RMSE

13%

Memberikan interpretasi lebih intuitif karena satuannya sama dengan variabel target.

R-Squared (R^2)

92.7%

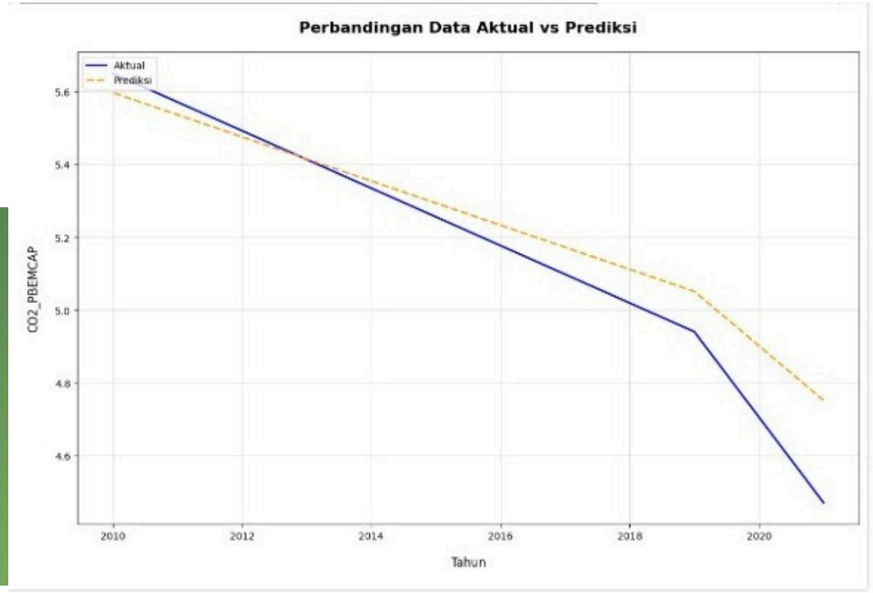
Menunjukkan seberapa baik model menjelaskan variabilitas data.

Metrik ini digunakan untuk membandingkan **kinerja model**, seperti **Random Forest**, dan memilih model dengan kesalahan prediksi rendah **(MAE/MSE/RMSE kecil)** serta kemampuan penjelasan tinggi **(R^2 besar).**

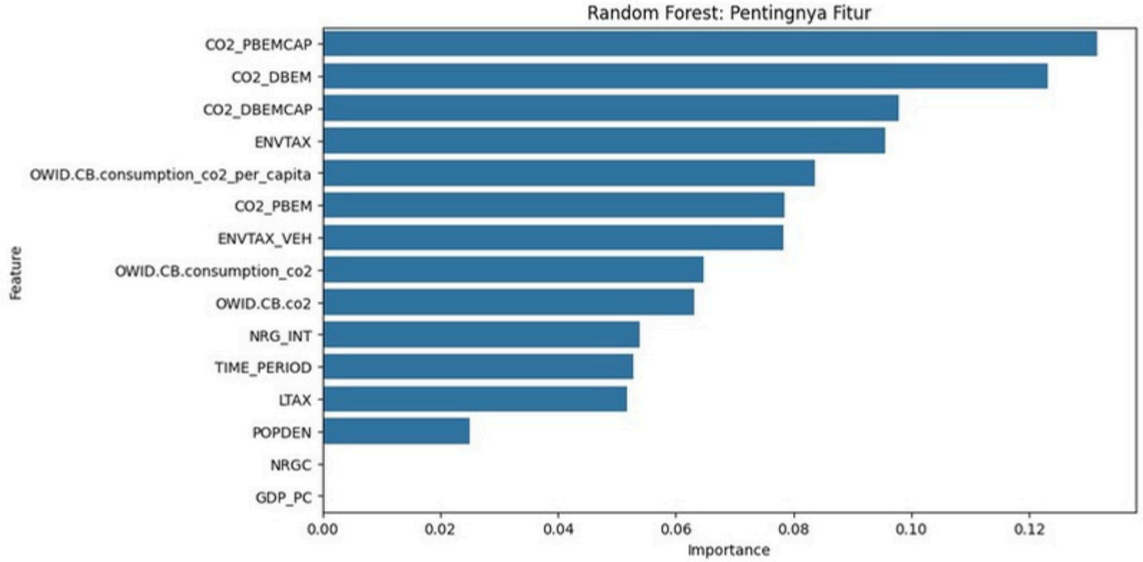
Feature Importance

Data Prediksi vs Data Aktual

Baik **data aktual** maupun **prediksi** menunjukkan adanya tren **penurunan emisi CO2,** yang menandakan kesesuaian antara hasil prediksi dan kondisi nyata. Hal ini menunjukkan bahwa **model** yang digunakan cukup baik dalam menangkap pola tren **penurunan emisi CO2**, sehingga dapat memberikan wawasan yang relevan untuk analisis lebih lanjut.



