Bridge: Jurnal Publikasi Sistem Informasi dan Telekomunikasi
Volume 4, Nomor 3, Agustus 2026
e-ISSN: 3046-725X; p-ISSN: 3046-7268, Hal. 25-34
DOI: https://doi.org/10.62951/bridge.v4i3.967
Tersedia: https://journal.aptii.or.id/index.php/Bridge

Klasifikasi *Compliance* dan *Clustering Risk Profiling Supplier* pada Data *Procurement*

Wahyu Rahman Hakim[1*], Fitriah[2]
[1-2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah, Indonesia
**Penulis Korespondensi: waraha93@gmail.com*

Abstract*. The cost efficiency and supply chain quality of a company are strongly influenced by how well purchase orders comply with established policies and how consistently supplier performance is maintained. This study builds a classification model to predict the Compliance status of a purchase order, while also grouping suppliers into risk profiles through Clustering. The data source used is the Procurement KPI Analysis Dataset from Kaggle, covering 777 rows of purchase order data from 5 suppliers. Prior to modelling, an investigation into missing value patterns pointed toward a Missing Not At Random (MNAR) mechanism on the Defective_Units attribute, so an imputation strategy accompanied by an indicator feature (missing flag) was applied instead of row deletion. Three classification algorithms were tested Logistic Regression, Random Forest, and Gradient Boosting while supplier risk grouping used the K-Means algorithm. Random Forest emerged as the best-performing model, recording an F1-Score of 0.899 and a Recall of 0.977, with Defect_Rate as the most decisive feature. On the Clustering side, two of the five suppliers fell into the Higher Risk category. These findings confirm that combining classification and Clustering, coupled with careful examination of Missing Data, produces a more complete understanding to support data-driven Procurement decision-making.*

Keywords*: Classification; Clustering; K-Means; Missing Data; Random Forest.*

Abstrak. Efisiensi biaya dan kualitas rantai pasok suatu perusahaan sangat dipengaruhi oleh sejauh mana purchase order dijalankan secara patuh (*Compliance*) serta bagaimana performa supplier terjaga dari waktu ke waktu. Studi ini membangun model klasifikasi untuk memprediksi status *Compliance* suatu purchase order, sekaligus mengelompokkan supplier ke dalam profil risiko melalui *Clustering*. Sumber data yang dipakai adalah *Procurement* KPI Analysis Dataset dari Kaggle, mencakup 777 baris data purchase order dari 5 supplier. Sebelum pemodelan, dilakukan penelusuran pola missing value yang mengarah pada mekanisme Missing Not At Random (MNAR) pada atribut Defective_Units, sehingga diterapkan strategi imputasi disertai fitur indikator (missing flag), bukan penghapusan baris. Tiga algoritma klasifikasi diuji *Logistic Regression*, Random Forest, dan Gradient Boosting sementara pengelompokan risiko supplier memakai algoritma K-Means. Random Forest tampil sebagai model terbaik, mencatatkan F1-Score 0,899 dan Recall 0,977, dengan Defect_Rate sebagai fitur paling menentukan. Pada sisi *Clustering*, dua dari lima supplier masuk kategori risiko tinggi (*Higher Risk*). Temuan ini menegaskan bahwa penggabungan klasifikasi dan *Clustering*, ditambah kecermatan menelaah *Missing Data*, menghasilkan pemahaman lebih utuh untuk mendukung pengambilan keputusan *Procurement* berbasis data.

Kata kunci: *Clustering*; K-Means; Klasifikasi; *Missing Data*; Random Forest.

