SMARTICS Journal, Vol.11 No.02 2025. p79-90
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v11i2.13760

Prediksi Keberlanjutan UMKM Menggunakan Pseudo-Labeling Berbasis Composite Index dan Model Ensemble Machine Learning

Terttiaavini[a*], Tedy Setiawan Saputra[b], Lesfandra[c]
[a]Universitas Indo Global Mandiri, Palembang, Indonesia
[b]Sekolah Tinggi Ilmu Ekonomi Aprin, Palembang, Indonesia
[c]Politeknik Sains Seni Rekakreasi, Bogor, Indonesia
[*]correspondence email : avini.saputra@uigm.ac.id

Abstract*— SME sustainability plays a crucial role in strengthening local and national economies; however, many enterprises face high risks due to limited access to capital, demand volatility, and disparities in operational capacity. This study aims to develop an SME sustainability prediction model using composite index–based pseudo-labeling and ensemble machine learning. Pseudo-labels are constructed from six key indicators revenue, profit, operating costs, average production volume, business scale, and number of employees and categorized into three sustainability classes. The dataset consists of 400 SMEs operating in Palembang City, with Random Forest based feature selection employed to identify the most relevant variables. The evaluated base models include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting, while ensemble approaches comprise Bagging, Boosting, and Stacking. The results indicate that Logistic Regression achieves perfect accuracy (100%) on the test set, suggesting potential overfitting, whereas the Stacking ensemble provides more stable predictions with an F1-score of 0.918. Statistical validation using the Friedman and Wilcoxon tests confirms the superior performance of ensemble models compared to single learners. The contributions of this study include an objective pseudo-labeling method for unlabeled datasets and the development of a robust predictive model that can support policymakers and SME stakeholders in identifying sustainability risks and implementing evidence-based interventions.

Index Terms*: *SMEs; Pseudo-labeling; Composite index; Ensemble learning; Random Forest

Abstrak*– Keberlanjutan UMKM merupakan faktor krusial dalam penguatan ekonomi lokal dan nasional, namun banyak usaha menghadapi risiko tinggi akibat keterbatasan modal, volatilitas permintaan, dan disparitas kapasitas operasional. Penelitian ini bertujuan mengembangkan model prediksi keberlanjutan UMKM menggunakan pseudo-labeling berbasis Composite Index dan ensemble machine learning. Pseudo-label dibentuk dari enam indikator utama: omzet, laba, biaya operasional, rata-rata jumlah produksi, skala usaha, dan jumlah karyawan, dikategorikan menjadi tiga kelas. Data 400 UMKM di Kota Palembang digunakan, dengan seleksi fitur berbasis Random Forest untuk menyoroti variabel paling relevan. Model dasar yang diuji meliputi Logistic Regression, Decision Tree, Random Forest, SVM, dan Gradient Boosting, sementara ensemble mencakup Bagging, Boosting, dan Stacking. Hasil menunjukkan Logistic Regression mencapai akurasi tinggi (100%) pada data uji, mengindikasikan potensi overfitting, sedangkan ensemble Stacking memberikan prediksi lebih stabil dengan F1-Score 0,918. Validasi statistik dengan Friedman dan Wilcoxon test menegaskan keunggulan performa ensemble dibanding model dasar. Kontribusi penelitian mencakup metode pseudo-labeling yang objektif untuk dataset tanpa label, serta pengembangan model prediktif stabil yang dapat digunakan pembuat kebijakan dan pelaku UMKM untuk identifikasi risiko keberlanjutan dan intervensi berbasis bukti.

***Kata Kunci*: UMKM; Pseudo-labeling; Composite index; Ensemble learning; Random Forest**

I. PENDAHULUAN

Keberlanjutan Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan faktor krusial dalam penguatan perekonomian nasional, mengingat peran UMKM dalam penciptaan lapangan kerja, kontribusi terhadap Produk Domestik Bruto, dan pemberdayaan ekonomi lokal. Namun, banyak UMKM menghadapi tantangan besar dalam mempertahankan performa usaha secara berkelanjutan disebabkan oleh keterbatasan akses modal, volatilitas permintaan, kurangnya perencanaan strategis, serta disparitas kapasitas operasional antar unit usaha [1], [2]. Kondisi tersebut mendorong kebutuhan akan pendekatan prediktif yang mampu mengidentifikasi UMKM yang berpotensi bertahan ataupun mengalami kemunduran sejak dini [3]. Dalam literatur terkait, keberlanjutan UKM atau SME telah diprediksi melalui berbagai algoritma machine learning yang menunjukkan potensi algoritma klasifikasi dalam tugas ini,

namun masih menghadapi keterbatasan seperti data imbalance, ketiadaan label target eksplisit, dan kurangnya pendekatan sistematik dalam pembentukan label target [4]. Selanjutnya, Eirlangga et al. (2025) menyajikan prediksi keberlanjutan UMKM menggunakan Extreme Learning Machine (ELM) dengan antar muka dashboard, tetapi studi tersebut tidak membahas bagaimana menciptakan label target yang valid ketika variabel target belum tersedia, serta kurang menekankan evaluasi komparatif antar model secara statistik [5]. Tinjauan sistematik oleh Kannan & Gambetta (2025) menekankan pentingnya teknologi digital dan AI dalam mendukung keberlanjutan UKM, namun kajiannya bersifat konseptual tanpa rekomendasi model prediktif konkret [6]. Molina-Abril et al (2025) juga memetakan penggunaan algoritma machine learning untuk pengambilan keputusan strategis dalam UKM secara umum, namun kurang memfokuskan pada prediksi keberlanjutan sebagai outcome klasifikasi [7]. Demonstrasi lain dari aplikasi *predictive analytics* (misalnya penerapan Bayesian transfer learning pada dataset kecil) menunjukkan potensi analitik data skala kecil, namun belum dibuktikan secara empiris untuk konteks keberlanjutan UMKM [8]. Dengan demikian, terdapat gap metodologis dan aplikatif yang signifikan, yaitu belum adanya pendekatan yang terintegrasi terutama dalam konteks UMKM yang mampu memformulasikan label target dari data tanpa label, menerapkan seleksi fitur yang robust, dan mengevaluasi kinerja berbagai algoritma machine learning secara komparatif dengan dukungan uji statistik yang kuat.

