Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2025129419
Vol. 12, No. 2, April 2025, hlm. 351-358 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENINGKATAN AKURASI PREDIKSI HARGA BARANG IMPOR MENGGUNAKAN XGBOOST DAN *PARTICLE SWARM OPTIMIZATION*

Asmuni Haris[*1], Mahrus Sholeh[2], Lailil Muflikhah[3], Novanto Yudistira[4]

[1,2,3,4]Universitas Brawijaya, Malang
Email: [1]asmuniharis@student.ub.ac.id, [2]mahrussholeh@student.ub.ac.id, [3]lailil@ub.ac.id, [4]yudistira@ub.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 8 November 2024, diterima untuk diterbitkan: 14 April 2025)

Abstrak

Impor di Indonesia dilakukan untuk memenuhi kebutuhan dalam negeri dan memastikan kelancaran produksi serta distribusi. Namun sering terjadi *under invoicing*, yaitu harga barang yang diimpor dilaporkan lebih rendah dari nilai sebenarnya, yang mengakibatkan kerugian penerimaan negara. Penelitian ini bertujuan untuk memprediksi harga barang impor yang sebenarnya guna mengurangi kerugian tersebut. Data yang digunakan diperoleh dari dataset barang impor yang tersedia di *platform Kaggle*, yang disediakan oleh *Data Analytics Community (Mof-DAC)* dari Kementerian Keuangan Indonesia. Metode yang diusulkan meliputi beberapa langkah, dimulai dengan ekstraksi fitur menggunakan *Large Language Model (LLM)* dan *Regular Expression (Regex)*, diikuti oleh optimasi *hyperparameter* XGBoost menggunakan *Particle Swarm Optimization (PSO).* Hasil penelitian menunjukkan bahwa model dengan ekstraksi fitur menggunakan metode Regex mengungguli LLM berdasarkan nilai *Mean Squared Error (MSE), Root Mean Squared Error (RMSE),* dan *Mean Absolute Percentage Error (MAPE).* Kombinasi ekstraksi fitur menggunakan Regex dan TFIDF memberikan hasil yang optimal dalam hal waktu pemrosesan dan akurasi prediksi. *Hyperparameter* terbaik untuk XGBoost ditemukan dengan *max-depth* 51,49, *subsample* 0,89, dan *min_child_weight* 0,65, yang meningkatkan akurasi MAPE menjadi 14,6%. Meskipun model *Random Forest* memiliki akurasi prediksi sedikit lebih baik dengan MAPE sebesar 12,8%, namun waktu pemrosesannya sangat lama sekitar 3 jam membuatnya kurang efisien. Sebaliknya, XGBoost dengan waktu pemrosesan hanya 51,49 detik dan MAPE 14,6% dipilih sebagai model terbaik karena akurasi yang cukup baik dengan waktu komputasi yang cepat.

Kata kunci: *Impor, Optimisasi Particle Swarm, Pembelajaran Mesin, Prediksi, XGBoost*

IMPROVING PREDICTION ACCURACY OF IMPORTED GOODS PRICES USING XGBOOST AND PARTICLE SWARM OPTIMIZATION

Abstract

Imports in Indonesia fulfill domestic needs and sustain manufacturing and distribution. Under invoicing, where imported products are purposely underpriced, reduces state revenue. This study predicts imported goods prices to reduce financial losses. The Data Analytics Community (Mof-DAC) of the Indonesian Ministry of Finance provided the Kaggle imported products dataset. The Large Language Model (LLM) and Regular Expression are used to extract features in the suggested method. XGBoost hyperparameters are then optimized using Particle Swarm Optimization. Research shows that the Regex-extracted feature model outperforms the LLM model in MSE, RMSE, and MAPE. Regex feature extraction and TFIDF produce the best processing time and prediction accuracy. The ideal XGBoost hyperparameters were a maximum depth of 51.49, a subsample value of 0.89, and a minimum child weight of 0.65. These hyperparameters increased MAPE accuracy to 14.6%. The Random Forest model has a Better Prediction Accuracy (MAPE) of 12.8%, but its processing time is 3 hours, lowering its efficiency. XGBoost was chosen as the best model due to its 51.49-second processing time and 14.6% MAPE. High accuracy and efficient computing make this model effective.

Keywords: *Import, Machine Learning, Particle Swarm Optimization, Prediction, XGBoost*

1. PENDAHULUAN

Di Indonesia, impor digunakan untuk memenuhi kebutuhan dalam negeri dan memastikan produksi dan distribusi dalam negeri tidak terganggu. Hal ini terutama diperlukan karena ketergantungan pada komoditas atau bahan mentah tertentu yang tidak dapat diproduksi di dalam negeri atau tidak cukup untuk memenuhi permintaan pasar dalam negeri (Hanifah, 2022).

Setiap barang yang diimpor akan dilakukan pemeriksaan untuk mengetahui nilai pabeannya yang akan menjadi dasar penghitungan bea masuk dan pajak dalam rangka impor. Namun demikian, harga barang yang diberikan importir tidak sesuai dengan nilai sebenarnya, hal ini biasa disebut dengan *under invoicing*. Dalam dunia perdagangan global, kesalahan faktur perdagangan merupakan salah satu komponen penghindaran pajak. Pada tahun 2016, perkiraan kerugian penerimaan negara dari bea dan cukai adalah sekitar USD 302 juta, dari total perkiraan kerugian sebesar USD 6,5 miliar (Heydt, 2019).

