

COMPARISON OF EXTRA TREES REGRESSION AND CART IN MODELLING E-COMMERCE PRODUCT SHIPPING COSTS IN INDONESIA

PERBANDINGAN EXTRA TREES REGRESSION DAN CART DALAM PEMODELAN BIAYA PENGIRIMAN PRODUK E-COMMERCE DI INDONESIA

Dewa Rizky Meidiva Pratama¹, Bilal Muhammad Thalib¹,
Arantxa Aulia¹, Sofia Lorena¹, and Septian Rahardiantoro¹

¹Department of Statistics, IPB University, Indonesia

‡corresponding author: billythalib@apps.ipb.ac.id

Copyright © 2026 Dewa Rizky Meidiva Pratama, Bilal Muhammad Thalib, Arantxa Aulia, Sofia Lorena, and Septian Rahardiantoro. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

This study aims to compare the performance of two tree-based machine learning methods in modeling e-commerce shipping costs in Indonesia, namely Extra Trees Regression and Classification and Regression Tree (CART). The dataset used consisted of 20,848 sales transaction observations obtained from an e-commerce business, of which 19,033 observations were retained for analysis after the preprocessing stage. Predictor variables included total product weight, quantity, destination province, shipping option, product category, payment method, as well as engineered features such as the weight-to-quantity ratio and logarithmic transformation of product weight. Model evaluation was conducted using RMSE and MAPE metrics on a 20% test dataset. The results indicate that both models demonstrated competitive performance with different strengths. CART achieved lower RMSE (5,159.55 vs. 5,256.43) and higher R-Squared (0.6385 vs. 0.6319), while Extra Trees outperformed CART on MAE (3,604.92 vs. 3,677.67) and MAPE (35.69% vs. 36.06%). Based on RMSE as the primary evaluation metric, CART was selected as the best-performing model. Feature importance analysis revealed that destination province and weight-related variables were the most influential predictors of shipping costs. These findings suggest that tree-based methods have strong potential for modeling the nonlinear characteristics of e-commerce shipping costs.

Keywords: CART Regression, E-Commerce Logistics, Extra Trees Regression, Machine Learning, Shipping Cost Prediction

Abstrak

Penelitian ini bertujuan untuk membandingkan performa dua metode *machine learning* berbasis pohon keputusan dalam memodelkan biaya pengiriman e-commerce di Indonesia, yaitu Extra Trees Regression dan Classification and Regression Tree (CART). Dataset yang digunakan terdiri atas 20.848 observasi transaksi penjualan yang diperoleh dari bisnis e-commerce dan setelah proses preprocessing, sebanyak 19.033 observasi digunakan dalam analisis. Variabel prediktor meliputi total berat produk, kuantitas, provinsi tujuan, opsi pengiriman, kategori produk, metode pembayaran, serta fitur hasil rekayasa seperti rasio berat per kuantitas dan transformasi logaritma berat. Evaluasi model dilakukan menggunakan metrik RMSE dan MAPE pada data uji sebesar 20%. Hasil penelitian menunjukkan bahwa kedua model memiliki keunggulan pada metrik yang berbeda. CART mencatatkan RMSE lebih rendah (5.159,55 vs. 5.256,43) dan R-Squared lebih tinggi (0,6385 vs. 0,6319), sementara Extra Trees unggul pada MAE (3.604,92 vs. 3.677,67) dan MAPE (35,69% vs. 36,06%). Berdasarkan RMSE sebagai indikator utama, CART ditetapkan sebagai model terbaik. Analisis *feature importance* menunjukkan bahwa provinsi tujuan dan variabel terkait berat produk merupakan prediktor paling berpengaruh terhadap biaya pengiriman. Temuan ini menunjukkan bahwa metode berbasis pohon keputusan memiliki potensi yang baik dalam memodelkan karakteristik nonlinier pada biaya pengiriman e-commerce.