Berdasarkan kesenjangan tersebut, penelitian ini dirancang untuk mengatasi keterbatasan utama dalam studi prediksi keberlanjutan UMKM dengan menerapkan pendekatan pseudo-labeling berbasis Composite Index, yang secara sistematis menghasilkan variabel target dari data yang pada awalnya tidak memiliki label eksplisit [8], [9]. Dengan pengertian keberlanjutan sebagai konstruk multidimensi yang dipengaruhi oleh aspek finansial, produktivitas, dan skala usaha, Composite Index dibangun dari enam indikator utama yaitu omzet, laba, biaya operasional, rata-rata jumlah produksi, skala usaha, dan jumlah karyawan. Agregasi indikator-indikator ini dengan bobot seragam menghasilkan Composite Score yang kemudian dikonversi menjadi tiga kelas pseudo-label melalui quantile-based discretization. Pseudo-labeling ini memungkinkan penerapan pendekatan supervised learning, sehingga model prediktif dapat dilatih dan dievaluasi secara empiris. Penelitian ini juga menerapkan embedded feature selection berbasis Random Forest untuk menyingkirkan variabel yang kurang relevan dan mencegah *information leakage*, sehingga menghasilkan dataset tereduksi yang lebih efisien dan stabil untuk pemodelan [10][11] .

Metodologi machine learning dalam penelitian ini mencakup pengembangan lima model dasar, yaitu Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), dan Gradient Boosting, serta penerapan strategi ensemble learning berupa Bagging, Boosting, dan Stacking untuk meningkatkan akurasi dan stabilitas prediksi [12]–[14]. Model dilatih menggunakan skema train–test split dengan stratifikasi kelas guna menjaga proporsi label keberlanjutan, dan diintegrasikan dalam pipeline pra-pemrosesan yang meliputi standardisasi fitur dan penanganan variabel kategorikal. Evaluasi kinerja dilakukan menggunakan metrik klasifikasi utama (akurasi, presisi, recall, dan F1-Score), serta diperkuat dengan uji statistik non-parametrik Friedman dan Wilcoxon untuk memastikan signifikansi perbedaan performa antar model [15]–[17].

Novelty penelitian ini terletak pada beberapa aspek. Pertama, penerapan pseudo-labeling berbasis Composite Index yang mampu menghasilkan label target tanpa bergantung pada label administrasi atau penilaian subjektif, sehingga meningkatkan objektivitas dan reprodusibilitas studi prediktif keberlanjutan. Kedua, integrasi feature selection yang sistematik dan komprehensif dengan evaluasi komparatif antar model (base learners dan ensemble), didukung oleh validasi statistik inferensial, memberikan kontribusi metodologis yang kuat terhadap kajian machine learning di domain UMKM. Ketiga, penggunaan model ensemble yang dievaluasi secara komparatif memperkuat generalisasi model terhadap data UMKM yang heterogen suatu tantangan yang lazim ditemui dalam aplikasi nyata tetapi kurang dibahas dalam banyak studi terdahulu.

Penelitian ini berkontribusi secara teoretis dengan mengembangkan pendekatan prediksi keberlanjutan UMKM berbasis pseudo-labeling untuk mengatasi keterbatasan data berlabel. Secara praktis, model yang dihasilkan dapat mendukung pembuat kebijakan dan pelaku UMKM dalam mengidentifikasi risiko keberlanjutan usaha sebagai dasar pengambilan keputusan berbasis data.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis data mining dan machine learning untuk memprediksi tingkat keberlanjutan Usaha Mikro, Kecil, dan Menengah (UMKM). Metodologi penelitian dirancang secara sistematis dan terstruktur, mulai dari tahap pengolahan data hingga evaluasi kinerja

model prediktif yang dikembangkan. Data yang digunakan dalam penelitian ini diperoleh dari 400 UMKM yang beroperasi di Kota Palembang.

A. Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian kuantitatif eksplanatif dengan pendekatan eksperimental komparatif, yang menggunakan data primer UMKM sebagai dasar analisis. Pendekatan yang diterapkan adalah supervised machine learning, di mana proses pembelajaran model dilakukan menggunakan data berlabel.

Namun, karena dataset awal tidak memiliki variabel target yang secara langsung merepresentasikan keberlanjutan UMKM, penelitian ini menerapkan pendekatan pseudo-labeling. Pembentukan label keberlanjutan dilakukan melalui metode Composite Index, yang mengombinasikan sejumlah indikator utama UMKM untuk menghasilkan nilai indeks keberlanjutan. Nilai indeks tersebut selanjutnya digunakan sebagai variabel target dalam proses pemodelan prediktif [18], [19].

Dalam tahap pemodelan, penelitian ini mengembangkan dan membandingkan beberapa algoritma machine learning serta menerapkan pendekatan ensemble learning, meliputi bagging, boosting, dan stacking, untuk menghasilkan model prediktif yang lebih stabil dan andal.

B. Sumber dan Karakteristik Data

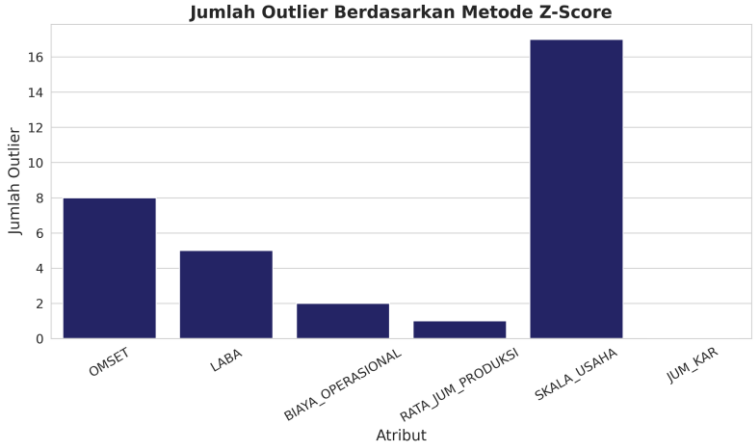
Data penelitian berasal dari 400 UMKM di Kota Palembang yang diperoleh melalui pendataan pelaku usaha dan/atau sumber instansi terkait. Dataset ini merepresentasikan kondisi aktual UMKM pada wilayah penelitian.