Selain itu, kesalahan faktur perdagangan, biasanya disebut sebagai *under invoicing* atau *under valuation*, adalah teknik penipuan yang digunakan oleh importir selama pemeriksaan pabean untuk mengubah nilai yang dilaporkan dari produk impor, dengan tujuan menghindari pajak dan denda. Praktik kriminal ini mengacu pada tindakan yang sengaja memberikan gambaran yang salah tentang nilai atau kuantitas produk, yang mungkin mencakup impor yang meningkat dan ekspor yang dinilai terlalu rendah (Lai and Hou, 2023).

Importir berupaya menurunkan pajak dan tarif yang harus mereka bayarkan di perbatasan dengan mendistorsi nilai komoditas, yang pada akhirnya mengurangi pendapatan pemerintah (Thiao, 2021). Perilaku ini berdampak negatif pada kesejahteraan finansial suatu negara dan juga merusak data perdagangan, sehingga menghasilkan evaluasi yang salah terhadap perekonomian dan pilihan kebijakan (Asmah, Andoh and Titriku, 2020). Kesalahan faktur perdagangan menurunkan pendapatan pajak dan upaya untuk mendorong transparansi dan tata kelola yang baik dalam transaksi perdagangan dengan menimbulkan inkonsistensi antara nilai ekspor dan impor yang dilaporkan (Asmah, Andoh and Titriku, 2020).

Pada tahun 2021, Kementerian Keuangan RI mengadakan hackathon data melalui Ministry of Finance Data Anlytics Community (MofDac) untuk mencari model machine learning terbaik dalam memprediksi harga impor serta pada tahun 2023, Direktorat Jenderal Bea Cukai juga menyelenggarakan data hackathon. persaingan proyek analitik dan salah satu temanya adalah deteksi penipuan pada impor berdasarkan faktur.

Untuk memprediksi harga impor secara efisien, kami memilih model model *machine learning eXtreme Gradient Boosting (XGBoost)* sering digunakan untuk prediksi di platform Kaggle karena akurasinya yang tinggi. Namun, meskipun kami menggunakan XGBoost, hasilnya masih kurang dibandingkan dengan *random forest*. Oleh karena itu, penulis bertujuan untuk melakukan *tuning hyperparameter* pada algoritma XGBoost.

Particle Swarm Optimization (PSO) dipilih sebagai metode untuk proses *tuning* karena diakui sebagai teknik optimasi yang sangat efektif di beberapa domain, seperti *machine learning* dan biologi komputasi (Qin et al., 2021). PSO dapat dimanfaatkan untuk mengoptimalkan *hyperparameter* dalam ruang pencarian berkelanjutan ketika diterapkan pada algoritma seperti XGBoost (Qin et al., 2021). Kapasitas PSO untuk menangani nilai riil berkelanjutan sangat cocok untuk mengoptimalkan XGBoost untuk pekerjaan seperti penilaian kredit (Qin et al., 2021).

Berdasarkan uraian di atas, maka penelitian ini memiliki rumusan masalah yaitu apakah model *machine learning* XGBoost dan PSO dapat meningkatkan akurasi prediksi harga barang impor sebenarnya yang selanjutnya diharapkan akan memberikan kontribusi sebagai berikut:

1) Penerapan teknik *feature engineering* yang efektif untuk variabel DESKRIPSI_BARANG yaitu uraian tentang barang yang diimpor pada data yang digunakan.

2) Mengoptimalkan *hyperparameter* XGBoost menggunakan PSO untuk meningkatkan akurasi prediksi harga barang impor di Indonesia.

2. LANDASAN PUSTAKA

2.1 XGBoost

EXtreme Gradient Boosting (XGBoost) adalah model *machine learning* yang sangat efisien dan akurat yang dirancang khusus untuk menangani data terstruktur. Pendekatannya menggunakan *ensemble learning*, dimana rangkaian model prediksi yang lemah, biasanya pohon keputusan, dibangun secara berurutan untuk membentuk model prediksi yang kuat (Li, Zhang and Wang, 2023). Proses pelatihan setiap pohon baru dalam kelompok yang berupaya untuk memperbaiki kesalahan apa pun yang disebabkan oleh kelompok saat ini, sehingga menghasilkan peningkatan akurasi prediksi (Li, Zhang and Wang, 2023).

Dataset D didefinisikan sebagai himpunan pasangan (x_i, y_i). Dataset tersebut berisi N sampel, dimana setiap sampel terdiri dari xi ∈ RM dan y_i∈R. Setiap sampel masukan terdiri dari fitur berdimensi M dan label berdimensi satu. Subpohon pada algoritma CART dibentuk dan diberi nomor dengan K. Selanjutnya hasil prediksi setelah integrasi seluruh subpohon pada XGBoost dapat direpresentasikan sebagai (Fang et al., 2022):

$$\widehat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k(\cdot) \in F \qquad (1)$$

Fungsi pemetaan subpohon k dilambangkan dengan $f_k(\cdot)$, F mewakili himpunan fungsi pemetaan semua subpohon, dan $\phi(x_i)$ mewakili fungsi pemetaan subpohon model setelah mengintegrasikan semua subpohon. XGBoost tidak menghasilkan semua subpohon secara bersamaan, melainkan menambahkan satu subpohon di setiap putaran berdasarkan algoritma serakah yang menggunakan metode peningkatan gradien. Proses pemasangan dilakukan secara bertahap. Perkiraan hasil untuk putaran t adalah:

$$\widehat{y_i^{(t)}} = \sum_{k=1}^{t} f_k(x) = \widehat{y_i^{(t-1)}} + f_t(x_i) \quad (2)$$

2.2 *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization (PSO) adalah teknik pengoptimalan yang terinspirasi dari alam dan digunakan secara luas di banyak domain, termasuk *machine learning* dan biologi komputasi. PSO adalah teknik komputasi yang meniru perilaku sosial yang diamati dalam kawanan burung atau kawanan ikan. Ini secara berulang meningkatkan solusi dengan memodifikasi partikel sesuai dengan lokasi pribadinya yang paling terkenal dan posisi paling terkenal secara keseluruhan dalam ruang pencarian (c).

