Optimasi Akurasi Koefisien Pajak Kendaraan Bermotor di Indonesia Menggunakan Metode Klasifikasi dan Regresi

Joiner Tennye Ariel Togatorop[1], Antonius Rachmat Chrismanto[2], Willy Sudiarto Raharjo[3]
Program Studi Informatika, Universitas Kristen Duta Wacana
JL. Dr. Wahidin Sudiro Husodo No. 5 – 25, Yogyakarta
[1]joiner.tennye@ti.ukdw.ac.id
[2]anton@ti.ukdw.ac.id
[3]willysr@staff.ukdw.ac.id

Abstract*— The growing awareness of the impact of motor vehicle emissions on the environment has encouraged Indonesia's Ministry of Environment and Forestry to enforce emission testing regulations. These emission standards serve as a basis for calculating Motor Vehicle Tax (PKB). The Transportation Technology Research Center (BRIN) developed a tax coefficient prediction system to support this policy. Initial research utilized Orange Data Mining for machine learning analysis with algorithms like Random Forest, Neural Network, and AdaBoost. However, Orange Data Mining has limitations in flexibility, particularly in parameter tuning and preprocessing data, as well as inefficiencies in handling large datasets. This study adopts a more flexible approach, employing AutoML LazyPredict for quick identification of optimal models and GridSearchCV for hyperparameter optimization. The methodology involves two approaches: classification and regression. Classification employs models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Extra Tree, and LightGBM, while regression utilizes Support Vector Regressor (SVR) optimized with GridSearchCV. Both approaches enable a comprehensive comparison and analysis of model performance. The results indicate that SVM and Decision Tree excelled in classification, achieving an accuracy of 100%. In regression, the models demonstrated high 16 performance with R^2 values ranging from 0.95 to 1.00, indicating exceptional predictive accuracy. Evaluations were conducted using metrics such as MAE, MSE, and R^2 for regression, along with accuracy scores and classification reports for classification tasks. This research underscores the effectiveness of machine learning model optimization, with both analyzed algorithms showing outstanding performance for classification and regression tasks.

Keywords : ***AutoML, vehicle emissions, GridSearchCV, machine learning, motor vehicle tax.***

Intisari— Peningkatan kesadaran akan dampak emisi kendaraan bermotor terhadap lingkungan mendorong Kementerian Lingkungan Hidup dan Kehutanan Indonesia untuk menerapkan regulasi uji emisi. Nilai standar emisi ini dapat menjadi dasar perhitungan Pajak Kendaraan Bermotor (PKB). Tim Pusat Riset Teknologi Transportasi (BRIN) mengembangkan model prediksi koefisien pajak untuk mendukung kebijakan ini. Penelitian awal menggunakan *Orange Data Mining* untuk analisis machine learning dengan algoritma seperti *Random Forest*, *Neural Network*, dan *AdaBoost*. Namun, *Orange* memiliki kelemahan pada fleksibilitas, terutama dalam *tuning parameter*, *preprocessing dataset*, serta keterbatasan menangani dataset besar. Dalam penelitian ini, *Google Colab* digunakan sebagai platform utama untuk memaksimalkan fleksibilitas analisis dan kustomisasi model. Algoritma *AutoML LazyPredict* diterapkan untuk mengidentifikasi model terbaik secara cepat, sedangkan *GridSearchCV* digunakan untuk optimasi hyperparameter. Pendekatan pertama adalah klasifikasi menggunakan model seperti *Support Vector Machine* (*SVM*), *K-Nearest Neighbors* (*KNN*), *Decision Tree*, *Extra Tree*, dan *LightGBM*. Pendekatan kedua adalah regresi menggunakan *Support Vector Regressor* (*SVR*), yang dioptimalkan dengan *GridSearchCV*. Kedua metode ini digunakan untuk membandingkan performa model dan menganalisis hasil. Hasil penelitian menunjukkan bahwa *SVM* dan *Decision Tree* memberikan hasil klasifikasi terbaik dengan akurasi mencapai 100%. Untuk regresi, nilai R^2 berkisar antara 0,95 hingga 1,00 menunjukkan tingkat akurasi prediksi yang sangat tinggi. Evaluasi dilakukan dengan metrik seperti *MAE*, *MSE*, dan R^2 untuk regresi, 14 serta *accuracy score* dan *classification report* untuk klasifikasi. Penelitian ini membuktikan bahwa optimasi model machine learning berhasil, dengan kedua algoritma dianalisa menunjukkan kinerja yang sangat baik untuk klasifikasi dan regresi.

***Kata Kunci*— *AutoML, emisi kendaraan, GridSearchCV, machine learning, pajak kendaraan bermotor*.**

I. PENDAHULUAN

Karbon adalah senyawa gas yang dihasilkan dari berbagai aktivitas makhluk hidup dalam kurun waktu tertentu. Salah satu bentuknya, karbon monoksida (*CO*), merupakan gas berbahaya yang tidak berwarna dan tidak berbau, dikenal sebagai *silent killer* karena dampaknya yang mematikan bagi kesehatan manusia. Sumber utama emisi karbon, termasuk *CO*, di lingkungan perkotaan adalah kendaraan bermotor berbahan bakar fosil. Peningkatan jumlah kendaraan bermotor di Indonesia, yang mencapai lebih dari 148 juta unit pada tahun 2024, berkontribusi signifikan terhadap polusi udara dan berbagai dampak negatif bagi kesehatan serta lingkungan, seperti perubahan iklim dan risiko keracunan karbon monoksida. Paparan gas *CO* dapat menyebabkan masalah serius, termasuk penurunan kapasitas darah untuk mengangkut oksigen dan risiko penyakit kronis, terutama pada individu dengan penyakit

jantung atau paru-paru.