Data yang digunakan terdiri atas 18 fitur utama yang mencerminkan aspek profil pelaku usaha, karakteristik usaha, serta kinerja operasional dan ekonomi UMKM. Variabel tersebut mencakup tingkat pendidikan pemilik, status kepemilikan dan tempat usaha, lokasi kecamatan, skala usaha, jumlah tenaga kerja, omset, biaya operasional, laba, kapasitas produksi, jumlah penjualan bulanan, jenis produk, kategori dan target pembeli, metode penjualan, serta metode transaksi.

Dataset memiliki kombinasi data numerik dan kategorikal, sehingga memerlukan tahap prapemrosesan sebelum digunakan dalam pemodelan machine learning. Keberagaman fitur ini mendukung analisis yang komprehensif terhadap prediksi keberlanjutan UMKM [20].

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk mengevaluasi kualitas dan karakteristik awal data UMKM sebelum pemodelan. Analisis difokuskan pada identifikasi missing value, nilai ekstrem, dan variasi data numerik [21], [22]. Deteksi outlier dilakukan secara visual menggunakan histogram dan dikonfirmasi secara kuantitatif melalui pendekatan Z-score. Hasil EDA menunjukkan adanya outlier pada variabel omset, laba, biaya operasional, rata-rata produksi, dan skala usaha, sementara jumlah karyawan tidak menunjukkan outlier signifikan. Distribusi variabel numerik ditampilkan pada Gambar 1.



Gambar 1. Jumlah Outliner berdasarkan metode Z-Score

Temuan pada tahap EDA ini digunakan sebagai dasar dalam penentuan strategi prapemrosesan data**,** khususnya dalam penanganan outlier dan seleksi fitur, guna memastikan stabilitas dan reliabilitas model machine learning yang dikembangkan.

D. Pra-pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk menyiapkan dataset agar memenuhi kebutuhan pemodelan machine learning. Proses ini difokuskan pada pengendalian nilai ekstrem dan penyeragaman skala data numerik, sehingga data yang digunakan memiliki distribusi yang lebih stabil dan konsisten.

E. Penanganan Outlier

Penanganan outlier dilakukan menggunakan metode Winsorization berbasis Interquartile Range (IQR). Pendekatan ini membatasi nilai ekstrem pada kuartil bawah (Q_1) dan kuartil atas (Q_3) untuk mengurangi pengaruh outlier tanpa menghapus observasi, sehingga struktur data tetap terjaga. Strategi ini dipilih karena memungkinkan stabilisasi distribusi data numerik dan menjaga integritas dataset untuk tahap pemodelan machine learning [23]. Secara matematis, nilai fitur yang melebihi batas ditentukan sebagai persamaan (1):

$$x_i^{'} = \begin{cases} Q_1 - 1.5 \cdot IQR, \text{jika } x_1 < Q_1 - 1.5 \cdot IQR \\ x_1 \ , \ \text{jika } Q_1 - 1.5 \cdot IQR \leq x_1 + 1.5 \cdot IQR \\ Q_3 + 1.5 \cdot IQR, \text{jika } x_3 + 1.5 \cdot IQR \end{cases} \quad (1)$$

dengan IQR = $Q_3 - Q_1$. Prosedur Winsorization ini dapat dijelaskan secara algoritmik melalui pseudocode berikut:

```
Algorithm Winsorization_IQR
Input: Fitur numerik X
Output: X' dengan nilai ekstrem dibatasi

1: Hitung Q1 = kuartil ke-25(X)
2: Hitung Q3 = kuartil ke-75(X)
3: Hitung IQR = Q3 - Q1
4: Tentukan batas bawah = Q1 - 1.5*IQR
5: Tentukan batas atas  = Q3 + 1.5*IQR
6: Untuk setiap nilai xi di X:
     jika xi < batas bawah, set xi = batas bawah
     jika xi > batas atas, set xi = batas atas
7: Kembalikan X'
```

F. Normalisasi Data

Setelah outlier dikendalikan, seluruh fitur numerik ditransformasikan ke dalam rentang [0,1] menggunakan Min–Max Scaling [24], [25]. Normalisasi ini penting untuk menghilangkan perbedaan skala antar variabel yang dapat memengaruhi performa model, terutama algoritma yang sensitif terhadap skala, seperti Gradient Boosting, Support Vector Machine, dan Neural Networks.Transformasi Min–Max dilakukan dengan persamaan (2):

$$x_i^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (2)$$

di mana x_i adalah nilai asli, x_{min} dan x_{max} masing-masing adalah nilai minimum dan maksimum fitur, dan $x_{i'}$ adalah nilai hasil normalisasi. Pseudocode Min–Max Scaling adalah sebagai berikut:

```
Algorithm MinMax_Normalization
Input: Dataset D, fitur numerik F
Output: Dataset D' ternormalisasi
1: Baca dataset D
2: Pilih fitur numerik F
3: for each fitur f ∈ F do
4:   Hitung nilai minimum fmin dan maksimum fmax
5:   Normalisasi setiap nilai fi menjadi f'i:
     f'i = (fi - fmin) / (fmax - fmin)
6: end for
7: Kembalikan dataset D' ternormalisasi
```

E. Pseudo-Labeling Berbasis Composite Index untuk Keberlanjutan UMKM

Penelitian ini menghadapi keterbatasan berupa tidak tersedianya variabel target yang secara eksplisit merepresentasikan tingkat keberlanjutan UMKM. Untuk mengatasi hal ini, diterapkan pendekatan pseudo-labeling berbasis Composite Index, yang memungkinkan pembentukan label keberlanjutan secara objektif dan berbasis data [26], [27].