Arsitektur PSO terdiri dari populasi partikel yang bergerak melalui ruang pencarian untuk menemukan solusi optimal. Setiap partikel menyesuaikan posisinya berdasarkan pengalamannya sendiri dan pengalaman kelompok, dipandu oleh dua komponen utama: komponen kognitif, yang mewakili memori partikel tentang posisi terbaiknya, dan komponen sosial, yang mewakili pengaruh posisi terbaik kelompok tersebut (Qin et al., 2021).

Struktur algoritma PSO terdiri dari sekumpulan partikel yang menjelajahi ruang pencarian untuk menemukan solusi yang paling optimal. Setiap partikel memodifikasi posisinya dengan mempertimbangkan pengalaman masa lalunya dan pengalaman kolektif kawanannya. Proses ini dikendalikan oleh dua faktor kunci: komponen kognitif, yang mewakili ingatan partikel akan posisi terbaiknya, dan komponen sosial, yang mewakili dampak dari posisi terbaik kelompok tersebut (Qin et al., 2021). Selama setiap iterasi algoritma PSO, setiap partikel menyimpan catatan posisi terbaik sebelumnya *(pbest)* dan memiliki akses ke posisi terbaik yang tercatat secara global *(gbest).* Akibatnya, setiap partikel memodifikasi posisi dan kecepatannya dengan memanfaatkan Persamaan (3) dan (4) dengan tujuan mencapai solusi terbesar yang mungkin ada dalam kelompok.

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (3)$$

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * (p_{id} - x_{id}^{t}) + c_2 * r_2 * (p_{gd} - x_{id}^{t}) \ (2) \quad (4)$$

Dimana x_{id}^{t} dan v_{id}^{t+1} masing-masing menyatakan posisi dan kecepatan partikel i pada iterasi sedangkan d termasuk dalam himpunan D = {1, 2, 3, ...}. Dimensi ruang pencarian dilambangkan dengan D. Personal best dan global best masing-masing diwakili oleh variabel pid dan pgd, serta c_1 dan c_2 adalah konstanta positif yang digunakan untuk mewakili kecepatan pembelajaran, yang biasanya diberi nilai 2,0. Istilah percepatan stokastik menggambarkan bobot yang menarik setiap partikel menuju tempat terbaik pribadinya (pbest) dan terbaik global (gbest). Simbol 'w' menunjukkan bobot inersia, sedangkan . r_1 dan r_2 ditetapkan secara acak bilangan bulat nyata dalam rentang (Heydt, 2019; Hanifah, 2022).

2.3 Pemeriksaan Barang Impor di Indonesia

Barang yang diimpor ke Indonesia akan menjalani pemeriksaan oleh petugas Bea Cukai yang berada di bawah Kementerian Keuangan. Pemeriksaan ini akan dilakukan baik di pelabuhan maupun bandara internasional. Sesuai Peraturan Menteri Keuangan Nomor 185 Tahun 2022, proses pemeriksaan meliputi pemeriksaan fisik dan penilaian nilai produk impor. Tujuan pemeriksaan pabean adalah untuk memperoleh informasi yang tepat dan mengevaluasi Pemberitahuan Pabean Impor atau Dokumen Pelengkap yang telah diserahkan.

Penilaian terhadap harga atau nilai barang impor akan menjadi dasar penetapan nilai pabean, yang kemudian digunakan untuk menghitung bea masuk dan pajak atas impor. Terkadang, nilai barang yang diimpor tidak sesuai dengan nilai sebenarnya, yang disebut *under invoicing*. Fenomena ini dikenal dengan *istilah trade mis-invoicing*, yaitu strategi yang digunakan untuk menghindari pajak dalam bidang perdagangan internasional. Indonesia diperkirakan mengalami kerugian sebesar USD 302 juta dalam pendapatan bea cukai dan pajak. Kerugian ini merupakan bagian dari potensi kerugian yang lebih besar yaitu sebesar USD 6,5 miliar pada tahun 2016 (Global Financial Integrity, 2019).

Pejabat kantor bea cukai mempunyai kewenangan untuk menganalisis nilai pabean kiriman dengan menggunakan metode resmi, seperti membandingkan data atau berkonsultasi dengan sumber lain. Mereka juga dapat menyelidiki kasus-kasus di mana terdapat kecurigaan bahwa nilai yang dinyatakan pada nota konsinyasi lebih rendah dari yang seharusnya. Apabila penerima barang tidak puas dengan penyelesaian yang dicapai oleh kantor pabean, ia mempunyai pilihan untuk mengajukan permohonan pembetulan atau keberatan sesuai dengan norma yang mengatur keberatan di bidang kepabeanan.