Sebagai upaya pengendalian, pemerintah Indonesia menerapkan regulasi emisi karbon dengan standar *Euro 4* dan menggunakan hasil uji emisi untuk menghitung koefisien Pajak Kendaraan Bermotor (PKB). Penelitian terkait, seperti yang dilakukan oleh tim BRIN, menggunakan data uji emisi untuk memodelkan koefisien pajak berbasis pembelajaran mesin. Namun, penelitian sebelumnya memiliki beberapa keterbatasan, seperti bias pada dataset dan kurangnya fleksibilitas dalam platform analisis.

Penelitian ini bertujuan untuk mengoptimalkan akurasi model prediksi koefisien pajak menggunakan algoritma *AutoML LazyPredict* dan optimisasi *hyperparameter* melalui *GridSearchCV*. Model yang dikembangkan mencakup pendekatan klasifikasi dan regresi dengan algoritma seperti *SVM*, *Decision Tree*, dan *LightGBM*. Dengan pendekatan ini, diharapkan dapat dihasilkan model prediksi yang lebih efektif untuk mendukung kebijakan pengelolaan PKB di Indonesia.

II. TINJAUAN PUSTAKA

Penelitian terkait prediksi menggunakan pembelajaran mesin (*machine learning*) telah banyak dikembangkan untuk berbagai kebutuhan di berbagai bidang. Kajian ini bertujuan untuk mengidentifikasi berbagai pendekatan, algoritma, serta hasil evaluasi yang telah dilakukan pada beberapa penelitian terkait prediksi harga saham, *cryptocurrency*, emisi karbon, dan optimasi model pembelajaran mesin lainnya. Penelitian terkait memprediksi suatu nilai [1] dengan menggunakan algoritma *AutoML H2O* untuk memprediksi harga saham PT BRI. Dalam penelitian ini, algoritma dasar yang digunakan adalah *Gradient Boosting Machine* (*GBM*), dan hasil evaluasi menunjukkan nilai *Mean Absolute Percentage Error* (*MAPE*) sebesar 2,327 serta R^2 sebesar 0,973. Selain itu, teknik stacking digunakan untuk meningkatkan akurasi prediksi, dengan ensemble algoritma seperti *Distributed Random Forest* (*DRF*), *Extremely Randomized Trees* (*XRT*), *Generalized Linear Model* (*GLM*), dan GBM. Kombinasi ini menghasilkan prediksi yang lebih baik dalam menjelaskan variabilitas data harga saham.

Pendekatan yang lain, model pembelajaran mesin untuk memprediksi harga Bitcoin. Penelitian [2] menggunakan algoritma *AutoML H20* untuk memprediksi harga jual *Bitcoin*. Dalam penelitian mereka, algoritma *AutoML* digunakan untuk mengotomatisasi proses pelatihan dan penyesuaian parameter model. Hasil evaluasi menunjukkan nilai R^2 sebesar 0,968 dan nilai error sebesar 3,48%. Tantangan utama dalam penelitian ini adalah mengidentifikasi pola fluktuasi harga *Bitcoin* yang kompleks. Penelitian mengenai optimasi model [3] dengan library *Sklearn*, untuk mengeksplorasi *Sklearn Model Selection* untuk mencapai performa optimal pada beberapa algoritma pembelajaran mesin. Teknik *GridSearchCV* digunakan untuk *hyperparameter tuning*, dengan evaluasi menggunakan pengukuran *Mean Cross Validation*. Dari tujuh algoritma yang diuji, *XGBoost* menunjukkan performa terbaik, sementara *Decision Tree* memiliki skor terendah. Penelitian ini menekankan pentingnya penyesuaian parameter dalam mencapai akurasi prediksi yang tinggi.

Penelitian mengenai prediksi emisi karbon di Indonesia terdapat beberapa seperti penelitian ini [4], menggunakan model pembelajaran ini membahas prediksi karbon monoksida di Jakarta menggunakan model time series dari *Library PyCaret*. Model yang dibandingkan antara lain *Huber Regressor*, *Linear Regressor*, dan *Ridge Regressor*, dengan *Huber Regressor* menunjukkan akurasi terbaik. Evaluasi model menunjukkan nilai *MAE* sebesar 5,3187, *RMSE* sebesar 8,9838, serta *MAPE* sebesar 0,2364. Penelitian ini memberikan kontribusi signifikan pada pemantauan polusi udara. Dan penelitian sebelumnya mengenai prediksi koefiesen pajak kendaraan bermotor di Indonesia.

Dalam penelitian ini [5], menggunakan pembelajaran mesin untuk memprediksi koefisien pajak berdasarkan data uji emisi. Algoritma yang digunakan mencakup *Random Forest* (*RF*), *AdaBoost* (*AB*), dan *Neural Network* (*NN*) melalui platform *Orange Data Mining*. *Neural Network* memberikan hasil terbaik dengan R^2 sebesar 0,987 untuk dataset diesel dan 0,970 untuk dataset bensin. Namun, untuk nilai *Mean Absolute Error* (*MAE*), algoritma *Random Forest* menunjukkan hasil terbaik pada dataset diesel.

A. Mesin Pembelajaran untuk Klasifikasi dan Regresi

Dalam penelitian ini menggunakan berbagai metode pembelajaran mesin telah dikembangkan untuk mendukung analisis data dan pengambilan keputusan, terutama dalam tugas klasifikasi dan regresi. Algoritma yang digunakan untuk eksplorasi berbagai metode pembelajaran yaitu Algoritma *LazyPredicts.* Algoritma *LazyPredicts* [6], merupakan bagian dari *Supervised Learning* yang melakukan pendekatan terhadap pembelajaran mesin. Setiap model yang dilatih, kemudian menggunakan data berlabel untuk memprediksi *output* berdasarkan *input* baru yang belum terlihat, Algoritma ini dapat dapat digunakan sebagai model klasifikasi untuk memprediksi label diskrit dan regresi untuk memprediksi nilai numerik.