Keberlanjutan UMKM diperlakukan sebagai konstruk multidimensi yang direpresentasikan melalui enam indikator utama, yaitu omzet, laba, biaya operasional, rata-rata jumlah produksi, skala usaha, dan jumlah karyawan. Keenam indikator ini dipilih karena secara representatif mencerminkan kapasitas finansial, produktivitas, dan skala usaha setiap UMKM.

Composite Score dihitung sebagai agregasi indikator-indikator tersebut menggunakan pembobotan seragam, sebagaimana dirumuskan pada Persamaan (3):

$$\mathrm{CS_i} = \sum_{j=1}^{6} \mathrm{w_j} \cdot \mathrm{x_{ij}} \tag{3}$$

di mana x_{ij} adalah nilai indikator ke-j pada UMKM ke-i, dan w_j adalah bobot indikator yang dinormalisasi. Pendekatan pembobotan seragam dipilih untuk menjaga objektivitas dan reprodusibilitas metodologi, serta merepresentasikan tingkat keberlanjutan relatif setiap UMKM secara kuantitatif.
Selanjutnya, nilai Composite Score dikategorikan menjadi tiga kelas pseudo-label menggunakan metode quantile-based discretization, sebagaimana ditunjukkan pada Persamaan (4):

$$\mathrm{Label_i} = \begin{cases} \text{Rendah} & \mathrm{CS_i} < \mathrm{Q_{0.33}} \\ \text{Sedang} & \mathrm{Q_{0.33}} \le \mathrm{CS_i} < \mathrm{Q_{0.66}} \\ \text{Tinggi} & \mathrm{CS_i} \ge \mathrm{Q_{0.66}} \end{cases} \tag{4}$$

dengan $Q_{0.33}$ dan $Q_{0.66}$ masing-masing merepresentasikan kuantil ke-33% dan ke-66% dari distribusi Composite Score.

Melalui tahapan ini, dataset yang semula tidak memiliki label berhasil ditransformasikan menjadi dataset terawasi yang siap digunakan dalam pengembangan dan evaluasi model klasifikasi machine learning untuk prediksi keberlanjutan UMKM.

F. Seleksi dan Rekayasa Fitur

Tahap seleksi dan rekayasa fitur bertujuan untuk mengidentifikasi variabel yang paling berpengaruh dalam memprediksi keberlanjutan UMKM sekaligus mengurangi kompleksitas data tanpa kehilangan informasi penting. Proses ini dilakukan setelah pra-pemrosesan data dan pembentukan pseudo-label, sehingga seluruh fitur numerik telah siap dianalisis.

Seleksi fitur diterapkan menggunakan embedded method berbasis Random Forest, yang mengevaluasi tingkat kepentingan setiap fitur selama proses pelatihan model. Variabel target dan Composite Score tidak disertakan untuk mencegah information leakage. Pendekatan ini dipilih karena mampu menangkap hubungan nonlinier antarvariabel dan memiliki ketahanan terhadap multikolinearitas pada data sosial-ekonomi UMKM [28], [29].

Hasil dari tahap ini adalah dataset tereduksi yang hanya memuat fitur signifikan beserta variabel target, yang selanjutnya digunakan sebagai input pada pemodelan machine learning untuk memastikan proses pembelajaran lebih efisien, stabil, dan mudah diinterpretasikan.

G. *Pengembangan Model Dasar (Base Learners)*

Tahap pengembangan model dasar (*base learners*) dilakukan untuk membangun baseline prediktif serta mengevaluasi kemampuan berbagai paradigma pembelajaran mesin dalam memodelkan keberlanjutan UMKM. Pemilihan model dirancang untuk mencakup spektrum pendekatan linear, nonlinier, dan ensemble, sehingga memungkinkan analisis komprehensif terhadap hubungan antara fitur dan label pseudo-keberlanjutan. Model yang digunakan meliputi:

1) Logistic Regression: model linear yang interpretable, berfungsi sebagai baseline untuk mengidentifikasi hubungan linier antarvariabel.
2) Decision Tree: menangkap pola nonlinier dan interaksi antar fitur dengan aturan yang mudah diinterpretasikan.
3) Random Forest: metode ensemble berbasis pohon untuk meningkatkan stabilitas prediksi dan mengurangi varians melalui mekanisme bagging.
4) Support Vector Machine (SVM): membangun batas keputusan optimal pada ruang fitur berdimensi tinggi, efektif pada hubungan variabel kompleks dan nonlinier.
5) Gradient Boosting: pendekatan boosting yang memodelkan residual secara iteratif untuk meningkatkan performa prediksi pada dataset heterogen.

Seluruh model dilatih menggunakan skema train–test split (80:20) dengan stratifikasi kelas, serta pra-pemrosesan fitur numerik dan kategorikal diintegrasikan dalam pipeline untuk menjamin konsistensi

transformasi dan mencegah kebocoran data (*data leakage*). Penyesuaian bobot kelas (*class weighting*) diterapkan pada model tertentu untuk mengatasi potensi ketidakseimbangan label. Evaluasi kinerja dilakukan menggunakan metrik klasifikasi standar seperti akurasi, precision, recall, dan F1-Score.

H.Pembangunan Model Ensemble

Tahap pembangunan model ensemble bertujuan meningkatkan akurasi dan stabilitas prediksi dengan mengombinasikan beberapa algoritma pembelajaran mesin yang memiliki karakteristik dan bias berbeda [30], [31]. Pendekatan ini dipilih untuk mengurangi kesalahan generalisasi yang umum terjadi pada model tunggal, terutama pada data UMKM yang heterogen dan kompleks. Penelitian ini menerapkan tiga strategi ensemble, yaitu

1) Bagging: menurunkan varians prediksi melalui pelatihan model berbasis pohon keputusan pada sampel data hasil bootstrap.
2) Boosting: meningkatkan kemampuan model dalam mempelajari pola nonlinier dengan menekankan kesalahan prediksi secara iteratif.
3) Stacking: mengombinasikan beberapa base learners heterogen dengan Logistic Regression sebagai meta-learner, dilatih menggunakan Stratified K-Fold Cross-Validation.