2.4 Evaluasi

Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), dan Root Mean Squared

Error (RMSE) adalah metrik evaluasi yang banyak digunakan di berbagai domain termasuk penilaian kredit perangkat lunak (Qin et al., 2021), perkiraan kecepatan angin (Shi, 2024), prediksi konsumsi bahan bakar (Su et al., 2023), dan prediksi kegagalan jaringan distribusi tenaga listrik (Fang et al., 2022).

Indikator-indikator ini sangat penting untuk mengevaluasi efektivitas model dan algoritma prediktif. Para ilmuwan telah menggunakan metode pembelajaran mesin seperti XGBoost dan PSO untuk meningkatkan ketepatan perkiraan di berbagai bidang (Gu, Zhang and Bao, 2021; Mai et al., 2021; Nan, 2023)

3. METODE PENELITIAN

3.1. Pengumpulan Data

Data diperoleh dengan mengakses dataset barang impor yang tersedia di platform Kaggle. Dataset ini disediakan oleh *Data Analytics Community (Mof-DAC)*, yang merupakan komunitas data analitik resmi di Kementerian Keuangan Indonesia.

URL untuk mengakses kumpulan data disediakan di https://www.kaggle.com/competitions/nilai-impor/data.

3.1.1 Karakteristik Data

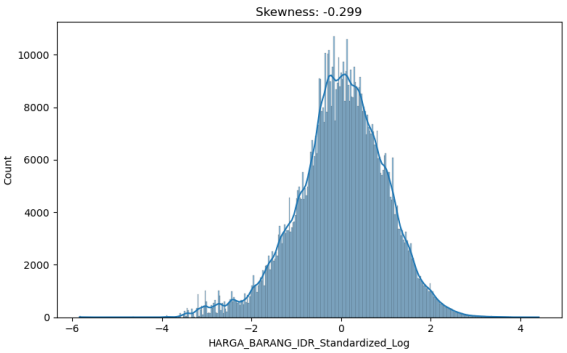
Dataset terdiri dari 9 variabel. Semua variabel ditampilkan pada Tabel 1.

Tabel 1. Variabel Pada Dataset Yang Digunakan

No	Variabel (Tipe Data)	Deskripsi	Karakteristik
1	KODE_BARANG (String)	Kode klasifikasi barang berdasarkan sistem *HS Code*	Variabel ini digunakan sebagai identifikasi unik bagi setiap jenis barang yang diimpor.
2	DESKRIPSI_BARANG (Long Text)	Uraian tentang barang yang diimpor	Variabel ini digunakan untuk memberikan informasi tambahan yang tidak cukup jelas dari kode barang.
3	ASAL_BARANG (String)	Negara asal barang impor	Variabel ini digunakan untuk mengatahui negara asal barang impor
4	JUMLAH_KEMASAN (Numerik)	Jumlah kemasan barang	Variabel ini penting untuk menentukan skala transaksi atau volume fisik dari barang yang diimpor.
5	JENIS_KEMASAN (String)	Jenis kemasan barang	Variabel ini mempengaruhi biaya pengiriman, penanganan, dan penyimpanan barang
6	HARGA_BARANG_IDR (Numerik)	Total harga barang dalam rupiah	Variabel ini berisi nilai finansial barang yang diimpor.
7	KODE_SATUAN_BARANG (String)	Jenis satuan barang	Variabel ini menentukan satuan barang yang digunakan untuk mengukur jumlah barang.
8	JUMLAH_BARANG (Numerik)	Jumlah barang per kemasan	Variabel ini menunjukkan jumlah unit barang di dalam setiap kemasan.
9	MENGGUNAKAN_FASILITAS (Boolean)	Apakah dalam impor tersebut ada fasilitas pembebasan pajak impor	Variabel biner ini penting untuk menganalisis apakah barang impor tersebut mendapat fasilitas fiskal atau insentif, seperti pembebasan pajak.

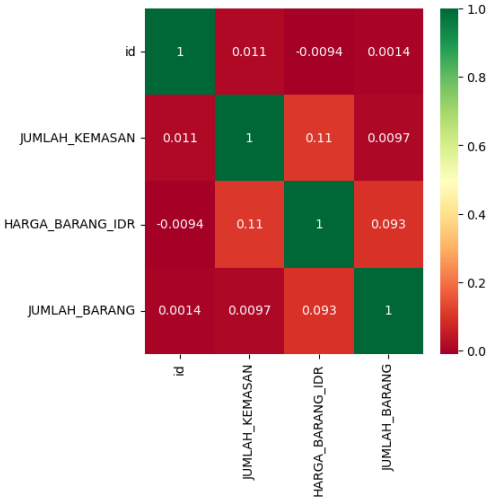
3.1.2 Eksplorasi Data

Dataset terdiri dari 797.269 baris data. Distribusi sebaran harga barang diperlihatkan dalam Gambar 1.



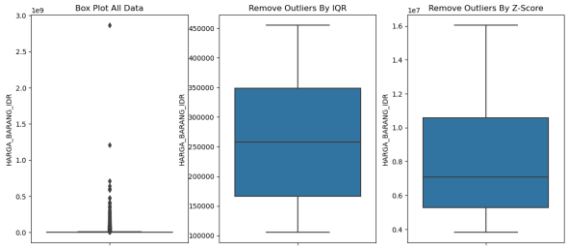
Gambar 1. Distribusi Harga Barang Impor

Berdasarkan data yang tersaji pada Gambar 1, sebaran HARGA PRODUK menunjukkan nilai *skewness* yang sedikit negatif yaitu -0,299. Adapun korelasi antar variabel bertipe numerik ditunjukkan pada Gambar 2.



