ABSTRAK

Santrisna, I Putu (2026). *Analisis Faktor Dominan dan Prediksi Emisi Karbon di Indonesia Menggunakan Pendekatan Hybrid: Recursive Feature Elimination dan Extreme Gradient Boosting (XGBoost)*. Tesis, Program Studi Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Tesis ini sudah disetujui dan diperiksa oleh Pembimbing I: Prof. Dr. Ir. Dewa Gede Hendra Divayana, S.Kom., M.Kom., IPU., ASEAN.Eng., APEC.Eng. dan Pembimbing II: Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs.

Kata-kata kunci: ekstrapolasi, emisi karbon, Recursive Feature Elimination, seleksi fitur, XGBoost, Pembelajaran Mesin.

Penelitian ini bertujuan untuk (1) mengidentifikasi variabel prediktor dengan kontribusi prediktif tertinggi terhadap emisi karbon di Indonesia melalui seleksi fitur Recursive Feature Elimination (RFE) dengan estimator Random Forest, (2) membangun dan mengevaluasi kinerja model prediksi emisi karbon berbasis XGBoost menggunakan fitur hasil seleksi RFE yang diukur dengan RMSE dan R^2 pada data uji, serta (3) mengevaluasi apakah penerapan seleksi fitur RFE menghasilkan RMSE yang lebih rendah dibandingkan model XGBoost baseline yang menggunakan seluruh fitur.

Penelitian ini menggunakan pendekatan kuantitatif dengan data deret waktu tahunan emisi CO_2 per kapita Indonesia periode 1990–2023 (34 observasi) yang bersumber dari World Bank. Lima variabel prediktor digunakan, yaitu PDB per kapita, konsumsi energi primer, persentase energi baru terbarukan, luas hutan, dan populasi. Data dibagi secara kronologis menjadi tiga lapis: data latih (1990–2011), data validasi (2012–2016), dan data uji (2017–2023). Seleksi fitur dilakukan dengan RFECV menggunakan TimeSeriesSplit hanya pada data latih. Model dilatih pada gabungan data latih dan validasi (1990–2016), kemudian dievaluasi pada data uji menggunakan skema hold-out dan walk-forward. Multiple Linear Regression digunakan sebagai model pembanding, dan multikolinearitas antar prediktor diperiksa melalui nilai Variance Inflation Factor (VIF).

Hasil penelitian menunjukkan tiga hal. Pertama, konsumsi energi primer merupakan variabel prediktor dengan kontribusi prediktif tertinggi dan menjadi satu-satunya fitur terpilih oleh RFECV. Kedua, model XGBoost dengan fitur hasil seleksi RFE menghasilkan RMSE 0,359 dan R^2 −2,238 pada data uji; nilai R^2 negatif ini bukan disebabkan kegagalan model dalam belajar (R^2 in-sample mencapai 1,000 dan R^2 walk-forward positif +0,579), melainkan keterbatasan struktural model berbasis pohon dalam melakukan ekstrapolasi, karena seluruh nilai data uji berada di atas rentang data latih. Ketiga, seleksi fitur RFE tidak menghasilkan RMSE yang lebih rendah dibandingkan model XGBoost baseline (0,357). Disimpulkan bahwa pada data deret waktu bertren monoton yang periode ujinya melampaui rentang latih, keberhasilan prediksi ditentukan oleh kemampuan ekstrapolasi model. Disarankan agar penelitian selanjutnya mempertimbangkan model dengan kemampuan ekstrapolasi sebagai pembanding utama sejak tahap perancangan, serta menerapkan skema walk-forward apabila model berbasis pohon tetap digunakan.

ABSTRACT

Santrisna, I Putu (2026). *Analysis of Dominant Factors and Prediction of Carbon Emissions in Indonesia Using a Hybrid Approach: Recursive Feature Elimination and Extreme Gradient Boosting (XGBoost)*. Thesis, Computer Science Study Program, Postgraduate Program, Universitas Pendidikan Ganesha.

This thesis has been approved and examined by Supervisor I: Prof. Dr. Ir. Dewa Gede Hendra Divayana, S.Kom., M.Kom., IPU., ASEAN.Eng., APEC.Eng. and Supervisor II: Dr. Ir. I Made Gede Sunarya, S.Kom., M.Cs

Keywords: carbon emissions, extrapolation, feature selection, Recursive Feature Elimination, XGBoost, Machine learning.

This study aims to (1) identify the predictor variable with the highest predictive contribution to carbon emissions in Indonesia through Recursive Feature Elimination (RFE) with a Random Forest estimator, (2) build and evaluate the performance of an XGBoost-based carbon emission prediction model using the RFE-selected features, measured by RMSE and R^2 on the test set, and (3) evaluate whether applying RFE feature selection yields a lower RMSE than the baseline XGBoost model using all features.

This study employs a quantitative approach using annual time-series data of Indonesia's per-capita CO_2 emissions for the period 1990–2023 (34 observations) sourced from the World Bank. Five predictor variables are used: GDP per capita, primary energy consumption, renewable energy share, forest area, and population. The data are split chronologically into three layers: training (1990–2011), validation (2012–2016), and testing (2017–2023). Feature selection is performed with RFECV using TimeSeriesSplit on the training data only. The models are trained on the combined training and validation data (1990–2016), then evaluated on the test data using hold-out and walk-forward schemes. Multiple Linear Regression is used as a comparison model, and multicollinearity among predictors is examined through the Variance Inflation Factor (VIF).

The results show three findings. First, primary energy consumption is the predictor variable with the highest predictive contribution and the only feature selected by RFECV. Second, the XGBoost model with RFE-selected features yields an RMSE of 0.359 and an R^2 of −2.238 on the test set; this negative R^2 is not caused by the model's failure to learn (in-sample R^2 reaches 1.000 and walk-forward R^2 is positive, +0.579), but rather by the structural limitation of tree-based models in extrapolation, as all test values lie above the range of the training data. Third, RFE feature selection does not produce a lower RMSE than the baseline XGBoost model (0.357). It is concluded that for monotonically trending time-series data whose test period exceeds the training range, prediction success is determined by the model's extrapolation capability. It is recommended that future research consider models with extrapolation capability as the main comparison from the design stage, and apply a walk-forward scheme if tree-based models are retained.