Kata Kunci: Logistik E-Commerce, Machine Learning, Prediksi Biaya Pengiriman, Regresi Cart, Regresi Extra Trees

1. Pendahuluan

Perkembangan teknologi digital telah mendorong transformasi signifikan dalam sektor perdagangan global, termasuk di Indonesia. Laporan e-Conomy SEA dari Google, Temasek Holdings, dan Bain & Company mencatat Indonesia sebagai pasar e-commerce utama di Asia Tenggara dengan transaksi yang terus tumbuh, sehingga meningkatkan kompleksitas sistem logistik dan distribusi barang. Dalam hal ini, biaya pengiriman dapat memengaruhi keputusan pembelian konsumen (Bansal & Muzatko, 2021). Struktur biaya pengiriman dipengaruhi oleh berbagai faktor operasional seperti berat barang, kuantitas pesanan, lokasi tujuan, serta metode pengiriman, di mana berat kiriman dan jarak distribusi merupakan determinan utama dalam pembentukan tarif logistik (Harani *et al.*, 2025).

Pemodelan biaya pengiriman sering kali dihadapkan pada kompleksitas struktural karena data logistik umumnya heterogen, tidak berdistribusi normal, serta mengandung outliers yang dapat memengaruhi stabilitas estimasi model (Zhao, 2024; Cebeci *et al.*, 2025). Keterkaitan antara

variabel operasional dan biaya logistik juga cenderung bersifat nonlinier, sehingga pendekatan machine learning berbasis pohon menjadi alternatif karena mampu menangkap hubungan nonlinier tanpa bergantung pada asumsi parametrik yang ketat (Chowdhury *et al.*, 2021).

Salah satu pendekatan *ensemble tree-based* yang diimplementasikan adalah *Extremely Randomized Trees* (Extra Trees) sebagaimana dipelopori oleh Geurts *et al.* (2006). Algoritma ini mengonstruksi sekumpulan pohon keputusan dengan tingkat randomisasi tinggi pada seleksi atribut dan penentuan titik pemisahan guna meningkatkan kemampuan generalisasi serta mereduksi varians estimasi model. Sebagai pembanding, digunakan Classification and Regression Tree (CART) yang diperkenalkan oleh Breiman *et al.* (1984), yaitu pendekatan pohon keputusan tunggal dengan interpretabilitas tinggi dan kerap digunakan sebagai baseline model dalam studi machine learning.

Beberapa penelitian sebelumnya telah mengkaji prediksi biaya logistik menggunakan pendekatan machine learning. Jang (2023) menggunakannya untuk memprediksi biaya pengiriman pada platform freight brokerage, Arfeen (2024) memodelkan biaya transportasi dengan mempertimbangkan variabel operasional pengiriman, sementara Harani *et al.* (2025) menunjukkan bahwa model ensemble tree-based memberikan hasil prediksi yang lebih baik dibandingkan regresi linier. Meskipun demikian, penelitian yang secara khusus membandingkan Extra Trees Regression dan CART untuk memodelkan biaya pengiriman pada data transaksi e-commerce di Indonesia masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan mengembangkan dan membandingkan model Extra Trees Regression dan CART untuk menentukan model paling optimal dalam memprediksi biaya pengiriman e-commerce di Indonesia.

2. Metodologi

2.1 Bahan dan Data

Penelitian ini menggunakan data sekunder dari dataset Indonesia E-Commerce Sales & Shipping 2023-2025 yang tersedia pada platform Kaggle (Baki Tacos 2025). Dataset ini bersumber dari operasional ritel di e-commerce yang telah dianonimkan sebelum dipublikasikan. Perubahan yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Dataset yang digunakan pada penelitian ini terdiri atas 20.848 amatan dan 19 variabel. Dari keseluruhan variabel tersebut, penelitian ini menggunakan 7 variabel yang relevan dengan tujuan pemodelan biaya pengiriman, yang mencakup variabel numerik maupun kategorik.

Tabel 1: Daftar Peubah Penelitian.

Kode	Variabel	Satuan	Sumber
Y	perkiraan_ongkos_kirim	Rp (Rupiah)	Kaggle
X1	total_weight_gr	g (Gram)	Kaggle
X2	total_qty	Unit	Kaggle
X3	provinsi	-	Kaggle
X4	opsi_pengiriman	-	Kaggle
X5	product_categories	-	Kaggle
X6	metode_pembayaran	-	Kaggle

2.2 Metode Penelitian

Penelitian ini menggunakan pendekatan pemodelan nonlinier berbasis machine learning dengan metode Extra Trees Regression dan CART untuk memprediksi biaya pengiriman produk e-commerce. Kedua metode dipilih karena mampu menangkap keterkaitan nonlinier serta interaksi variabel yang kompleks tanpa bergantung pada asumsi parametrik yang ketat seperti normalitas dan homoskedastisitas (Breiman et al., 1984). Analisis dilakukan menggunakan perangkat lunak R.