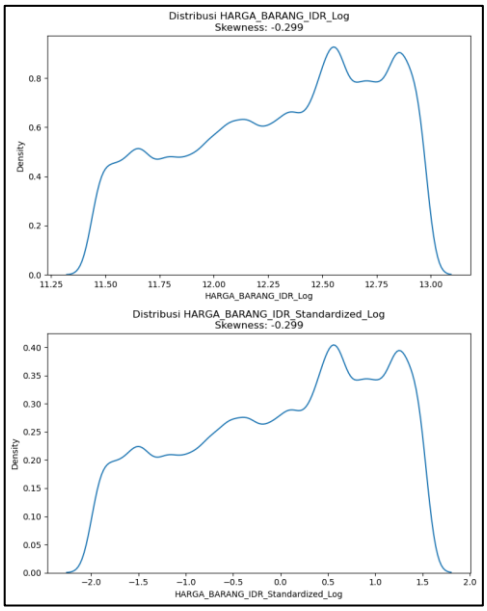
Gambar 2. Korelasi Antar Variabel Numerik

Dari Gambar 2 dapat dilihat kurangnya korelasi antara variabel dependen yang menjadi target prediksi yaitu HARGA_BARANG_IDR dengan variabel numerik lainnya.



Gambar 3. Box-Plot Distribusi Harga Barang

Dari sebaran harga barang pada boxplot di Gambar 3, terlihat banyak *outlier*. Untuk menghilangkan pencilan ini kami mencoba menggunakan perhitungan *Inter Quartile Range (IQR)* dan metode *Z Score* dan membandingkan hasil keduanya seperti yang ditunjukkan pada Gambar.3, lalu memutuskan menggunakan IQR karena hasilnya yang lebih baik untuk proses selanjutnya.



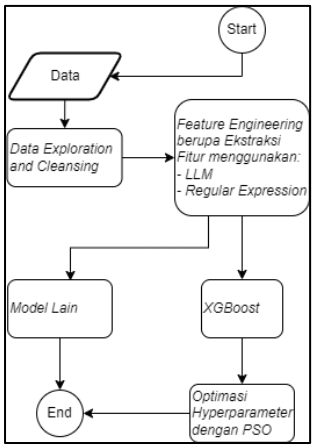
Gambar 4. Distribusi Harga Barang
Setelah Normalisasi dan Standardisasi

Selanjutnya kami melakukan normalisasi dan stadardisasi menggunakan fungsi log dengan hasil seperti yang di tunjukkan pada Gambar 4.

3.2. Alur Proses Eksperimen

Kami menggunakan beberapa langkah untuk memprediksi harga barang impor sebenarnya yang secara umum diperlihatkan pada Gambar 5.

Proses ini dimulai dengan langkah *Data Exploration and Cleansing*, yang mencakup eksplorasi awal data serta pembersihan data dari anomali atau *noise* ysng tidak relevan untuk analisis lebih lanjut. Selanjutnya dilakukan *feature engineering* berupa ekstaksi fitur pada variabel DESKRIPSI_BARANG untuk mendapatkan informasi detil barang menggunakan *Large Language Model* (LLM) dan *Regular Expression (Regex)* kemudian membandinganken hasil keduanya untuk dipakai pada proses training dan testing dengan XGBoost. Setelah mendapatkan rentang hyperparameter yang tidak terlalu jauh, lalu dilakukan optimasi hyperparameter XGBoost menggunakan PSO serta model algoritma lain sebagai pembanding.



Gambar 5. Alur Proses Eksperimen

3.2.1. Ekstraksi Fitur Dengan *Large Language Model*

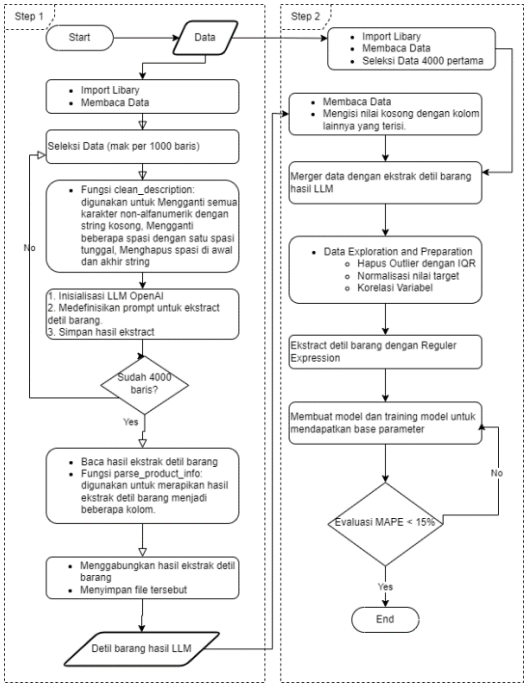
Proses menggunakan LLM OpenAI dilakukan untuk melakukan ekstraksi pada kolom DESKRIPSI_BARANG dari 4000 data pertama untuk mendapatkan informasi detil barang berupa jenis, model, dan merek barang. Terbatasnya data yang digunakan karena keterbatasan dana penelitian mengingat penggunaan LLM ini bersifat berbayar. Untuk memulai langkah 1, LLM Open AI perlu dimulai dengan menyediakan *api_key*. Setelah *api_key* dimasukkan, kemudian menentukan prompt untuk pemrosesan LLM. Data yang diperoleh dari LLM akan dimasukkan ke dalam data dan kemudian disusun dalam kolom jenis, model, dan merek barang.