LazyPredict adalah pustaka *Python* yang dirancang untuk mempercepat eksplorasi dan evaluasi berbagai algoritma *supervised learning* secara otomatis, tanpa memerlukan pengaturan parameter manual. Pustaka ini memiliki dua modul utama, *LazyClassifier* untuk klasifikasi dan *LazyRegressor* untuk regresi, yang secara otomatis membandingkan model seperti *SVM*, *KNN*, *Decision Tree*, dan *LightGBM*. Dengan fitur pemilihan model otomatis, evaluasi kinerja melalui metrik seperti akurasi dan *MAE*, serta efisiensi waktu, *LazyPredict* memungkinkan pengguna untuk dengan cepat menemukan model terbaik untuk dataset tertentu. Beberapa model – model dipilih sebagai berikut *Support Vector Machine* (*SVM*), *Support Vector Regressor* (*SVR*), *Decision Tree Classifier*, *ExtraTrees Classifier*, *LightGLBM Classifier*, *Decision Tree Regressor*, *ExtraTrees Regressor*, dan *LightGLBM Regressor*. Evaluasi metrik yang digunakan dalam penelitian menggunakan *library SkicitLearn.*. Library *SkicitLearn* [15] adalah pustaka *Python*

JUTEI Edisi Volume. 9 No.1 April 2025
ISSN 2579-3675, e-ISSN 2579-5538
DOI 10.21460/jutei.2025.91.389

yang populer dalam pengembangan model *machine learning*, khususnya dalam evaluasi kinerja model. Salah satu fitur utama *scikit-learn* adalah modul *metrics*, yang menyediakan berbagai metrik untuk mengukur performa model pada masalah klasifikasi, regresi, dan *clustering*.

Dalam proses evaluasi model *machine learning*, metrik yang digunakan berbeda-beda tergantung pada jenis masalahnya, seperti klasifikasi atau regresi. Berikut adalah penjelasan metrik-metrik utama yang digunakan:

1) Metrik untuk Klasifikasi

1. *Accuracy*: Mengukur proporsi prediksi yang benar dari total data. Cocok digunakan ketika data seimbang.
2. *Precision*: Mengukur proporsi prediksi positif yang benar terhadap semua prediksi positif. Berguna saat kesalahan prediksi positif harus diminimalkan.
3. *Recall* (atau Sensitivitas): Mengukur kemampuan model dalam mendeteksi semua sampel positif yang sebenarnya ada.
4. *F1-Score*: Rata-rata harmonis dari precision dan recall. Sangat bermanfaat pada data yang tidak seimbang, karena menggabungkan informasi dari precision dan recall.
5. *Confusion Matrix*: Matriks yang menggambarkan performa model dengan menunjukkan jumlah prediksi benar dan salah untuk setiap kelas.

2) Metrik untuk Regresi

1. *Mean Squared Error* (*MSE*): Mengukur rata-rata kuadrat dari selisih antara nilai prediksi dan nilai sebenarnya. Memberikan penalti yang lebih besar untuk kesalahan besar.
2. *Mean Absolute Error* (*MAE*): Mengukur rata-rata dari selisih absolut antara nilai prediksi dan nilai sebenarnya. Lebih robust terhadap data dengan pencilan.
3. *R-squared* (R^2): Mengukur seberapa baik model menjelaskan variasi dalam data. Nilai mendekati 1 menunjukkan model yang baik.

III. METODOLOGI PENELITIAN

Metodologi penelitian ini dirancang untuk mengevaluasi performa model pembelajaran mesin dalam mengklasifikasi dan memprediksi data emisi karbon kendaraan bermotor berbahan bakar bensin dan diesel. Pemodelan tersebut menggunakan komputasi dengan spesifikasi, yang dapat dilihat Tabel I.

Langkah-langkah metodologi yang dilakukan meliputi:

A. Implementasi Awal

Tahap ini mencakup pengumpulan dan pemrosesan awal data:

1) Data Understanding: Dataset emisi karbon kendaraan bermotor diperoleh dari uji emisi Pusat Riset Teknologi Transportasi, yang sesuai dengan rancangan regulasi Kementerian Lingkungan Hidup dan Kehutanan.

2) Ekspor Dataset: File dataset dalam format XLSX diekspor ke format CSV, dan dilakukan pelabelan kelas data serta penentuan fitur yang relevan.

TABEL I
SPESIFIKASI SISTEM MODEL PREDIKSI DAN GOOGLE COLAB

Spesifikasi	**Fitur**	**AutoML**	**Pre-Processing**
Sistem Prediksi	Tahun Uji, Tahun Pembuatan, Usia, Opasitas, CO,HC, RHC, RCO	LazyPredict	Data Resampling
	Dataset	**Tuning Hyperparameter**	**Metrik Evaluasi**
	80 % *Training* 20 % *Testing*	GridSearchCV	MSE, MAE, Classification Report, Accuracy Score (F1 Score), R-Squared
Google Colab (Versi Gratis)	**Processor**	**GPU**	**TPU**
	Intel Xeon 2 vCPU	NVIDIA Tesla K80, T4, P100 dan P4 (Bervariasi)	TPU v2-8
	Memori	**Durasi Runtime**	**Runtime type**
	Standar 12 GB dan High RAM 25 GB	12 Jam per Sesi	Python 3 dan R

B. Data Preprocessing

Untuk memastikan data siap digunakan, dilakukan beberapa langkah pemrosesan:

1) Cleaning Data: Menghapus data yang hilang (null) dan duplikasi.
2) Data Resampling: Mengatasi ketidakseimbangan data dengan mengurangi data mayoritas dalam dataset.
3) Data Proporsional: Mengatasi ketidakseimbangan data dengan membagi semua data menjadi 1 proporsi yang sama.

C. Pelatihan dan Evaluasi Model

Pada tahap ini, model pembelajaran mesin diterapkan dan dievaluasi:

1) Model Klasifikasi: Model yang digunakan meliputi Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree Classifier, Extra Tree Classifier, dan LightGBM Classifier.