1. LATAR BELAKANG

Peran pengadaan (*Procurement*) berada pada posisi krusial dalam operasional perusahaan, karena terkait langsung dengan efisiensi biaya, kualitas produk, dan kelangsungan rantai pasok. Salah satu tantangan yang sering dihadapi tim *Procurement* adalah memastikan setiap purchase order (PO) tetap sesuai dengan kebijakan yang berlaku (*Compliance*), sambil terus memantau kinerja setiap supplier secara terus-menerus. Saat sebuah pesanan tidak sesuai (*non-Compliance*), dampaknya dapat mengakibatkan kerugian finansial, penundaan proses produksi, hingga penurunan kualitas produk akhir. Kehadiran Data *Science* menciptakan kesempatan bagi proses pengambilan keputusan pengadaan untuk beralih dari kebiasaan lama yang mengandalkan intuisi menuju pendekatan yang lebih objektif dan berdasarkan data (data-

Naskah Masuk: 27 Mei 2026; Revisi: 20 Juni 2026; Diterima: 30 Juli 2026; Terbit: 05 Agustus 2026

driven decision making). Penelitian sebelumnya telah membahas permasalahan klasifikasi dan pengelompokan dalam pengelolaan pemasok maupun proses pengadaan. Di antaranya memanfaatkan metode *MARCOS* sebagai pendekatan dalam penilaian dan pemilihan pemasok (Abdulla et al., 2023) menerapkan algoritma *Random Forest* pada sistem pendukung keputusan untuk mengklasifikasikan kriteria pemasok (Ali et al., 2023), mengembangkan model analitik *prediktif* guna mengurangi risiko pada rantai pasok (Aljohani, 2023), serta menggunakan algoritma *K-Means* untuk melakukan segmentasi pemasok berdasarkan karakteristik keberlanjutan (Rahiminia et al., 2026).

Sebagian besar penelitian tersebut masih membahas proses klasifikasi dan *Clustering* sebagai dua pendekatan yang berdiri sendiri. Selain itu, informasi yang hilang (*missing values*) umumnya hanya diperlakukan sebagai permasalahan kualitas data dan belum banyak dimanfaatkan sebagai bagian dari strategi rekayasa fitur. Padahal, ketiga aspek tersebut berpotensi saling melengkapi dalam menghasilkan gambaran kondisi risiko pengadaan yang lebih menyeluruh. Penelitian ini bertujuan untuk mengembangkan dan membandingkan model klasifikasi dalam memprediksi status kepatuhan *purchase order*, mengelompokkan pemasok ke dalam profil risiko dengan metode *Clustering*, serta menganalisis pola nilai yang hilang untuk menetapkan strategi penanganan yang paling sesuai.

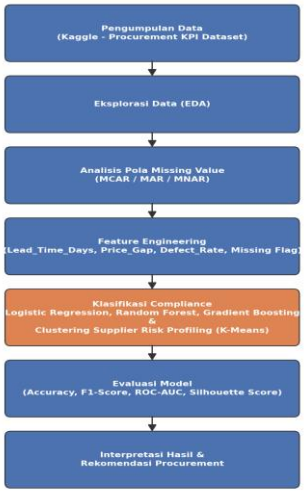
2. KAJIAN TEORITIS

Ilmu data merupakan bidang lintas disiplin yang menggabungkan statistik, ilmu komputer, dan pemahaman domain untuk memperoleh wawasan dari data, mengikuti langkah-langkah *Knowledge Discovery in Databases* (KDD): pemilihan, pra-pengolahan, transformasi, penambangan data, dan interpretasi. *Logistic Regression* memodelkan probabilitas kelas melalui fungsi sigmoid sebagai baseline klasifikasi biner. *Random Forest*, algoritma *ensemble* berbasis *bagging decision tree*, dikenal tangguh terhadap data *imbalanced*, sedangkan *Gradient Boosting* membangun model secara bertahap dengan tiap model memperbaiki kesalahan model sebelumnya. *K-Means* mengelompokkan data ke dalam *k cluster* berdasarkan jarak *Euclidean* terhadap *centroid*, dengan jumlah *cluster* optimal ditentukan melalui *Elbow Method* dan divalidasi *Silhouette Score*. Penelitian terdahulu pada ranah supplier *management* dan *Procurement* banyak membahas klasifikasi dan *Clustering* secara terpisah, seperti integrasi *machine learning* dengan metode *MARCOS* untuk seleksi *supplier* (Abdulla et al., 2023), decision support system berbasis Random Forest (Ali et al., 2023), *predictive analytics* untuk mitigasi risiko rantai pasok (Aljohani, 2023), federated *machine learning* untuk prediksi risiko kolektif (Zheng et al., 2023), serta *K-Means* untuk segmentasi supplier Rahiminia et al., (2026)