Kolom yang diekstraksi akan digunakan pada langkah kedua untuk mengevaluasi hasil dalam model XGBoost dan membandingkannya dengan menggunakan pendekatan berbasis Regex untuk mengekstraksi detil barang tertentu seperti yang ditunjukkan pada Gambar 6.

Proses ekstraksi fitur pada DESKRIPSI _BARANG_IDR menggunakan *Regular Expression (Regex)* untuk nanti dibandingkan hasilnya dengan hasil ekstraksi fitur menggunakan LLM. Formula Regex juga akan menghasilkan kolom jenis, model, dan merek barang yang diekstraksi dari kolom DESKRIPSI_BARANG. Formula tersebut berisi daftar jenis, model, dan merek sebagai referensi pencocokan. Jika ekspresi reguler tidak menemukan kecocokan di satu kolom, maka formula akan mengambil nilai dari kolom lainnya yang sudah terisi.

Setelah itu, dilakukan *training model* menggunakan tambahan kolom informasi detil barang dari LLM dan Regex secara terpisah untuk dilihat nilai MAPE nya yang paling baik. *Training model* ini menggunakan XGBoost standar tanpa penyesuaian apa pun pada *hyperparameter*. Kolom informasi detil barang yang menghasilkan nilai MAPE terbaik akan di gunakakan pada proses selanjutnya yakni tuning *hyperparameter*.

3.2.2. Ekstraksi Fitur Dengan *Regular Expression*



Gambar 6. Prosedur Mencoba Beberapa Teknik Feature Engineering dan Membandingkan Hasilnya

3.2.3. Mencari *Hyperparameter* XGBoost menggunakan PSO

Tahap terakhir menggunakan fungsi PSO untuk mengoptimalkan *hyperparameter* model XGBoost seperti yang ditunjukkan pada Gambar 7. Fungsi PSO dirancang untuk memperbarui posisi dan kecepatan partikel melalui beberapa iterasi, sedangkan fungsi evaluasi_model menilai kinerja model yang menggunakan *hyperparameter* tertentu.

4.1. HASIL DAN PEMBAHASAN

4.1. Konfigurasi Eksperimen

Dengan bahasa Python, kami menggunakan modul XGBoost untuk *training* dan *testing* model. Pada proses pencarian *hyperparameter*, kami membuat algoritma PSO sendiri tanpa melibatkan *library* eksternal. Pendekatan *pipeline* dari *library* Scikit-learn digunakan untuk *training* dan *testing* model. Hal ini memungkinkan integrasi dengan efisien pada proses *data preparation*, seperti normalisasi, standardisasi pada beberapa kolom ke dalam satu alur *(pipe)* bersama dengan model.

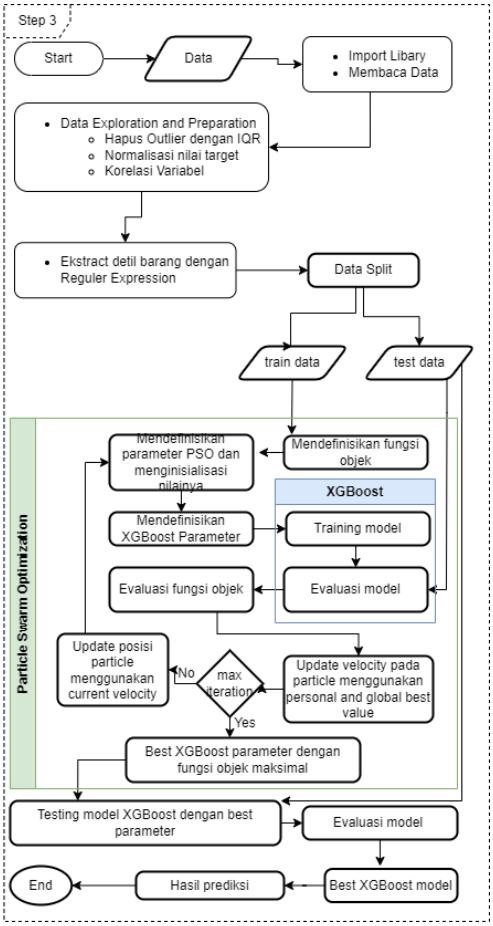
4.2. Hasil Evaluasi Penggunaan Ekstraksi Fitur

Tabel 2 Variabel Pada Dataset Yang Digunakan

Fitur	Waktu (detik)	MSE	RMSE	MAPE*
Ekstraksi DESKRIPSI_BARANG menggunakan LLM	133,00	0,97	0,98	36,7
Ekstraksi DESKRIPSI_BARANG menggunakan Regex	121	**0,93**	**0,96**	**35,1**
Tidak menggunakan DESKRIPSI_BARANG	**106**	1,12	1,06	37,5
Tidak menggunakan seluruh kolom bertipe numerik	109	0,99	0,99	38,3

Hasil ekstraksi fitur yang dilakukan kemudian dibandingkan dengan berbagai metode pemilihan fitur menggunakan model standar XGBoost ditunjukkan pada Tabel 2.

Tabel 2 menunjukkan bahwa model yang berisi rekayasa fitur melalui ekstraksi teks menggunakan metode regex mengungguli hasil yang diperoleh dengan LLM berdasarkan nilai MSE, RMSE, dan MAPE. Nilai yang lebih rendah dari ekstraksi fitur menggunakan LLM ini mungkin timbul karena penggunaan 4 ribu data pertama saja, yang jauh lebih rendah dibandingkan keseluruhan kumpulan data yang berjumlah lebih dari 700 ribu data. Sedangkan performa model saat memanfaatkan seluruh data yang tersedia ditunjukkan pada Tabel 3.