2) Model Regresi: Model regresi yang diterapkan meliputi Support Vector Regressor (SVR), K-Nearest Neighbor Regressor, Decision Tree Regressor, Extra Tree Regressor, dan LightGBM Regressor.

3) Optimisasi Model:

1. *AutoML:* Untuk pemilihan model terbaik secara otomatis.
2. *GridSearchCV:* Untuk tuning hyperparameter.

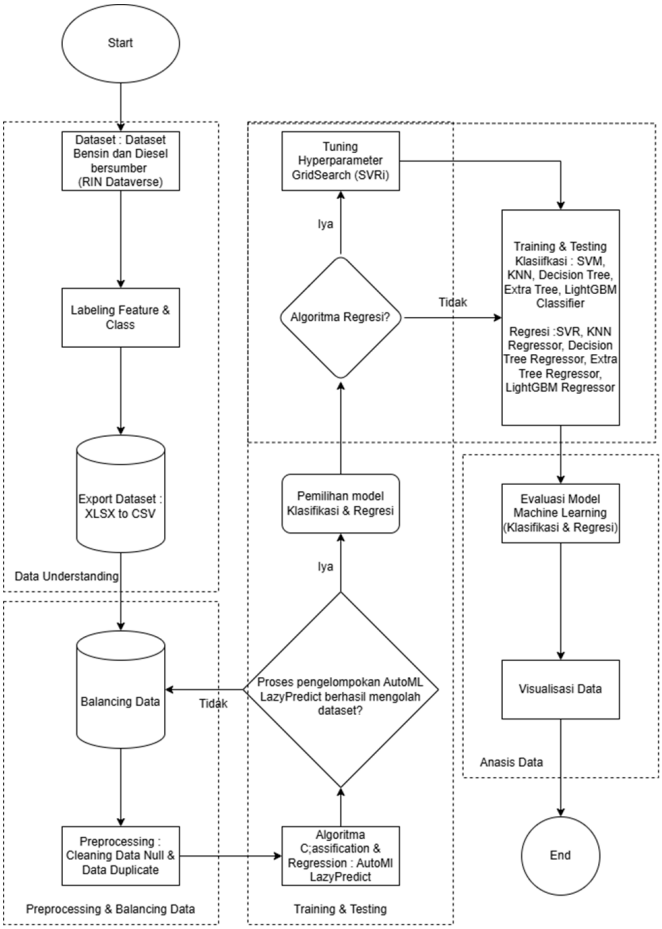
4) Evaluasi Model:

1. Regresi*:* Menggunakan metrik *Mean Absolute Error* (*MAE*), *Root Mean Squared Error* (*RMSE*), dan koefisien determinasi (R^2).
2. Klasifikasi*:* Menggunakan *accuracy score* dan laporan klasifikasi.

5) Analisis Hasil:

1. Mengukur peningkatan akurasi model menggunakan *AutoML* dan *GridSearchCV*.
2. Membandingkan hasil prediksi model baru dengan model sebelumnya, khususnya dalam konteks estimasi koefisien pajak emisi karbon, untuk menilai keunggulan model yang dikembangkan.
3. Visualisasi data menggunakan *scatter plot*, *box plot*, dan *residual plots* digunakan untuk memahami distribusi kesalahan prediksi, sementara grafik perbandingan model yang dioptimalkan dengan *AutoML* dan *GridSearchCV* menunjukkan peningkatan performa serta memberikan wawasan mendalam tentang kekuatan dan kelemahan model.

Metodologi ini dirancang untuk menghasilkan model yang lebih akurat dan efisien dalam menangani data emisi karbon, serta memberikan wawasan baru dalam penerapan pembelajaran mesin di bidang transportasi dan lingkungan. Langkah – langkah diatas dijelaskan pada diagram alur, dapat dilihat pada Gambar 1. yang diadaptasi dari *flowchart* metode penelitian [16].



Gambar. 1. Diagram Alur Metodologi Penelitian

IV. HASIL DAN ANALISIS

A. Analisis Data Eksploratif

Eksplorasi dataset untuk melihat deskripsi statistik, distribusi nilai variabel, dan korelasi variabel. Berikut adalah langkah-langkah dan hasil analisis eksploratif terhadap data yang terdiri dari variabel: Tahun Uji, Tahun Pembuatan, *CO*, *HC*, Usia, *RCO*, *RHC*, Opasitas dan *Rating*. Dapat dilihat pada Tabel II.

TABEL II
DESKRIPSI STATISTIK DATASET GASOLINE DAN DATASET DIESEL

Variabel	Mean	Median	Min	Max	Std. Deviasi
Dataset Gasoline					
Tahun Uji	2016.19	2016	2013	2022	2.584
Tahun Pembuatan	2010.89	2012	1944	2022	5.711
CO (g/km)	0,531	0,03	0	74,6	1,516
HC (g/km)	90,064	20	0	55018	286,970
Usia (tahun)	5,299	4	0	71	5,251
Dataset Diesel					
Tahun Uji	2015,949	2016	2013	2022	2,390
Tahun Pembuatan	2008,192	2010	1921	2022	6,884
Opasitas	50,587	46	0	100	28,816
Usia	7,757	6	0	94	6,528

JUTEI Edisi Volume. 9 No.1 April 2025
ISSN 2579-3675, e-ISSN 2579-5538
DOI 10.21460/jutei.2025.91.389

Dalam penelitian sebelumnya, distribusi nilai variabel diambil dari data Kementrian Lingkungan Hidup dan Kehutanan yang berkolaborasi dengan tim BRIN. Distribusi nilai variabel untuk dataset *gasoline*, dapat dilihat pada Tabel III dan Tabel IV untuk dataset diesel [5]