2.2.1 Preprocessing Data

Tahap ini meliputi penghapusan missing value, transformasi variabel kategorik, serta penanganan outlier pada variabel respon menggunakan metode Interquartile Range (IQR) untuk mengurangi pengaruh nilai ekstrem terhadap kestabilan model.

2.2.2 Feature Engineering

Feature engineering dilakukan melalui transformasi logaritma pada total berat produk (log_weight) untuk mengurangi skewness dan pengaruh nilai ekstrem. Selain itu, dibentuk variabel weight_per_qty untuk merepresentasikan rata-rata berat per item dalam suatu transaksi. Secara matematis dirumuskan sebagai berikut:

$$weight\ per\ qty = \frac{total\ weight\ (gr)}{total\ qty} \quad (1)$$

2.2.3 Eksplorasi Data

Eksplorasi data dilakukan untuk memahami karakteristik data sebelum proses pemodelan meliputi statistik deskriptif, visualisasi histogram, boxplot, dan analisis korelasi antar variabel numerik.

2.2.4 Pembagian data latih dan data uji

Pembagian data dilakukan secara acak dengan proporsi 80% data latih dan 20% data uji. Pada data latih diterapkan k-fold cross validation untuk meningkatkan stabilitas estimasi dan meminimalisasi overfitting. Optimasi hyperparameter Extra Trees Regression dilakukan melalui Grid Search pada rentang mtry 2 hingga 8, dengan konfigurasi terbaik ditentukan berdasarkan nilai RMSE validasi silang terkecil.

2.2.5 Pengembangan Model Extra Trees Regression

Extra Trees Regression merupakan metode ensemble yang membangun sejumlah besar pohon keputusan dengan tingkat randomisasi tinggi pada pemilihan fitur dan titik split (Hameed *et al.*, 2021). Prediksi akhir dihitung sebagai rata-rata seluruh prediksi pohon:

$$\hat{f}(x) = \left(\frac{1}{T}\right) \sum f_t(x) \quad (2)$$

di mana:

$f_t(x)$ adalah prediksi dari pohon ke- t

T adalah jumlah pohon dalam ensemble

2.2.6 Pengembangan Model Classification and Regression Tree (CART)

Sebagai model pembanding digunakan CART (Breiman *et al.*, 1984), metode nonparametrik berbasis pohon keputusan yang membentuk struktur pohon secara rekursif untuk meminimalkan galat prediksi pada setiap node. Proses pruning dilakukan untuk mengurangi kompleksitas pohon dan menghindari overfitting.

2.2.7 Evaluasi dan Penentuan Model Terbaik

Evaluasi performa model dilakukan menggunakan metrik evaluasi untuk memberikan penilaian yang lebih komprehensif terhadap akurasi prediksi, dengan Root Mean Squared Error (RMSE) dan Mean Absolute Percentage Error (MAPE) sebagai indikator utama (Chharia *et al.*, 2024).

a. RMSE

RMSE digunakan untuk mengukur rata-rata besar kesalahan prediksi model terhadap nilai aktual dalam satuan yang sama dengan variabel respon. Semakin kecil nilai RMSE maka semakin baik performa model.

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

Dengan:

y_i = nilai aktual
 $\hat{y}_i$ = nilai prediksi
 n = jumlah observasi

b. MAPE

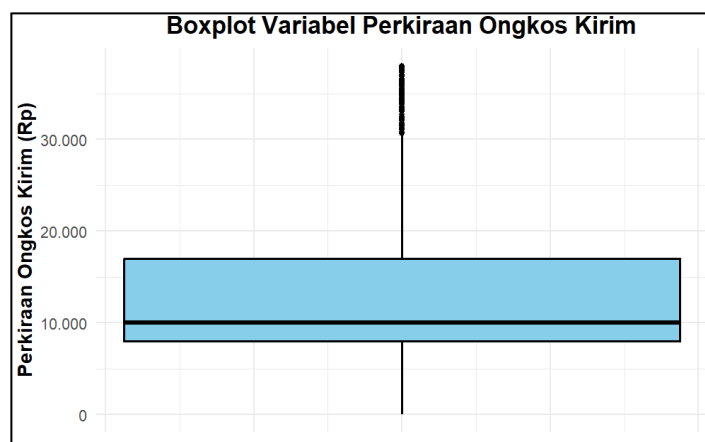
MAPE digunakan untuk mengukur rata-rata persentase kesalahan prediksi model terhadap nilai aktual. Semakin kecil nilai MAPE maka semakin baik tingkat akurasi model.