dan profiling pelanggan (Barus et al., 2023). Pada isu data imbalanced, algoritma berbasis pohon keputusan seperti *Random Forest* dan *Gradient Boosting* terbukti lebih tangguh dibanding model linear (Dube & Verster, 2023; Hanisah et al., 2023; Kaope & Pristyanto, 2023), sementara teknik *SMOTE* efektif meningkatkan deteksi kelas minoritas (Hairani et al., 2023; Wongvorachan & He, 2023). Pada isu *missing value*, sejumlah studi menekankan perlunya melampaui klasifikasi *MCAR/MAR/MNAR* sederhana (Lee et al., 2023), sebab mekanisme *MNAR* yang tak termodelkan tepat berpotensi menimbulkan (Chen & Wang, 2023), sehingga *missing indicator method* terbukti efektif terutama pada model berbasis pohon keputusan (Alam et al., 2023; Ness et al., 2023; Sisk et al., 2023). Berdasarkan tinjauan tersebut, teridentifikasi celah penelitian: klasifikasi dan *Clustering* jarang dibahas bersamaan, dan pola *missing value* jarang dianalisis sebagai bagian strategi *feature engineering*, sebagaimana coba dijawab oleh penelitian ini.

3. METODE PENELITIAN

Riset ini dikategorikan sebagai penelitian terapan (*applied research*) yang mengusung paradigma kuantitatif bertumpu pada data mining. Materi penelitian diperoleh dari *Procurement KPI Analysis Dataset* yang tersedia di platform *Kaggle*, memuat 777 baris data *purchase order* dengan 11 atribut, meliputi identitas pesanan, nama pemasok, tarikh order maupun distribusi barang, klasifikasi komoditas, status pesanan, volume, nominal harga, jumlah unit yang cacat, serta status *Compliance* (Ya/Tidak) sebagai variabel target klasifikasi. Rangkaian alur penelitian secara garis besar divisuali sasikan pada Gambar 1.



Gambar 1. Kerangka penelitian.

Sumber Data dan Penanganan *Missing Value*

Pemeriksaan awal mendapati dua atribut yang mengandung *missing value*, yakni *Delivery_Date* (87 baris, 11,2%) dan *Defective_Units* (136 baris, 17,5%). Ketika hubungan antara status *missing* tersebut dengan target *Compliance* diuji, terlihat bahwa baris dengan *Defective_Units* kosong memiliki proporsi *Compliance* = No sebesar 25,7%, jauh melampaui baris yang datanya lengkap (15,9%), sebuah indikasi kuat adanya mekanisme Missing Not At Random (MNAR). Bertolak dari temuan tersebut, *missing value* tidak dihapus melainkan ditangani lewat imputasi yang disertai fitur *flag* penanda (*Defective_Units_Missing_Flag, Lead_Time_Missing_Flag*), mengikuti pendekatan *missing indicator method* (Ness et al., 2023; Sisk et al., 2023).