Fitur yang berpengaruh



- **ENVTAX_VEH:** Kebijakan pajak kendaraan sangat memengaruhi variabel prediksi.
- **CO2_PBEMCAP, CO2_PBEM, CO2_DBEM:** Menunjukkan pentingnya isu lingkungan.
- **GDP_PC:** Menggambarkan hubungan antara ekonomi dan lingkungan

Model mengungkap kompleksitas masalah dan dapat menjadi dasar pengambilan kebijakan yang lebih baik, terutama terkait lingkungan dan pajak kendaraan.

Feature Selection

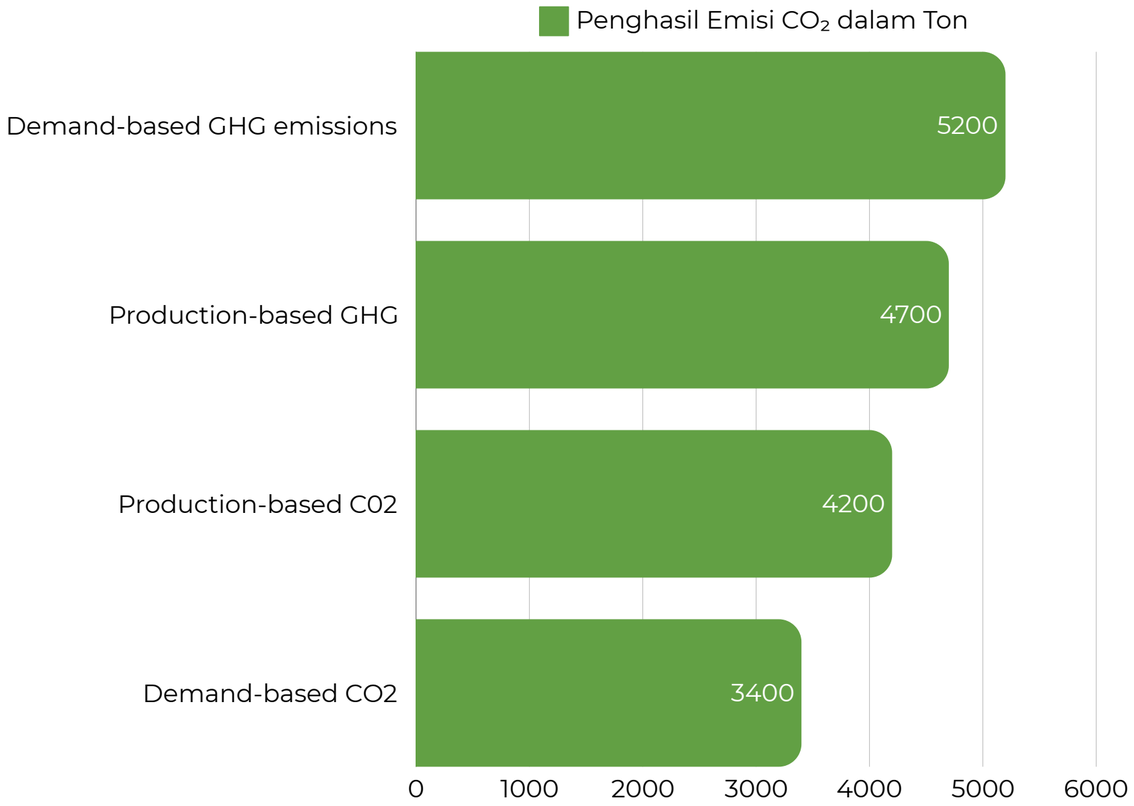
Memahami **feature importance** memungkinkan kita untuk mengarahkan perhatian pada **faktor yang memiliki dampak besar terhadap hasil**, sehingga menghasilkan **model yang lebih relevan** secara bisnis.

Sementara itu, **VIF** membantu mengidentifikasi **variabel redundan** yang dapat dihapus atau dikombinasikan untuk **mengurangi multikolinieritas**, memastikan **stabilitas** dan **akurasi model**.