Gambar 7. Penggunaan XGBoost dan PSO

Tabel 3 Performa XGBoost Standar

Fitur	Waktu (detik)	MSE	RMSE	MAPE*
Ekstraksi DESKRIPSI_BARANG menggunakan Regex	2,11	0,74	0,86	32,50
Tidak menggunakan DESKRIPSI_BARANG	1,43	0,80	0,89	34,20
Tidak menggunakan seluruh kolom bertipe numerik	1,49	0,83	0,91	35,50
Ekstraksi DESKRIPSI_BARANG menggunakan Regex dan TFIDF	**11,00**	**0,61**	**0,78**	**28,50**

Tabel 3 menunjukkan bahwa kombinasi ekstraksi fitur menggunakan Regex dan TFIDF memberikan nilai waktu, MSE, RMSE, dan MAPE yang paling optimal. Hal ini menunjukkan bahwa XGBoost mampu menangani semua jenis data secara efektif, bahkan ketika data tersebut tidak memiliki distribusi yang lebih baik dibandingkan pengujian sebelumnya yang menghilangkan data anomali.

Selanjutnya nilai *hyperparameter* optimal untuk fitur dalam model ini akan ditentukan menggunakan PSO. Kami mengatur nilai rentang *hyperparameter* pada XGBoost seperti pada Tabel 4.

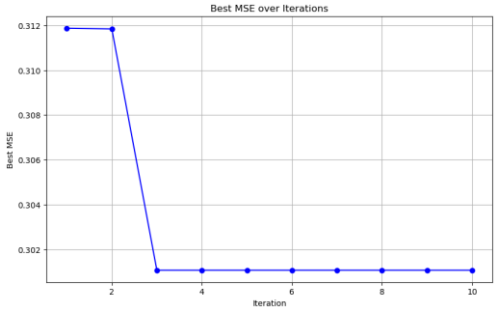
Tabel 4. Rentang Nilai *Hyperparameter* XGBoost

Hyperparameter	***Deskripsi***	***Range***
Maximum depth	Maksimum kedalaman *tree* pada model	49 – 51
Subsample	proporsi data pelatihan yang digunakan untuk membangun setiap *tree* dalam model	0.7 – 1
Minimum child weight	Minimum bobot yang dibuuthkan dalam setiap *node*	1 – 3

Kami mengatur PSO dengan jumlah iterasi maksimum 10 *(n_iterations)*, jumlah partikel 5 (n_particles). Sedangkan untuk rentang nilai *hyperparameter* XGBoost kami tentukan berdasarkan acuan pada web dokumentasi XGBoost dan berdasarkan beberapa kali percobaan training hingga menemukan nilai evaluasi yang cukup tinggi. Adapun rentang nilai *hyperparameter* XGBoost adalah nilai batas kedalaman *tree* maksimum (*maximum depth*) 49 sampai 51, rentang subsampel *(subsample)* 0,7 hingga 1, dan bobot anak minimum *(Minimum child weight*) 1 sampai 3.

Hasil iterasi ditampilkan pada Gambar 7 yang menggambarkan nilai *hyperparameter* optimal yang dicapai mulai iterasi ketiga, dengan nilai MSE sebesar 0,31. Hasil detilnya ditampilkan pada Tabel 5.

Tabel 5 menunjukkan hyperparamter terbaik XGBoost yaitu *Max-depth* adalah 6, *subsample* adalah 1, dan *Min_child_weight* adalah 1. Perubahan ini menghasilkan peningkatan akurasi MAPE menjadi 14,6%. Setelah itu, kami mengevaluasi kinerjanya dengan membandingkannya dengan algoritma lain yang ditunjukkan pada Tabel 6.



Gambar 7. Nilai MSE

Tabel 5. Performa Kombinasi XGBoost dan PSO

Hyperparameter	***Standard XGBoost***	**Kombinasi PSO: *Max iteration* = 10 *Particle* = 5**
Max-depth	6,00	51,49
subsample	1,00	0,89
Min_child_weight	1,00	0.65
Waktu	11,00	105,00
MSE Normalized	0,61	0,31
RMSE Normalized	0,78	0,55
MAPE Normalized	201,00	199,00
MSE (Miliar)	6,10	3,00
RMSE	78673,00	55415,00
MAPE	28,50	**14,60**

Tabel 6 Hasil Dengan Model/Algoritma Lain

Perbandingan Dengan Algoritma Lain

Algoritma	Waktu (detik)	MSE	RMSE	MAPE*
XGBoost + PSO	105,00	0,31	0,78	14,60
Regresi Linier	*107,00*	1,01	1.00	66,60
Decision Tree	*87,00*	0,44	0,66	15,00
Random Forest	***10.560,00***	0,26	0,51	**12,80**

Tabel VI menunjukkan bahwa pendekatan *Random Forest* memiliki akurasi prediksi terbaik dengan MAPE sebesar 12,8%. Namun perlu dicatat bahwa waktu pemrosesannya lebih dari 10.560 detik atau sekitar 3 jam. Karena pentingnya efisiensi waktu pemrosesan, peneliti lebih memilih XGBoost yang memiliki waktu pelatihan tercepat dengan waktu 51,49 detik, dengan MAPE 14,6%. Nilai MAPE tersebut hanya sekitar 1,8% lebih rendah dibandingkan temuan yang diperoleh dari model *Random Forest.*

4. KESIMPULAN

Penelitian ini bertujuan untuk memprediksi harga barang impor sebenarnya di Indonesia guna mengurangi kerugian negara akibat *under invoicing*. Melalui *feature engineering* berupa ekstraksi fitur menggunakan LLM dan Regex, serta pengoptimalan

hyperparameter XGBoost dengan PSO, penelitian ini mengidentifikasi metode terbaik untuk memprediksi nilai pabean barang impor.