***Feature Engineering* dan Algoritma Klasifikasi**

Tiga fitur tambahan diracik untuk memperkaya informasi yang tersedia bagi model, yaitu Lead_Time_Days (selisih hari antara Order_Date dan Delivery_Date), *Price_Gap* (selisih Unit_Price dengan *Negotiated_Price*), serta Defect_Rate (perbandingan Defective_Units terhadap *Quantity*). Untuk memprediksi status *Compliance*, tiga algoritma dijajal, yaitu Logistic Regression sebagai pembanding dasar, Random Forest, dan Gradient Boosting, dipilih mengingat ketahanan Random Forest dan Gradient Boosting terhadap data imbalanced dibanding model linear yang lebih sederhana (Dube & Verster, 2023; Hanisah et al., 2023; Wongvorachan & He, 2023). Pembagian data mengikuti skema 80% data latih dan 20% data uji melalui stratified split. Parameter yang diterapkan: Random Forest dengan n_estimators=200 dan class_weight='balanced'; Gradient Boosting dengan n_estimators=200.

Algoritma *Clustering* dan Metode Evaluasi

Pengelompokan risiko supplier (*supplier risk profiling*) mengandalkan algoritma *K-Means* yang bekerja pada data agregat tiap supplier (rata-rata *defect rate, Compliance rate, lead time, price gap,* dan *cancel rate*) setelah dinormalisasi memakai *StandardScaler*. *K-Means* bekerja secara *iteratif* menghitung jarak *Euclidean* tiap data terhadap *centroid*, dirumuskan pada Persamaan (1):

$$d(x, c) = \sqrt{\sum(x_1 - c_1)^2} \quad \text{(i)}$$

dengan x menyatakan data, c menyatakan *centroid*, dan i adalah indeks dimensi fitur. Performa model klasifikasi diukur menggunakan *Accuracy, Precision, Recall, F1-Score,* dan *ROC-AUC*, dengan bobot perhatian lebih besar pada *Recall* dan *F1-Score* kelas minoritas mengingat data yang dipakai bersifat *imbalanced*, dirumuskan berturut-turut pada Persamaan (2) sampai (6):

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad \text{.....(ii)}$$

$$Precision = \frac{TP}{(TP+FP)} \quad \text{.....(iii)}$$

$$Recall = \frac{TP}{(TP+FN)} \quad \text{.....(iv)}$$

$$F1 - Score = \frac{2 \text{ x (Precision x Recall)}}{Precision+Recall} \quad \text{.....(v)}$$

$$AUC = \int 0^1 TPR(FPR)\, d(FPR) \quad \text{.....(vi)}$$

Dengan TP, TN, FP, dan FN berturut-turut menyatakan *True Positive*, *True Negative, False Positive, dan False Negative*, sementara TPR dan FPR menyatakan *True Positive Rate dan False Positive Rate*. Kualitas hasil *Clustering* diukur *menggunakan Silhouette Score* pada Persamaan (7), sedangkan penentuan jumlah *cluster* optimal (k) turut divalidasi memakai *Elbow Method* yang mengamati penurunan nilai *Within-Cluster Sum of Squares (WCSS)* seiring bertambahnya k, dirumuskan pada Persamaan (8):

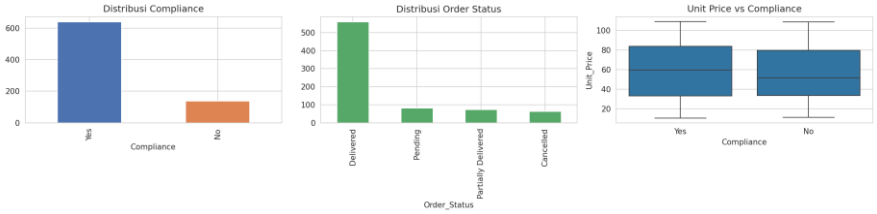
$$s(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \quad \text{.....(vii)}$$

$$WCSS = \sum k = 1^K \sum x \in C_k \|x - \mu_k\|^2 \quad \text{.....(viii)}$$

dengan a(i) dan b(i) berturut-turut menyatakan rata-rata jarak titik i terhadap titik lain dalam *cluster* yang sama (*intra-cluster*) dan rata-rata jarak titik i terhadap cluster tetangga terdekat, sementara Ck dan μk menyatakan anggota dan *centroid cluster* ke-k. Titik k optimal dipilih pada posisi "siku" (*elbow*), yaitu saat penurunan nilai *WCSS* mulai melandai secara signifikan.