Fitur-fitur yang dipilih yakni :

- **CO2_PBEMCAP**
- **CO2_DBEMCAP**
- **CO2_PBEM**
- **CO2_DBEM**
- **NRG_INT**
- **LTAX**
- **ENVTAX_VEH**
- **ENVTAX**
- **TIME_PERIOD**
- **OWID.CB.consumption_co2**
- **OWID.CB.consumption_co2_per_capita**
- **OWID.CB.co2**

Insight



Demand- GHG

5,2 rb

Demand-based GHG

* **Menunjukkan pola konsumsi masyarakat adalah kontributor utama emisi.**

Demand- CO2

3,4 rb

Demand-based CO2

* **Demand-based CO2 emissions berada di angka terendah yaitu 3,4 ribu ton, namun tetap menunjukkan angka yang mengkhawatirkan.**

Production-GHG

4,7 rb

Production-based GHG

* **Angka ini lebih tinggi 0,5 ribu ton dibanding production-based CO2, menunjukkan adanya gas rumah kaca lain selain CO2 yang dihasilkan dari proses produksi**

Production- CO2

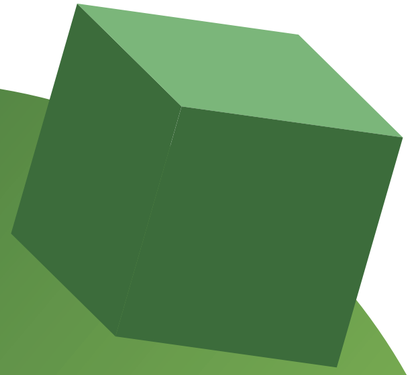
4,2 rb

Production-based CO2

* **Mencerminkan intensitas penggunaan bahan bakar fosil dalam proses produksi**
* **Aktivitas produksi dan industri di Spanyol menghasilkan jejak karbon yang cukup besar**

> **Spanyol menghadapi tantangan serius dalam emisi CO_2, dimana sektor konsumsi (demand-based) menempati posisi tertinggi, diikuti oleh emisi dari sektor produksi. Tingginya emisi dari kedua sektor ini menunjukkan perlunya transformasi menyeluruh, terutama dalam peralihan ke energi terbarukan dan adopsi teknologi rendah karbon untuk mencapai masa depan yang lebih berkelanjutan.**

Rekomendasi



Terkait Model yang Dikembangkan

Data Rangers

Finance

Optimalisasi Model

Gunakan data temporal untuk meningkatkan prediksi dengan menangkap pola jangka panjang dan musiman. Model seperti LSTM dapat menjadi opsi.

Evaluasi Lebih Mendalam

Lakukan evaluasi tambahan menggunakan metrik seperti Residual Plots, MAE, dan cross-validation untuk memastikan generalisasi model.

Pemilihan dan Pengolahan Fitur

- Terapkan seleksi fitur menggunakan feature importance untuk memastikan hanya variabel yang relevan digunakan.
- Tangani missing values dan outliers dengan teknik seperti imputation atau winsorizing.

Eksplorasi Model Lain

Coba model seperti Gradient Boosting Machines (GBM) atau XGBoost untuk pola yang lebih kompleks.

Integrasi ke Workflow Bisnis

Pastikan hasil model mudah dipahami dengan visualisasi seperti feature importance, sehingga dapat diimplementasikan untuk kebijakan.

Business Solution

Membuat model machine learning menggunakan Random Forest Regressor untuk **memprediksi tingkat emisi CO2 di masa depan dengan mempertimbangkan tren masa lalu dan hubungan di antara faktor industri, kepadatan penduduk, energi, dan ekonomi (pajak) Spanyol**. Untuk mencapai keberlanjutan lingkungan **jangka panjang** dan strategis, keluaran **model ini digunakan untuk mendorong kebijakan berbasis data, seperti peraturan energi yang lebih ketat atau adopsi teknologi rendah karbon**.

Glossary

Variabel	Deskripsi
CO2_PBEM	Production-based CO2 emissions
CO2_DBEMCAP	Demand-based CO2 intensity energy-related CO2 per capita
CO2_PBEMCAP	Production-based CO2 intensity, energy-related CO2 per capita
CO2_DBEM	Demand-based CO2 emissions
ENVTAX_VEH	Road transport related tax revenue
OWID.CB.consumption_co2	Konsumsi CO2
OWID.CB.co2	Emisi CO2 Total per Tahun
OWID.CB.consumption_co2_per_capita	Konsumsi CO2 per kapita
ENVTAX	Environment related taxes revenue
NRG_INT	Energy intensity per capita
LTAX	Labour tax revenue
TIME_PERIOD	Periode Waktu (Tahun)

THANK YOU

Data Rangers