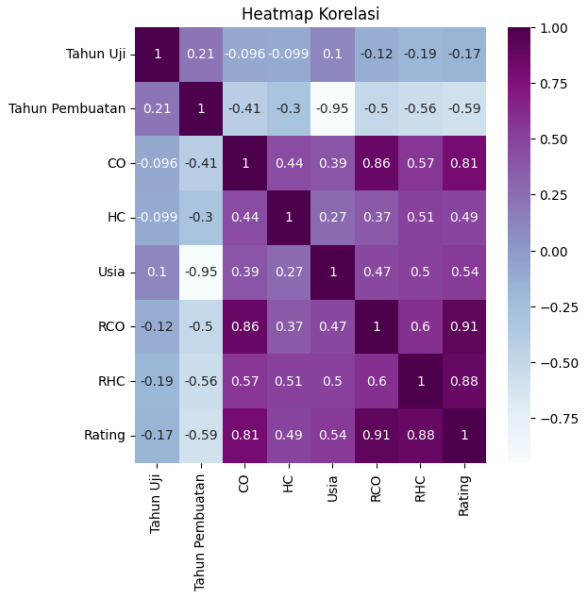
TABEL III
DISTRIBUSI VARIABEL DATASET GASOLINE

Tahun	CO				HC			
	<- 0.5 -		- 1 ->		< 100	100 – 150	150 – 1000	> 1000
	0.5	1	4	4				
< 2007	1	2	3	4	1	2	3	4
2007 - 2018	0.5	1.5	2.5	3.5	0.5	1.5	2.5	3.5
> 2018	0	1	2	3	0	1	2	3

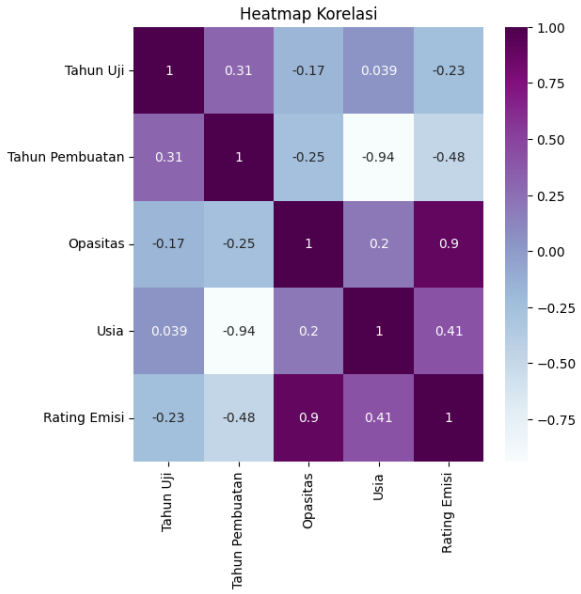
TABEL IV
DISTRIBUSI VARIABEL DATASET DIESEL

Tahun	Opasitas			
	<		>	
	30	30 - 40	40 – 65	65
< 2010	1	2	3	4
2010 – 2021	0.5	1.5	2.5	3.5
> 2021	0	1	2	3

Dari kedua dataset tersebut, dilakukan analisis untuk melihat korelasi variabel. Gambar 2 menunjukkan korelasi untuk dataset *gasoline*. Gambar 3 menunjukkan korelasi untuk dataset diesel. Kesimpulan nya, variabel yang memengaruhi nilai *rating* atau kelas dataset *gasoline* adalah *RCO*, *RHC*, *CO* dan *HC*. Untuk dataset diesel variabel Opasitas.



Gambar. 2. Heatmap Korelasi Dataset Gasoline



Gambar. 3. Heatmap Korelasi Dataset Diesel

B. Data Preprocessing

Proses data resampling pada dataset emisi karbon kendaraan bermotor berbahan bakar bensin dan diesel dilakukan setelah melalui tahap pembersihan untuk menghapus nilai null dan data duplikat. Dengan menggunakan pustaka *Python* seperti *pandas*, langkah ini bertujuan untuk menyeimbangkan dataset yang memiliki perbedaan rasio signifikan antara kelas mayoritas dan minoritas. Misalnya, pada dataset bensin, rasio antara jumlah data dengan *rating* terkecil dan terbesar mencapai 16.000:1, yang dapat menyebabkan ketidakseimbangan dalam pelatihan model. Informasi terkait dataset dapat dilihat Tabel V & Tabel VI.

Proses dimulai dengan memisahkan data yang berlebihan menggunakan fungsi seperti *.append*, *.sample*, dan *.concat*. Data yang tidak diperlukan dihapus, dan dataset disusun kembali dengan memastikan setiap kelas memiliki jumlah data yang sesuai. Setelah data digabungkan, langkah berikutnya adalah mengacak urutan dataset menggunakan fungsi *.sample*. Langkah ini penting untuk menghindari pola berurutan yang dapat menyebabkan model pembelajaran mesin mengalami *overfitting*. Data yang sudah diacak kemudian disimpan ke *file* baru, memastikan dataset siap untuk diolah lebih lanjut.

Tahapan data *resampling* dan data proporsional secara rinci mencakup:
1) Persiapan Data: Mengecek apakah nilai rating berada dalam rentang yang telah ditentukan.

2) Fitur Data Rating: Memisahkan data rating dari fitur lain untuk mempermudah pengelolaan.

3) Ekstraksi Data: Mengambil jumlah data yang diperlukan dari kelas dengan jumlah data yang berlebih.

4) Ekstraksi Data Proporsional: Untuk data proporsional, jumlah data yang diambil dengan ukuran 15% dari setiap rating daalam dataset semula.

5) Gabungkan Data: Menggabungkan kembali seluruh dataset setelah proses pemisahan.

6) Susun Data: Mengacak urutan data berdasarkan nilai rating agar distribusi lebih merata.

7) Simpan file: Menyimpan dataset yang telah diacak ke dalam file baru untuk tahap analisis berikutnya.