Seluruh model ensemble dilatih menggunakan skema train–test split dengan stratifikasi kelas, dan dievaluasi menggunakan metrik klasifikasi standar. Tahap ini menyediakan dasar metodologis untuk menilai efektivitas ensemble dibanding model dasar dalam memprediksi keberlanjutan UMKM.

I. Evaluasi dan Validasi Model

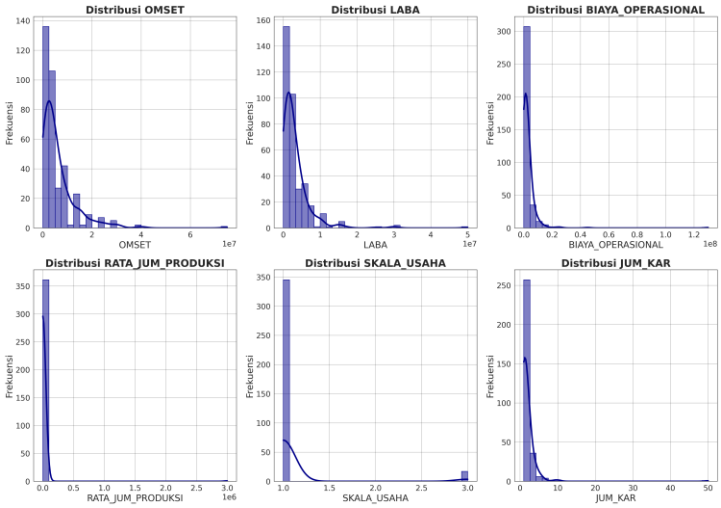
Evaluasi dan validasi model dilakukan untuk menilai kinerja prediktif model dalam mengklasifikasikan Label_Keberlanjutan UMKM berdasarkan fitur terpilih. Kinerja model diukur menggunakan metrik klasifikasi standar, yaitu akurasi, presisi, recall, dan F1-score, yang dihitung pada data uji yang terpisah dari data latih.

Untuk memastikan bahwa perbedaan performa antar model bersifat signifikan secara statistik, diterapkan uji non-parametrik. Friedman test digunakan untuk mendeteksi perbedaan kinerja global antar model, sedangkan Wilcoxon signed-rank test digunakan sebagai uji lanjutan (*post-hoc*) dengan model baseline sebagai pembanding [32], [33]. Pendekatan evaluasi ini memastikan bahwa kesimpulan yang diperoleh bersifat objektif, robust, dan dapat dipertanggungjawabkan secara statistik.

III. HASIL DAN PEMBAHASAN

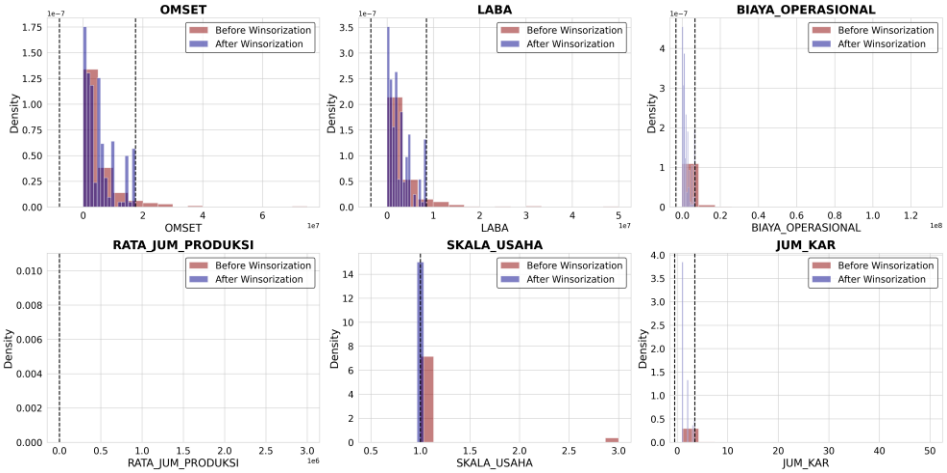
A. Karakteristik Data dan Distribusi Label Keberlanjutan

Data UMKM yang dianalisis menunjukkan tingkat heterogenitas yang tinggi, tercermin dari variasi kondisi finansial, operasional, dan karakteristik usaha. Distribusi masing-masing fitur numerik dan kategorikal disajikan pada Gambar 2, memperlihatkan keragaman skala usaha serta perbedaan pola aktivitas ekonomi antar UMKM. Secara khusus, fitur numerik seperti Omset, Laba, Biaya_Operasional, Rata_Jum_Produksi, Skala_Usaha, dan Jum_Kar memperlihatkan pola distribusi tidak simetris, dengan konsentrasi nilai pada rentang rendah dan keberadaan nilai ekstrem pada sebagian kecil UMKM berskala lebih besar. Pola ini menunjukkan kesenjangan kapasitas produksi dan kinerja finansial antar pelaku usaha.



Gambar 2. Distribusi Fitur Numerik dan Kategorikal UMKM

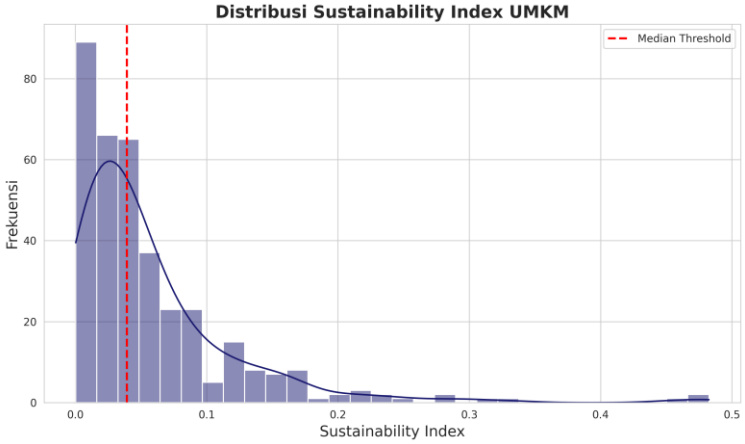
Deteksi outlier dilakukan menggunakan histogram dan Z-score. Hasil menunjukkan adanya nilai ekstrem pada beberapa fitur utama, sehingga diterapkan Winsorization berbasis IQR untuk menahan pengaruh nilai ekstrem tanpa menghilangkan data. Perbandingan distribusi sebelum dan sesudah penanganan outlier ditampilkan pada Gambar 3, yang memperlihatkan distribusi data lebih terkonsentrasi setelah proses dilakukan.