$$MAPE = 100\% \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

dengan notasi yang sama seperti persamaan RMSE di atas.

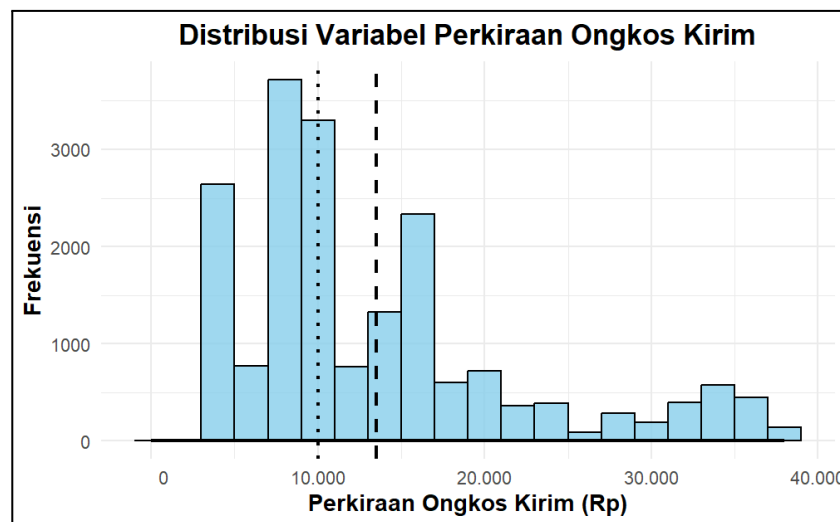
3. Hasil dan Pembahasan

3.1 Eksplorasi



Gambar 1: Visualisasi Boxplot Variabel Respon.

Visualisasi melalui boxplot pada Gambar 1 mengidentifikasi keberadaan sejumlah amatan ekstrem yang mengindikasikan distribusi variabel respon bersifat asimetris. Fenomena ini dipertegas oleh *histogram* pada Gambar 2 yang menunjukkan pola *right-skewed* dengan konsentrasi observasi dominan pada rentang Rp 8.000–Rp 11.000, di mana posisi rata-rata berada di sebelah kanan median. Karakteristik tersebut menunjukkan bahwa data respon tidak berdistribusi normal dan mengandung nilai ekstrem, sehingga pendekatan pemodelan berbasis pohon seperti Extra Trees Regression dan CART dipilih karena lebih fleksibel dalam menangani data dengan distribusi yang tidak normal tanpa bergantung pada asumsi parametrik yang ketat.



Gambar 2: Visualisasi Histogram Variabel Respon.

3.2 Hasil Model Extra Trees Regression

Model Extra Trees dibangun melalui Grid Search dengan konfigurasi terbaik pada $mtry = 8$ dan $min.node.size = 5$. Temuan ini mengindikasikan bahwa informasi prediktif tersebar merata di seluruh peubah sehingga penggunaan semua peubah pada setiap split menghasilkan performa terbaik. Model akhir mencatatkan OOB MSE sebesar 22.576.644 dan R^2 OOB sebesar 0,6924, yang menjelaskan sekitar 69,24% variasi biaya pengiriman. Kontribusi variabel dihitung menggunakan permutation importance yang dinormalisasi ke bentuk persentase, disajikan pada Tabel 2.