Dengan metode ini, dataset menjadi lebih seimbang, mengurangi risiko *bias*, dan memastikan model dapat belajar secara optimal untuk memprediksi hasil dengan akurasi yang lebih baik.

TABEL V
INFO DATASET EMISI (BENSIN)

No	Kolom	Jumlah Data	Dtype
1	Tahun Uji	273120	int64
2	Tahun Pembuatan	273120	int64
3	CO	273120	float64
4	HC	273120	float64
5	Usia	273120	int64
6	RCO	273120	float64
7	RHC	273120	float64
8	Rating	273120	int64

TABEL VI
INFO DATASET EMISI (DIESEL)

No	Kolom	Jumlah Data	Dtype
1	Tahun Uji	80103	int64
2	Tahun Pembuatan	80103	int64
3	Opasistas	80103	int64
4	Usia	80103	float64
5	Rating Emisi	80103	int64

C. Algoritma AutoML & Tuning Parameter

Dalam penelitian ini, AutoML digunakan untuk mengeksplorasi berbagai model pembelajaran mesin dan mencari model terbaik yang dapat memberikan performa optimal dalam memprediksi data.

Berikut adalah tahapan-tahapan yang dilakukan dalam menggunakan *AutoML* dan *GridSearchCV*:

1) Tahapan Penggunaan LazyPredict (AutoML):

1. Mengimpor *Library LazyPredict* dengan mengimpor pustaka *LazyPredict* yang berisi modul *LazyClassifier* dan *LazyRegressor* untuk evaluasi model klasifikasi dan regresi.
2. Instalasi model *LazyClassifier* dan *LazyRegressor*: dengan menginstal kedua modul tersebut sesuai dengan jenis model yang ingin diuji klasifikasi dan regresi.
3. Pengaturan Parameter Model:
    1. *Verbose* yang menentukan tingkat rincian informasi yang ditampilkan selama proses pelatihan (nilai 0, 1, atau 2).
    2. *ignore_warnings* yang mengatur apakah peringatan yang muncul selama proses eksekusi akan ditampilkan (*True* atau *False*). Parameter model lainnya, dapat dilihat pada Tabel VII.
4. Memasukkan data pelatihan dan pengujian: Memasukkan dataset pelatihan dan pengujian ke dalam model untuk digunakan dalam evaluasi.
5. Menjalankan *LazyPredict* dengan melakukan eksekusi *LazyPredict* untuk mengevaluasi dan membandingkan berbagai model pembelajaran mesin secara otomatis tanpa konfigurasi parameter manual.
6. Menganalisis *output* dengan menilai hasil evaluasi untuk berbagai model yang diuji dan memilih model dengan performa terbaik berdasarkan hasil tersebut.

2) Tahapan Penggunaan GridSearchCV untuk Tuning Hyperparameter:

1. Menentukan parameter model dengan memilih parameter model yang akan diuji, misalnya *Support Vector Regressor* (*SVR*).
2. Memanggil fungsi *GridSearchCV* yaitu memanggil fungsi *GridSearchCV* dengan memasukkan parameter berikut:
    1 Model pembelajaran mesin yang digunakan.
    2 Parameter model yang ingin diuji untuk meningkatkan performa.
    3 *CV* (*Cross Validation*) yaitu menentukan jumlah lipatan validasi silang.
    4 *Scoring* yaitu metrik evaluasi yang digunakan untuk menilai kinerja model (misalnya, *accuracy*, *MAE*, atau R^2).
3. *Fitting* data pelatihan: Menerapkan data pelatihan pada model untuk proses evaluasi tanpa menggunakan data pengujian, karena fokus utama adalah menguji kombinasi parameter.
4. Menjalankan *GridSearchCV* dengan melakukan eksekusi *GridSearchCV* untuk menemukan kombinasi parameter terbaik yang meningkatkan kinerja model.
5. Menganalisis hasil *GridSearchCV* yaitu mengkaji hasil evaluasi untuk melihat kombinasi parameter mana yang memberikan performa terbaik berdasarkan metrik yang dipilih.

Dengan tahapan ini, penelitian ini menggunakan pendekatan *AutoML* untuk mengevaluasi berbagai model pembelajaran mesin dan *GridSearchCV* untuk melakukan

JUTEI Edisi Volume. 9 No.1 April 2025
ISSN 2579-3675, e-ISSN 2579-5538
DOI 10.21460/jutei.2025.91.389

optimasi *hyperparameter*. Pendekatan ini membantu dalam memilih model yang memberikan hasil terbaik dalam hal akurasi dan efisiensi, serta memungkinkan untuk mendapatkan wawasan lebih dalam tentang kekuatan dan kelemahan masing-masing model yang diuji.

TABEL VII
PARAMETER MODEL

No	Model Pembelajaran Mesin	Parameter	Nilai
1	Lazy Predict Classifier	Verbose	0
		Ignore_warnings	True
2	Lazy Predict Regressor	Verbose	0
		Ignore_warnings	False
		custom_metric	None
3	GridSearch CV	Model()	SVR
		param_grid	C, epsilon, kernel
		cv	5
		Scoring	neg_mean_squared_error
4	SVM(Support Vector Machine)	Kernel	Linear
5	KNN (K-Neighbors Classifier)	n_neighbors	5
6	Decision Tree Classifier	random_state	42
7	ExtraTrees Classifier	random_state	42
8	LightGBM Classifier	random_state	42
9	SVR (Support Vector Regressor)	kernel	Linear
		C	10
		epsilon	0.01
10	KNN Regressor	n_neighbors	5
11	DecisionTree Regressor	random_state	42
12	ExtraTrees Regressor	random_state	42
13	LightGBM Regressor	random_state	42

D. Evaluasi & Pembahasan

Tabel VIII dan Tabel IX menunjukkan evaluasi dataset bensin menggunakan model klasifikasi dan regresi, dapat dijabarkan sebagai berikut:

TABEL VIII
EVALUASI MODEL KLASIFIKASI PADA DATASET GASOLINE

Model	Macro Average	Weighted Average	Support	F1 Score
Undersampling				
SVM	1.00	1.00	2638	100 %
KNN Classifier	1.00	1.00	2638	99.24 %
Decision Tree Classifier	1.00	1.00	2638	99.86 %
Extra Tree Classifier	1.00	1.00	2638	99.92 %
LightGBM Classifier	1.00	1.00	2638	99.96 %
Proporsi 15%				
SVM	1.00	1.00	5316	100 %
KNN Classifier	1.00	1.00	5316	99.64 %
Decision Tree Classifier	1.00	1.00	5316	99.94 %
Extra Tree Classifier	1.00	1.00	5316	100 %
LightGBM Classifier	1.00	1.00	5316	99.96 %

Dari hasil evaluasi diatas, bisa dapat disimpulkan bahwa optimasi akurasi menggunakan metode *AutoML* dan *GridSearchCV* berhasil. Dari Tabel VIII dapat dilihat bahwa model *SVM* mendapat nilai skor *f1-score* sempurna yaitu 100 %, ini membuktikan bahwa memprediksi koefisien pajak bisa menggunakan klasifikasi dan mendapat tingkat akurasi yang tinggi, menggunakan algoritma *AutoML Classifier*, bisa dengan mudah memilih model klasifikasi yang terbaik untuk dataset emisi karbon kendaraan bermotor di Indonesia. Selain itu, pada Tabel IX dapat dilihat bahwa model *Support Vector Regression* (*SVR*) mendapat *f1-score* sempurna yaitu 100%, skor ini mendapatkan nilai yang lebih tinggi dibanding dengan penelitian sebelumnya, menyimpulkan bahwa optimasi akurasi berhasil pada dataset bensin.

TABEL IX
EVALUASI MODEL REGRESI PADA DATASET GASOLINE

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared Score
Undersampling			
SVR	0.0049	0.0000	1.000
K-Nearest Regressor	0.0072	0.0037	0.9992
Decision Tree Regressor	0.0003	0.0003	0.9999
Extra Tree Regressor	0..0003	0.0001	0.9999
LightGBM Regressor	0.0012	0.0004	0.9999
Proporsi 15%			
SVR	0.0064	0.000	1.0
K-Nearest Regressor	0.003	0.0020	0.999
Decision Tree Regressor	0.005	0.0005	0.998
Extra Tree Regressor	3.197	6.5838	0.999
LightGBM Regressor	0.014	0.002	0.999

TABEL X
EVALUASI MODEL KLASIFIKASI PADA DATASET DIESEL

Model	Macro Average	Weighted Average	Support	F1 Score
SVM	0.93	0.93`	1604	93 %
KNN Classifier	0.93	0.93	1604	94.83 %
Decision Tree Classifier	1.00	1.00	1604	100 %
Extra Tree Classifier	0.99	0.99	1604	98.81 %
LightGBM Classifier	0.99	0.99	1604	99.31 %

TABEL XI
EVALUASI MODEL REGRESI PADA DATASET DIESEL

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-Squared Score
SVR	0.2173	0.0845	0.9565
K-Nearest Regressor	0.0780	0.0353	0.9817
Decision Tree Regressor	0.0	0.0	1.0
Extra Tree Regressor	0..0053	0.0007	0.9996
LightGBM Regressor	0.0090	0.0028	0.9985

Selanjutnya pada dataset diesel, dari Tabel X dan Tabel XI dapat dilihat dengan algoritma *AutoML*, kita dapat memilih model – model yang terbaik baik regresi dan klasifikasi, salah satu nya adalah *Decision Tree Classifier* & *Decision Tree Regressor* yang masing-masing mendapatkan *f1-score* 100%, dari angka tersebut, menunjukkan kenaikan angka akurasi dibandingkan dengan penelitian sebelumnya, menyimpulkan bahwa optimasi akurasi pada dataset diesel berhasil. Meskipun dengan terbatasi komputasi pada *Google Colab* dalam menjalankan algoritma *AutoML* dan *GridSearchCV*, dengan metode yang sederhana dan efektif diharapkan menjadi bentuk dukungan dalam pengembangan model prediksi koefisien pajak di Indonesia dan menjadikan salah satu alat bantu dalam membuat kebijakan pajak kendaraan di Indonesia oleh pemenrintah.

Penelitian ini dirancang untuk mengoptimasi akurasi model pembelajaran mesin terhadap prediksi koefisien pajak dari data emisi karbon kendaraan bermotor di Indonesia. Perbedaan penelitian dengan penelitian sebelumnya atau sejenis (Optimasi Model), yaitu penelitian ini menggunakan algoritma otomasi yaitu *LazyPredict*, yang memungkinkan penelitian ini, mendapat gambaran yang luas terhadap setiap kinerja model pembelajaran mesin yang dipakai. Tahapan eksplorasi ini, memungkinkan model pembelajaran menjadi lebih efektif. Dibandingkan dengan penelitian sebelumnya, metode *manual cross-validation* ketika mengeksekusi model pembelajaran mesin nya. Penelitian ini, metode *cross-validation* yang dilakukan secara otomatis menggunakan algoritma *GridSearchCV* sehingga validasi parameter nya teruji.

V. KESIMPULAN

Penelitian ini berhasil mengoptimalkan model pembelajaran mesin untuk memprediksi koefisien pajak kendaraan bermotor di Indonesia berdasarkan data emisi kendaraan. Dengan menggunakan metode *AutoML* dan optimasi *hyperparameter* melalui *GridSearchCV*, model yang diterapkan menunjukkan peningkatan performa yang signifikan. Model regresi dan klasifikasi yang digunakan,

JUTEI Edisi Volume. 9 No.1 April 2025
ISSN 2579-3675, e-ISSN 2579-5538
DOI 10.21460/jutei.2025.91.389

seperti *Support Vector Machine* (*SVM*) dan *Decision Tree*, berhasil mencapai akurasi tinggi, dengan *SVM* dan *Decision Tree Classifier* mencapai akurasi 100%, serta model regresi menunjukkan nilai R^2 yang sangat tinggi, yaitu 0,95 (*SVR*) dan *Decision Tree Regressor* mencapai akurasi 1,00.