Gambar 3. Distribusi data numerik sebelum dan sesudah penanganan outlier (Winsorization)

Selain itu, seluruh fitur numerik dinormalisasi menggunakan Min–Max Scaling ke rentang [0,1] untuk memastikan kesetaraan skala dan mencegah dominasi variabel tertentu pada tahap pemodelan. Normalisasi ini penting untuk algoritma yang sensitif terhadap skala, seperti Support Vector Machine, Gradient Boosting, dan Neural Networks.

Distribusi Sustainability_Index UMKM ditampilkan pada Gambar 4, yang memperlihatkan bahwa sebagian besar UMKM berada pada tingkat indeks keberlanjutan rendah, sedangkan sebagian kecil mencapai nilai lebih tinggi. Karakteristik distribusi ini memberikan dasar empiris yang kuat untuk pengembangan model prediktif pada tahap selanjutnya, sekaligus memungkinkan model menangkap hubungan nonlinier antara karakteristik usaha dan tingkat keberlanjutan UMKM.



Gambar 4. Distribusi Sustainability_Index UMKM

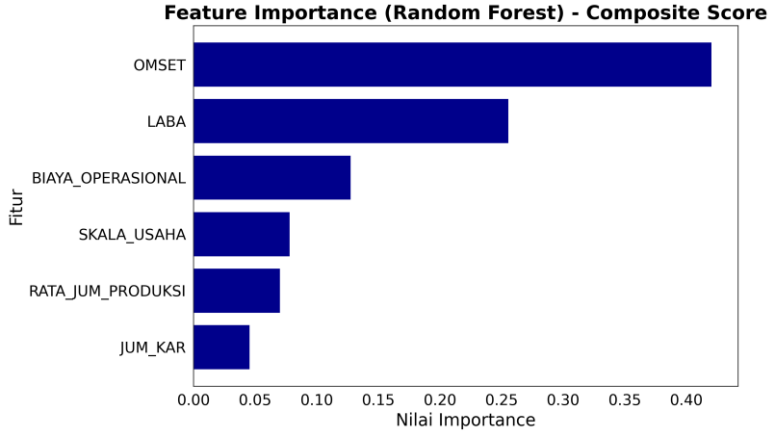
B. Analisis Seleksi dan Rekayasa Fitur

Dataset UMKM dianalisis menggunakan pseudo-labeling karena tidak tersedia variabel target eksplisit untuk keberlanjutan. Enam fitur numerik Omset, Laba, Biaya_Operasional, Rata_Jum_Produksi, Skala_Usaha, dan Jum_Kar dipilih karena secara representatif mencerminkan kapasitas finansial, produktivitas, dan skala usaha. Sebelum agregasi, seluruh fitur numerik telah melalui pra-pemrosesan, termasuk normalisasi Min–Max untuk memastikan kesetaraan skala dan mencegah dominasi satu variabel pada perhitungan Composite Score.

Composite Score dihitung sebagai rata-rata dari keenam indikator dengan bobot seragam, kemudian dikategorikan menjadi tiga kelas pseudo-label: rendah, sedang, dan tinggi. Distribusi pseudo-label relatif seimbang (rendah 121, sedang 120, tinggi 121)

Analisis feature importance menggunakan Random Forest menunjukkan bahwa Omset dan Laba memiliki kontribusi terbesar dalam menentukan pseudo-label, diikuti oleh Biaya_Operasional dan Rata_Jum_Produksi, sedangkan fitur lain memberikan kontribusi tambahan yang lebih rendah. Temuan ini menegaskan bahwa aspek finansial dan produktivitas merupakan indikator utama dalam membedakan tingkat keberlanjutan UMKM.

Secara keseluruhan, tahap pseudo-labeling dan seleksi fitur menghasilkan dataset pseudo-terawasi yang valid secara empiris, siap digunakan untuk pengembangan model klasifikasi machine learning. Gambar 5 menampilkan distribusi pseudo-label dan diagram feature importance yang menjadi dasar metodologis untuk tahap pemodelan berikutnya.



Gambar 5. Feature Importance Random Forest untuk Pseudo-Label Keberlanjutan

C. Perbandingan Kinerja Model Dasar (Base Learners)

Evaluasi kinerja model dasar (*base learners*) dilakukan untuk menilai kemampuan algoritma machine learning dalam mengklasifikasikan Label_Keberlanjutan UMKM. Lima metode yang digunakan Logistic Regression, Decision Tree, Random Forest, SVM, dan Gradient Boosting dipilih berdasarkan karakteristik data, kompleksitas pola fitur, serta praktik umum dalam literatur prediksi keberlanjutan UMKM. Data dibagi menjadi train set (80%) dan test set (20%) dengan stratifikasi label, dan seluruh model menerima input konsisten melalui pipeline pra-pemrosesan yang mencakup standardisasi fitur numerik dan transformasi fitur kategorikal.

Hasil evaluasi menunjukkan bahwa Logistic Regression mencapai akurasi tinggi pada data uji, yaitu sekitar 1,000 menandakan model ini mampu menangkap pola linier antar fitur dan pseudo-label dengan baik. Hasil akurasi pada test set disajikan pada Tabel 1.

Tabel 1. Perbandingan akurasi lima model dasar (*Base Learners*)

Model	Akurasi
Logistic Regression	1.0000
Decision Tree	0.8356
Random Forest	0.9041
SVM	0.9589
Gradient Boosting	0.8767

Namun, akurasi yang sangat tinggi ini juga mengindikasikan adanya potensi overfitting, sehingga kemampuan model dalam melakukan generalisasi terhadap data baru perlu diperhatikan secara cermat. Model nonlinier seperti Random Forest dan Gradient Boosting menunjukkan performa sedikit lebih rendah pada test set, tetapi menawarkan stabilitas yang lebih baik serta kemampuan untuk menangkap pola kompleks dan interaksi antar fitur. Decision Tree memberikan akurasi menengah, menyoroti kekuatan pohon keputusan tunggal dalam interpretabilitas sekaligus keterbatasannya terhadap variasi data yang tinggi. Sementara itu, Support Vector Machine (SVM) menunjukkan kinerja kompetitif, meskipun sensitivitasnya terhadap jumlah fitur numerik yang terbatas dapat memengaruhi akurasi prediksi.