4. HASIL DAN PEMBAHASAN

Penelusuran awal data memperlihatkan sebaran target *Compliance* yang timpang (*imbalanced*), yaitu 82,4% berstatus *Yes* berbanding 17,6% berstatus No. Ketimpangan semacam ini menjadi salah satu pertimbangan utama saat memilih algoritma maupun metrik evaluasi pada tahap klasifikasi.



Gambar 2. Distribusi *Compliance* , Order Status, dan Unit Price terhadap *Compliance.*

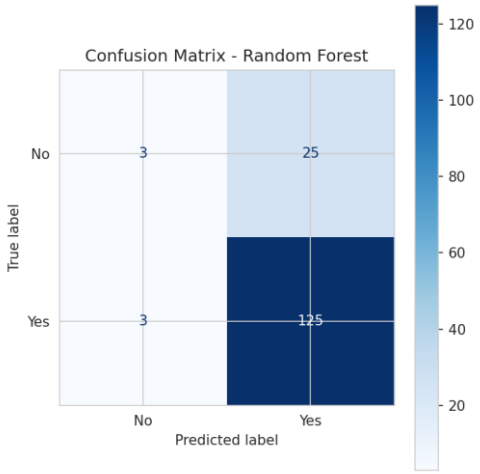
Hasil Klasifikasi *Compliance*

Perbandingan performa ketiga model klasifikasi pada data uji (156 baris) dirangkum pada Tabel 1.

Tabel 1. Perbandingan Performa Model Klasifikasi.

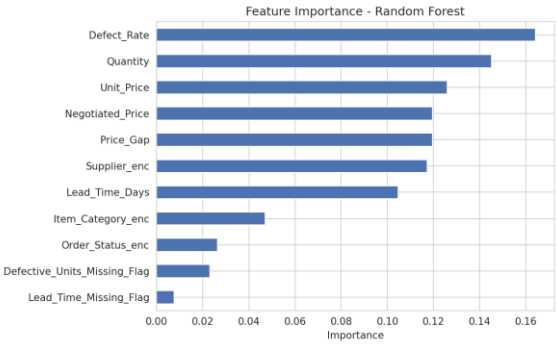
Model	***Accuracy***	***F1-Score***	***Recall***	***ROC-AUC***
Logistic Regression	0,673	0,765	0,648	0,728
Random Forest	0,821	0,899	0,977	0,716
Gradient Boosting	0,776	0,869	0,906	0,680

Dari ketiga model yang diuji, *Random Forest* tampil paling unggul dengan *F1-Score* 0,899 dan *Recall* 0,977. Penelusuran lebih jauh lewat *confusion matrix* (Gambar 3) memperlihatkan bahwa dari 28 order berstatus No pada data uji, model hanya sanggup mengenali 3 di antaranya dengan tepat, sementara 125 dari 128 order berstatus *Yes* berhasil dikenali dengan benar, sebuah indikasi bahwa model masih condong ke kelas mayoritas kendati parameter *class_weight='balanced'* sudah diaktifkan.



Gambar 3. *Confusion Matrix* model *Random Forest.*

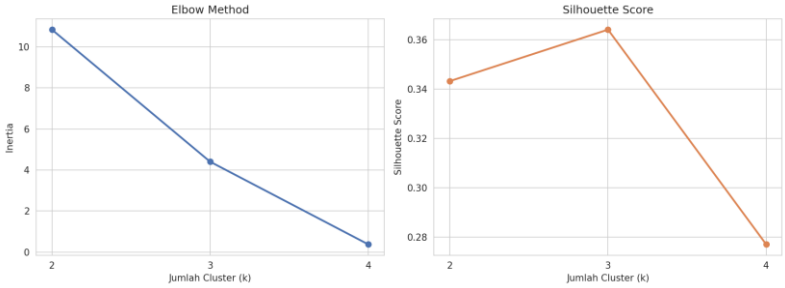
Tingkat kepentingan tiap fitur (*feature importance*) pada Gambar 4 menunjukkan *Defect_Rate* sebagai fitur paling berpengaruh, disusul *Quantity*, *Unit_Price*, *Negotiated_Price*, dan *Price_Gap*. *Fitur Defective_Units_Missing_Flag* turut memberi kontribusi berarti, memperkuat dugaan bahwa hilangnya data pada atribut tersebut bukan sekadar gangguan acak *(noise*), melainkan mengandung informasi (*MNAR*), selaras dengan Van Ness dkk (Ness et al., 2023).