Hasil penelitian menunjukkan bahwa metode ekstraksi teks menggunakan Regex menghasilkan performa yang lebih baik dibandingkan dengan menggunakan LLM. Kemudian jika dilakukan kombinasi ekstraksi fitur menggunakan Regex dan TFIDF, maka akan memberikan hasil waktu pemrosesan dan akurasi prediksi yang lebih optimal. *Hyperparameter* terbaik untuk XGBoost ditemukan dengan nilai *Max-depth* 51,49, *subsample* 0,89, dan *Min_child_weight* 0,65, yang meningkatkan akurasi prediksi menjadi 14,6%.

Meskipun algoritma *Random Forest* memiliki akurasi prediksi terbaik dengan MAPE sebesar 12,8%, waktu pemrosesannya yang sangat lama, sekitar 3 jam, membuatnya kurang efisien. Sebaliknya, XGBoost dengan waktu pemrosesan hanya 51,49 detik dan MAPE 14,6% dipilih sebagai algoritma terbaik karena keseimbangan antara akurasi dan efisiensi waktu.

Penelitian ini menyarankan penggunaan XGBoost dengan kombinasi ekstraksi fitur Regex dan TFIDF sebagai metode yang efisien dan akurat untuk memprediksi nilai pabean barang impor di Indonesia, guna mengurangi kerugian negara akibat *under invoicing*.

DAFTAR PUSTAKA

ASMAH, E.E., ANDOH, F.K. AND TITRIKU, E., 2020. Trade Misinvoicing Effects on Tax Revenue in Sub-Saharan Africa: The Role of Tax Holidays and Regulatory Quality. *Annals of Public and Cooperative Economics*, 91(4), pp.649–672.

FANG, J., WANG, H., YANG, F., YIN, K., LIN, X. AND ZHANG, M., 2022. A failure prediction method of power distribution network based on PSO and XGBoost. *Australian Journal of Electrical and Electronics Engineering*, [online] 19(4), pp.371–378.

GU, Y., ZHANG, D. AND BAO, Z., 2021. A New Data-Driven Predictor, PSO-XGBoost, Used for Permeability of Tight Sandstone Reservoirs: A Case Study of Member of Chang 4+5, Western Jiyuan Oilfield, Ordos Basin. *Journal of Petroleum Science and Engineering*, 199, p.108350.

HANIFAH, U., 2022. Pengaruh Ekspor Dan Impor Terhadap Pertumbuhan Ekonomi Di Indonesia. *Transekonomika: Akuntansi, Bisnis Dan Keuangan*, 2(6), pp.107–126.

HEYDT, M., 2019. *GFI: Indonesia lost estimated US$6.5 billion to trade misinvoicing in 2016.* [online] Washington DC. Available at: <https://gfintegrity.org/press-release/gfi-indonesia-lost-estimated-us6-5-billion-to-trade-misinvoicing-in-2016/>.

LAI, M. AND HOU, J., 2023. Let Us Misinvoice More? The Effect Of e Jure capital Controls on Trade Misinvoicing. *World Economy*, 46(7), pp.2157–2186.

LI, J., ZHANG, Z. AND WANG, X., 2023. Performance-Oriented Road Structure and Material Design Method Based on Enhanced XGBoost Algorithm. *International Journal of Pavement Engineering*, 25(1).

MAI, Y., SHENG, Z., SHI, H. AND LIAO, Q., 2021. Using Improved XGBoost Algorithm to Obtain Modified Atmospheric Refractive Index. *International Journal of Antennas and Propagation*, 2021, pp.1–11.

NAN, L., 2023. A Model for Analyzing Employee Turnover in Enterprises Based on Improved XGBoost Algorithm. *International Journal of Advanced Computer Science and Applications*, 14(11).

QIN, C., ZHANG, Y., BAO, F., ZHANG, C., LIU, P. AND LIU, P., 2021. XGBoost optimized by adaptive particle swarm optimization for credit scoring. *Mathematical Problems in Engineering*, 2021.

SHI, Y., 2024. Short-Term Wind Speed Forecasting Based on a Hybrid Model That Integrates PSO-LSSVM and XGBoost. *International Journal of Low-Carbon Technologies*, 19, pp.1138–1143.

SU, M., SU, Z., CAO, S., PARK, K. AND BAE, SI-HWA, 2023. Fuel Consumption Prediction and Optimization Model for Pure Car/Truck Transport Ships. *Journal of Marine Science and Engineering*, 11(6), p.1231.

THIAO, A., 2021. The Effect of Illicit Financial Flows on Government Revenues in the West African Economic and Monetary Union Countries. *Cogent Social Sciences*, 7(1).

Peraturan Menteri Keuangan Republik Indonesia (PMK) Nomor: 112/PMK.04/2018 Tentang Perubahan Atas PMK No. 182/PMK.04/2016 mengenai Ketentuan Impor Barang Kiriman. Jakarta: Kementerian Keuangan