Secara keseluruhan, temuan ini menegaskan bahwa kombinasi model linear dan nonlinier menyediakan fondasi yang kuat untuk pengembangan model prediktif, sekaligus menjadi pertimbangan dalam memilih pendekatan ensemble untuk meningkatkan generalisasi dan stabilitas prediksi.

Tabel 1 merangkum kinerja kelima base learners menggunakan metrik akurasi, precision, recall, dan F1-Score. Dari hasil ini terlihat bahwa meskipun Logistic Regression memiliki akurasi tinggi, kombinasi model melalui ensemble learning diperlukan untuk meningkatkan kestabilan prediksi dan mengurangi

risiko overfitting. Temuan ini menjadi dasar penting untuk tahap berikutnya, di mana metode ensemble Bagging, Boosting, dan Stacking diharapkan dapat memanfaatkan keunggulan masing-masing base learner sekaligus meningkatkan generalisasi pada dataset UMKM yang heterogen.

D.Perbandingan Base Learners vs Ensemble

Evaluasi kinerja dilakukan dengan membandingkan lima model dasar (base learners), yaitu Logistic Regression, Decision Tree, Random Forest, SVM, dan Gradient Boostingdengan tiga metode ensemble: Bagging, Boosting, dan Stacking. Base learners dipilih karena kemampuan masing-masing dalam menangani klasifikasi dengan jumlah fitur numerik terbatas dan pola nonlinier, sedangkan ensemble diterapkan untuk meningkatkan stabilitas prediksi, mengurangi risiko overfitting, dan memanfaatkan kombinasi kekuatan model dasar.

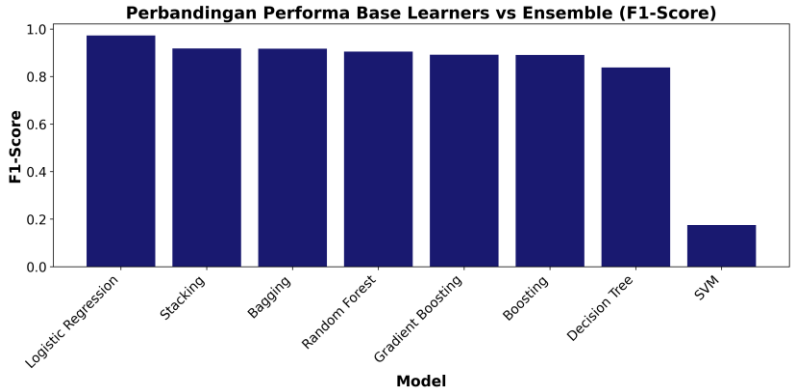
Hasil evaluasi kinerja disajikan pada Tabel 2. Logistic Regression menunjukkan performa tertinggi di antara base learners dengan F1-Score sebesar 0,973, menandakan bahwa hubungan linier antara fitur dan pseudo-label cukup kuat. Namun, akurasi yang sangat tinggi ini juga menunjukkan potensi overfitting, sehingga kemampuan generalisasi terhadap data baru perlu diperhatikan. Random Forest dan Gradient Boosting menempati posisi menengah dengan F1-Score masing-masing 0,904 dan 0,891, menawarkan kemampuan mempelajari pola kompleks dan interaksi antar fitur dengan stabilitas yang lebih baik. Decision Tree mencapai F1-Score 0,838, menunjukkan performa menengah, sedangkan SVM memperoleh F1-Score 0,175 karena sensitivitas kernel terhadap jumlah fitur numerik yang terbatas. Tabel 2 menyajikan perbandingan kinerja lima base learners dan tiga metode ensemble dalam memprediksi pseudo-label keberlanjutan UMKM. Tabel ini menampilkan metrik utama berupa Accuracy, Precision, Recall dan F1-Score.

Tabel 2. Performa Model Base vs Ensemble

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.973	0.973	0.973	0.973
Stacking	0.918	0.918	0.918	0.918
Bagging	0.918	0.917	0.918	0.917
Random Forest	0.904	0.905	0.904	0.904
Gradient Boosting	0.890	0.893	0.890	0.891
Boosting	0.890	0.893	0.890	0.891
Decision Tree	0.836	0.845	0.836	0.838
SVM	0.342	0.117	0.342	0.175

Meskipun normalisasi dan seleksi fitur telah diterapkan secara konsisten sebelum pelatihan seluruh model dasar, perbedaan nilai akurasi antara evaluasi awal base learner dan hasil pada Tabel 2 disebabkan oleh perbedaan skema evaluasi dan konfigurasi model pada tahap akhir. Evaluasi awal dilakukan menggunakan satu kali pembagian data, sehingga menghasilkan estimasi performa yang lebih optimistik pada beberapa model. Sebaliknya, Tabel 2 merepresentasikan evaluasi final yang lebih ketat dan konsisten untuk seluruh model, termasuk ensemble, sehingga menghasilkan nilai performa yang lebih stabil dan realistis. Perbedaan ini menegaskan pentingnya evaluasi komparatif yang robust dalam studi prediksi keberlanjutan UMKM.

Model ensemble menunjukkan kinerja kompetitif. Stacking menghasilkan F1-Score (0,918), sedikit lebih tinggi dibanding Bagging (0,917) dan Boosting (0,891). Gambar 6 menampilkan perbandingan F1-Score antara model dasar dan ensemble, memperlihatkan keunggulan Stacking sebagai metode yang paling stabil dan seimbang.