Gambar 4. Feature Importance model Random Forest.

Hasil *Clustering Supplier Risk Profiling*

Pengujian *Elbow Method* berikut *Silhouette Score* (Gambar 5) pada rentang k=2 sampai k=4 memperlihatkan bahwa k=3 mencatat *Silhouette Score* tertinggi (0,364), unggul tipis atas k=2 (0,343). Walau begitu, k=2 tetap dipilih sebab pertimbangan kepraktisan interpretasi bisnis, mengingat dengan hanya 5 supplier yang tersedia, k=3 berpotensi memunculkan *cluster* yang cuma diisi 1-2 supplier saja.



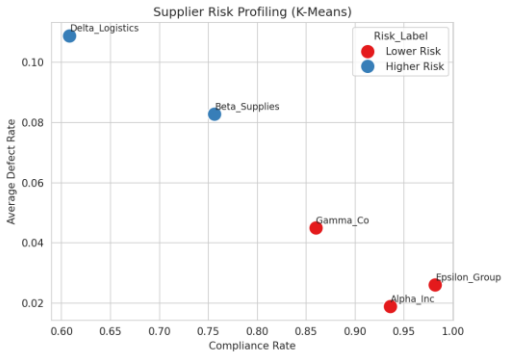
Gambar 5. Elbow Method dan Silhouette Score.

Tabel 2 memuat hasil agregasi performa tiap supplier berikut label risikonya menurut hasil *Clustering K-Means.*

Tabel 2. Hasil *Clustering* Supplier Risk Profiling.

Supplier	***Defect Rate***	***Compliance Rate***	***Cancel Rate***	**Label Risiko**
Alpha_Inc	1,9%	93,6%	9,2%	*Lower Risk*
Beta_Supplies	8,3%	75,6%	9,0%	*Higher Risk*
Delta_Logistics	10,9%	60,8%	9,4%	*Higher Risk*
Epsilon_Group	2,6%	98,2%	8,4%	*Lower Risk*
Gamma_Co	4,5%	86,0%	4,2%	*Lower Risk*

Dua supplier, *Beta_Supplies* dan *Delta_Logistics*, berada di kategori *Higher Risk* lantaran *Defect Rate-nya* lebih tinggi sekaligus *Compliance Rate-nya* lebih rendah dibanding tiga supplier lain. Pola sebaran pada Gambar 6 turut memperlihatkan konsistensi yang masuk akal secara logis, yaitu supplier dengan *defect rate tinggi* cenderung berada pada *Compliance* rate yang lebih rendah, memperkuat validitas hasil *Clustering*, sejalan dengan pendekatan Rahiminia dkk (Rahiminia et al., 2026).



Gambar 6. Visualisasi Supplier Risk Profiling hasil K-Means.