Optimasi ini diperoleh melalui metode *preprocessing* seperti *data resampling* dan algoritma *AutoML LazyPredict*, yang membuat sistem pembelajaran mesin lebih efisien dan efektif, serta memberikan peningkatan akurasi yang signifikan dibandingkan penelitian sebelumnya. Oleh karena itu, sistem pembelajaran mesin yang dikembangkan dalam penelitian ini dapat dianggap berhasil dalam meningkatkan akurasi dengan menggunakan dataset yang sama. Akan tetapi, karena masih ada beberapa faktor yang dapat ditingkatkan, seperti dibutuhkan pembaruan regulasi uji emisi untuk memperbarui dan menambah data yang dapat meningkatkan akurasi model di masa mendatang. Faktor lainnya, diperlukan komputasi yang lebih kuat, mengingat penggunaan platform online seperti *Google Colab* memiliki keterbatasan dalam mengelola dataset besar. Oleh karena itu, diperlukan perangkat keras yang lebih kuat untuk mengelola dataset yang terus berkembang.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Fakultas Teknologi Informasi Universitas Kristen Duta Wacana atas kesempatan yang telah diberikan dalam pengerjaan penelitian ini. Dan tim peneliti Badan Riset dan Inovasi Nasional terkhusunya tim peneliti Pusat Riset Teknologi Transportasi Pak Fitra Hidiyanto dan tim yang sudah berkenan untuk membagikan dataset untuk penelitian ini. Penulis juga sangat berterima kasih kepada Pak Antonius Rachmat Chrismanto dan Pak Willy Sudiarto Raharjo atas bimbingan, masukan, dan dukungan yang berharga selama pelaksanaan studi ini. Akhirnya, penulis juga ingin mengucapkan terima kasih kepada Keluarga, Nona Angela Josephine Pangemanan dan teman-teman yang telah memberikan dukungan moral selama proses penelitian.

DAFTAR PUSTAKA

[1] I. M. Tirta, A. Riski, and N. Sholikhah, "Prediksi Harga Saham PT Bank Rakyat Indonesia Tbk Menggunakan AUTOML H2O Pendahuluan," vol. 23, no. September, pp. 397–404, 2024.

[2] G. Geadalfa and S. Saidah, "Auto Machine Learning dengan Menggunakan H2O Pendahuluan Tinjauan Pustaka Metode Penelitian," *J. Ilm. KOMPUTASI*, vol. 20, no. 2, pp. 189–198, 2021.

[3] W. Nugraha and A. Sasongko, "Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search Hyperparameter Tuning on Classification Algorithm with Grid Search," *Sist. J. Sist. Inf.*, vol. 11, no. 2, pp. 391–401, 2022, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[4] N. Made, O. Wulaning, H. Elindra, and A. H. Saputra, "Prediksi Karbon Monoksida Menggunakan Model Machine Learning Berdasarkan Perbandingan Model Time Series Studi Kasus DKI Jakarta Carbon Monoxide Prediction Using Machine Learning Model Based on Time Series Model Comparison DKI Jakarta Case Study," vol. 7, no. 3, pp. 1116–1128, 2024, doi: 10.56338/jks.v7i3.4819.

[5] F. Hidiyanto, K. Fajar, A. Sukra, R. Fajar, N. S. Octaviani, and D. A. Sugeng, "Modeling Indonesian Motor Vehicle Tax Coefficients Based on Machine Learning Emission Data," *J. Ind. Res. Innov.*, vol. 17, no. 1, pp. 16–23, 2023, [Online]. Available: https://ejournal.brin.go.id/MIPI/article/view/1734

[6] S. R. Pandala, "Welcome to Lazy Predict's documentation! — Lazy Predict 0.2.12 documentation," 2022. https://lazypredict.readthedocs.io/en/latest/ (accessed Nov. 25, 2024).

[7] Vapnik and V. N., "The Nature of Statistical Learning," *Theory*. p. 334, 1995.

[8] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[9] M. M. Muttaqin, Wahyu Wijaya Widiyanto, A. W. Green Ferry Mandias, Stenly Richard Pungus, S. A. H. Wiranti Kusuma Hapsari, E. F. B. Aslam Fatkhudin, Pasnur, and N. S. Mochammad Anshori, Suryani, *Pengenalan Data Mining*, no. July. 2023.

[10] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[11] M. A. Muslim, Y. Dasril, M. Sam'an, and Y. N. Ifriza, "An improved light gradient boosting machine algorithm based on swarm algorithms for predicting loan default of peer-to-peer lending," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 1002–1011, 2022, doi: 10.11591/ijeecs.v28.i2.pp1002-1011.

[12] A. J. Smola and B. Sch, "Smola, Schölkopf - 2004 - Statistics and Computing - A tutorial on support vector regression.pdf," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[13] G. Louppe, "Understanding Random Forests: From Theory to Practice," no. July, 2014, [Online]. Available: http://arxiv.org/abs/1407.7502

[14] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[15] O. Chung, *Second edition*, vol. 69, no. 4. 2019.

[16] A. R. Chrismanto, A. K. Sari, and Y. Suyanto, "Enhancing Spam Comment Detection on Social Media With Emoji Feature and Post-Comment Pairs Approach Using Ensemble Methods of Machine Learning," *IEEE Access*, vol. 11, no. June, pp. 80246–80265, 2023, doi: 10.1109/ACCESS.2023.3299853.