Tabel 2: *Feature Importance* Model Extra Trees Regression

Feature	Importance (%)	Peringkat
Provinsi	50.49	1
log_weight	11.08	2
Opsi.Pengiriman	9.39	3
weight_per_qty	8.81	4
product_categories	8.47	5
total_weight_gr	7.18	6
total_qty	3.34	7
Metode.Pembayaran	1.24	8

Analisis kontribusi peubah pada model Extra Trees Regression menunjukkan bahwa variabel Provinsi merupakan prediktor paling dominan dengan tingkat kepentingan mencapai 50,49%. Temuan ini mengindikasikan bahwa faktor geografis merupakan determinan utama dalam pembentukan biaya pengiriman. Variabel log_weight menempati posisi kedua sebesar 11,08%, disusul oleh Opsi.Pengiriman sebesar 9,39%, weight_per_qty sebesar 8,81%, product_categories sebesar 8,47%, dan total_weight_gr sebesar 7,18%. Di sisi lain, variabel Metode.Pembayaran mencatatkan tingkat kepentingan terendah sebesar 1,24%, menunjukkan bahwa kontribusinya terhadap pembentukan biaya pengiriman relatif lebih kecil dibandingkan peubah lainnya.

3.3 Hasil Model Classification and Regression Tree (CART)

Model CART dibangun menggunakan 15.229 data latih dengan parameter awal cp 0,001, $minsplit$ 20, dan $maxdepth$ 10. Melalui mekanisme pruning validasi silang, dihasilkan pohon final dengan 41 titik pemisahan serta nilai galat relatif 0,3017 dan $xerror$ 0,3510. Kontribusi masing-masing variabel dihitung berdasarkan akumulasi reduksi RSS pada setiap titik pemisahan dan kemudian dinormalisasi ke dalam bentuk persentase. Hasil feature importance yang diperoleh disajikan pada Tabel 3.

Selaras dengan temuan pada model Extra Trees Regression, variabel Provinsi merupakan peubah dengan tingkat kepentingan tertinggi sebesar 37%, yang menunjukkan bahwa faktor geografis memiliki pengaruh dominan terhadap variasi biaya pengiriman. Variabel total_weight_gr, log_weight, dan Opsi.Pengiriman juga memiliki kontribusi yang cukup besar, sehingga karakteristik berat produk dan pilihan layanan pengiriman turut berperan dalam pembentukan tarif pengiriman. Sementara itu, variabel

total_qty memiliki tingkat kepentingan paling rendah sehingga kontribusinya relatif lebih kecil dibandingkan peubah lainnya.

Tabel 3: *Feature Importance* Model CART Regression (Penurunan RSS).

Feature	Importance (%)	Peringkat
Provinsi	37	1
total_weight_gr	13	2
log_weight	13	3
Opsi.Pengiriman	13	4
weight_per_qty	11	5
product_categories	10	6
total_qty	2	7

3.4 Evaluasi dan Penentuan Model Terbaik

Tahap evaluasi model diimplementasikan dengan memanfaatkan data uji sebanyak 3.804 amatan dengan enam metrik evaluasi disajikan secara komprehensif pada Tabel 4.

Tabel 4: Perbandingan Metrik Evaluasi Model Extra Trees Regression dan CART.

Metrik	Extra Trees	CART
MSE	27.630.048	26.620.986
RMSE	5.256,429	5.159,553
MAE	3.604,916	3.677,674
MAPE (%)	35,68919	36,06248
R-Squared	0,6319224	0,638549
Korelasi Pearson	0,7949354	0,7990926

Hasil komparasi menunjukkan bahwa kedua model memiliki keunggulan pada metrik yang berbeda. Model CART mencatatkan nilai MSE, RMSE, R-Squared, dan korelasi Pearson yang lebih baik dibandingkan Extra Trees, dengan RMSE sebesar Rp 5.159,55 dan R-Squared 0,6385. Di sisi lain, Extra Trees mengungguli CART pada metrik MAE sebesar Rp 3.604,92 dan MAPE sebesar 35,69%. Berdasarkan RMSE sebagai indikator utama, CART ditetapkan sebagai model terbaik. Hasil ini mengindikasikan bahwa pola hubungan dalam dataset dapat direpresentasikan secara efektif oleh struktur pohon tunggal hasil pruning. Nilai MAPE yang relatif tinggi perlu ditafsirkan dengan hati-hati karena sensitivitasnya terhadap amatan bernilai kecil.