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan *Random Forest* sebagai model klasifikasi terbaik untuk memprediksi *Compliance purchase order*, dengan *F1-Score* 0,899 dan *Recall* 0,977. Penelusuran pola *missing value* turut mengonfirmasi mekanisme *MNAR* pada atribut *Defective_Units*, sehingga strategi imputasi disertai missing flag terbukti lebih tepat dibanding menghapus baris data, sekaligus mendongkrak performa klasifikasi. *Clustering K-Means* juga menempatkan dua dari lima supplier, yaitu *Beta_Supplies* dan *Delta_Logistics*, ke kategori risiko tinggi *(Higher Risk)*, melengkapi hasil klasifikasi di level operasional dengan gambaran risiko di level strategis. Pengembangan lebih lanjut dapat diarahkan pada penerapan metode *balancing* data, misalnya *SMOTE* guna mempertajam deteksi kelas minoritas, penambahan dimensi time-series untuk menelusuri tren performa supplier, serta *hyperparameter tuning* guna meningkatkan nilai *ROC-AUC.* Penambahan jumlah sampel dan variasi supplier juga disarankan agar hasil *Clustering* lebih merepresentasikan kondisi lapangan.

DAFTAR REFERENSI

Abdulla, A., Baryannis, G., & Badi, I. (2023). An integrated machine learning and MARCOS method for supplier evaluation and selection. *Decision Analytics Journal*, *9*(October), 100342. https://doi.org/10.1016/j.dajour.2023.100342

Alam, S., Sohaib, M., Arora, S., & Asad, M. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing *Clustering* and classification analysis validity. *Decision Analytics Journal*, *9*(October), 100341. https://doi.org/10.1016/j.dajour.2023.100341

Ali, R., Ashiquzzaman, S., & Ahmed, S. (2023). A decision support system for classifying supplier selection criteria using machine learning and random forest approach. *Decision Analytics Journal*, *7*(March), 100238. https://doi.org/10.1016/j.dajour.2023.100238

Aljohani, A. (2023). *Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility*.

Barus, O. P., Nathasya, C., & Pangaribuan, J. J. (2023). *Mathematical Modelling of Engineering Problems The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering*. *10*(1), 298–303.

Chen, J., & Wang, P. (2023). Deep Generative Imputation Model for Missing Not At Random Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21â•fi25, 2023, Birmingham, United Kingdom* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3583780.3614835

Dube, L., & Verster, T. (2023). *Enhancing classification performance in imbalanced datasets : A comparative analysis of machine learning models*. *3*(September), 354–379.

Hairani, H., Anggrawan, A., & Priyanto, D. (2023). *Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link*. *7*(March), 258–264.

Hanisah, N., Malek, A., Fairos, W., Yaacob, W., Wah, Y. B., Nasir, S. A., Shaadan, N., & Indratno, S. W. (2023). *Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data*. *29*(1), 598–608. https://doi.org/10.11591/ijeecs.v29.i1.pp598-608

Kaope, C., & Pristyanto, Y. (2023). *The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance*. *22*(2), 227–238. https://doi.org/10.30812/matrik.v22i2.2515.

Lee, K. J., Carlin, J. B., Simpson, J. A., Moreno-betancur, M., & Moreno-betancur, M. (2023). *classification*.

Ness, M. Van, Bosschieter, T. M., Halpin-gregorio, R., & Udell, M. (2023). The Missing Indicator Method : From Low to High Dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6â•fi10, 2023, Long Beach, CA, USA* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3580305.3599911.

Rahiminia, M., Razmi, J., Farahani, S. S., & Sabbaghnia, A. (2026). *Cluster-based supplier segmentation : a sustainable data-driven approach*. *5*(3), 209–228. https://doi.org/10.1108/MSCRA-05-2023-0017.

Sisk, R., Sperrin, M., Peek, N., Smeden, M. Van, & Martin, G. P. (2023). *Imputation and missing indicators for handling Missing Data in the development and deployment of clinical prediction models : A simulation study*. *32*(8), 1461–1477. https://doi.org/10.1177/09622802231165001

Wongvorachan, T., & He, S. (2023). *A Comparison of Undersampling , Oversampling , and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining*.

Zheng, G., Kong, L., & Brintrup, A. (2023). *Federated machine learning for privacy preserving , collective supply chain risk prediction*. *7543*. https://doi.org/10.1080/00207543.2022.2164628