4. Simpulan dan Saran

Penelitian ini memodelkan biaya pengiriman e-commerce di Indonesia menggunakan Extra Trees dan CART Regression berbasis 19.033 data transaksi. Hasil evaluasi menunjukkan bahwa CART lebih unggul pada MSE, RMSE, R-Squared, dan korelasi Pearson, sementara Extra Trees unggul pada MAE dan MAPE. Menimbang RMSE sebagai indikator utama, CART Regression ditetapkan sebagai model terbaik dengan RMSE Rp 5.159,55 dan R-Squared 0,6385. Pada kedua model, variabel Provinsi secara konsisten menjadi peubah paling dominan, diikuti oleh variabel berbasis berat dan opsi pengiriman.

Keterbatasan studi ini memberikan peluang pengembangan ke depan. Pertama, penggunaan jarak pengiriman aktual sebagai prediktor dapat menggantikan variabel Provinsi sebagai proksi jarak, yang berpotensi meningkatkan presisi prediksi secara substantif. Kedua, perluasan perbandingan dengan metode *ensemble* lain seperti Gradient Boosting atau XGBoost disarankan guna memperoleh gambaran komprehensif mengenai performa berbagai model nonlinier pada logistik e-commerce. Ketiga, optimasi hiperparameter CART secara sistematis juga dapat dieksplorasi untuk memastikan perbandingan yang lebih setara antara kedua model.

Daftar Pustaka

- Afshar, F., Seyedabrishami, S., & Moridpour, S. (2022). Application of extremely randomised trees for exploring influential factors on variant crash severity data. *Scientific Reports*, 12, 11476. <https://doi.org/10.1038/s41598-022-15693-7>
- Arfeen, M. A. (2024). Machine learning-driven cost prediction in intermodal transport systems. *Journal of Information Systems Engineering and Management*. <https://jisem-journal.com/index.php/journal/article/view/8569>
- Baki Tacos. 2025. Indonesia E-Commerce Sales & Shipping 2023–2025. Kaggle. <https://www.kaggle.com/datasets/bakitacos/indonesia-e-commerce-sales-and-ship-ping-20232025>
- Bansal, G., & Muzatko, S. (2021). Shipping and return shipping prices and online purchase intentions: The role of gender and product price category. *International Journal of Business*, 26(3), 53–71. <https://ijb.cyut.edu.tw/var/file/10/1010/img/V26N3-4.pdf>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

<https://doi.org/10.1023/A:1010933404324>

- Cebeci, M. S., de Bok, M., Tapia, R., Nadi, A., & Tavasszy, L. (2025). Feasibility of crowdshipping for outlier parcels in last-mile delivery. *Research in Transportation Economics*, 112, 101607. <https://doi.org/10.1016/j.retrec.2025.101607>
- Chharia, A., et al. (2024). Accuracy of US CDC COVID-19 Forecasting Models. *Frontiers in Public Health*. <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1359368/full>
- Chowdhury, S., et al. (2021). Evaluation of tree-based regression over multiple linear regression for non-normally distributed data. *arXiv preprint arXiv:2111.02513*. <https://arxiv.org/abs/2111.02513>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Google, Temasek & Bain & Company. (2025). *e-Conomy SEA 2025: ASEAN's digital economy enters the AI reality* (10th edition). <https://economysea.withgoogle.com/>
- Hameed, M., et al. (2021). Application of extra trees regression for environmental modeling. *Data in Brief*, 34, 106824. <https://doi.org/10.1016/j.dib.2021.106824>
- Harani, N. H., Setyawan, M. Y. H., & Ferdinan, D. (2025). Predicting basic shipping tariff using machine learning. *Nuansa Informatika*, 19(2), 58–66. <https://journal.fkom.uniku.ac.id/ilkom/article/view/388>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://hastie.su.domains/ElemStatLearn/>
- Jang, H. S. (2023). Prediction of shipping cost on freight brokerage platform using machine learning. *Sustainability*, 15(2), 1122. <https://doi.org/10.3390/su15021122>
- Zhao, S. (2024). Research on logistics cost prediction based on random forest regression modeling. *Transactions on Economics, Business and Management Research*, 11, 367–376. <https://doi.org/10.62051/mv8p5485>
- Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G., & Wu, J. (2023). A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4), 2188–2216. <https://doi.org/10.3390/jtaer18040110>

Zhou, Z.-H. (2021). Ensemble methods: Foundations and algorithms (2nd ed.). CRC Press. <https://tjzhifei.github.io/links/EMFA.pdf>