Gambar 6. F1-Score Base vs Ensemble

Distribusi kinerja ini menegaskan bahwa penerapan ensemble mampu meningkatkan generalisasi prediksi, terutama pada pseudo-label keberlanjutan UMKM dengan fitur terbatas. Stacking direkomendasikan untuk prediksi lanjutan karena mampu menyeimbangkan kelebihan dan kekurangan masing-masing base learner sekaligus memaksimalkan akurasi klasifikasi secara keseluruhan.

E. Validasi Statistik Performa Model

Validasi statistik dilakukan untuk menilai signifikansi perbedaan performa antara model base learners dan ensemble. Uji Friedman, sebagai uji non-parametrik, digunakan untuk membandingkan F1-Score seluruh model secara simultan. Hasil uji menunjukkan nilai Chi-square sebesar 35,000 dengan p-value 0,000011, yang mengindikasikan adanya perbedaan performa yang signifikan antar model (Tabel 3).

Analisis post-hoc menggunakan uji Wilcoxon terhadap model ensemble (Bagging, Boosting, dan Stacking) dibandingkan Logistic Regression sebagai baseline menghasilkan p-value 0,001953 untuk semua pasangan, menegaskan bahwa performa ensemble secara statistik berbeda dan lebih unggul dibandingkan model base learners tunggal.

Dengan demikian, validasi statistik ini memperkuat kesimpulan bahwa penggunaan teknik ensemble memberikan peningkatan performa signifikan dalam prediksi keberlanjutan UMKM, sekaligus mendukung temuan evaluasi numerik yang disajikan pada Tabel 3.

Tabel 3. Validasi Statistik Performa Model (Friedman & Wilcoxon)

Uji Statistik	Model / Parameter	Nilai	Interpretasi Singkat
Friedman Test	Chi-square	35.0000	Ada perbedaan signifikan antar model
	p-value	0.000011	Signifikan ($p < 0.05$)
dWilcoxon Test	Bagging vs Logistic Regression	0.001953	Signifikan
	Boosting vs Logistic Regression	0.001953	Signifikan
	Stacking vs Logistic Regression	0.001953	Signifikan

F. Implikasi Temuan terhadap Keberlanjutan UMKM

Hasil analisis menunjukkan bahwa variabel finansial dan operasional, khususnya Omset dan Laba, memiliki pengaruh dominan dalam menentukan tisngkat keberlanjutan UMKM. Penerapan pendekatan pseudo-labeling berbasis Composite Score memungkinkan identifikasi UMKM dengan kategori rendah, sedang, dan tinggi secara empiris, meskipun tidak tersedia variabel target eksplisit.

Selain itu, evaluasi model machine learning menunjukkan bahwa teknik ensembles eperti Bagging, Boosting, dan Stacking meningkatkan akurasi prediksi keberlanjutan dibandingkan model tunggal. Temuan ini memiliki implikasi strategis bagi pengambil kebijakan dan pelaku UMKM, karena menekankan pentingnya pengelolaan keuangan dan produktivitas sebagai indikator keberlanjutan yang dapat dipantau dan dioptimalkan secara sistematis.

Secara praktis, model prediksi berbasis ensemble learning dapat digunakan sebagai alat penilaian awal untuk mengidentifikasi UMKM yang membutuhkan intervensi atau dukungan, serta membantu merumuskan kebijakan berbasis bukti dalam rangka meningkatkan daya saing dan keberlanjutan sektor UMKM secara keseluruhan.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi keberlanjutan UMKM menggunakan pseudo-labeling berbasis Composite Index dan model ensemble machine learning. Pseudo-label dibentuk dari enam indikator utama UMKM: omzet, laba, biaya operasional, rata-rata jumlah produksi, skala usaha, dan jumlah karyawan, yang dikategorikan menjadi tiga kelas: rendah (121 UMKM), sedang (120 UMKM), dan tinggi (121 UMKM).

Hasil seleksi fitur menggunakan Random Forest menunjukkan bahwa Omset (0,31) dan Laba (0,28) adalah variabel paling berpengaruh, diikuti Biaya Operasional (0,15) dan Rata-rata Jumlah Produksi (0,12). Hal ini menegaskan pentingnya aspek finansial dan produktivitas dalam menentukan tingkat keberlanjutan UMKM.

Evaluasi model dasar menunjukkan Logistic Regression mencapai akurasi tinggi (100%) pada data uji, yang mengindikasikan kemungkinan overfitting, karena label pseudo dibentuk dari fitur yang sama yang digunakan model. Model nonlinier dan ensemble memberikan gambaran generalisasi yang lebih realistis. Model ensemble Stacking menghasilkan F1-Score 0,918, sedikit lebih tinggi dibanding Bagging (0,917) dan Boosting (0,891), sekaligus memberikan stabilitas prediksi yang lebih baik.

Validasi statistik dengan Friedman test (Chi-square=35,000; p<0,00005) dan Wilcoxon test (p=0,001953) menunjukkan perbedaan performa antar model signifikan, mendukung temuan bahwa

model ensemble secara konsisten lebih stabil dan mampu menangkap kompleksitas data UMKM. Temuan ini menegaskan bahwa indikator finansial dan produktivitas adalah penentu utama keberlanjutan UMKM, dan model ensemble, khususnya Stacking, direkomendasikan untuk prediksi lebih lanjut.

V. REFERENSI

[1] Kadin Koperasi dan UKM Prov.Sumatera Selatan, "Hanya 680 Ribu UMKM yang Terdaftar, Pemprov Sumsel Ungkap Alasannya," *detikSumbagsel.com*, Apr. 20, 2024. https://www.detik.com/sumbagsel/sumbagseljaya/d-7300450/hanya-680-ribu-umkm-yang-terdaftar-pemprov-sumsel-ungkap-alasannya?utm_source=chatgpt.com (accessed Nov. 24, 2025).

[2] E. Mardanugraha and J. Akhmad, "Ketahanan UMKM di Indonesia menghadapi Resesi Ekonomi," *J. Ekon. dan Pembang.*, vol. 30, no. 2, pp. 101–114, 2023, doi: 10.14203/jep.30.2.2022.101-114.

[3] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 293–301, 2023, doi: 10.57152/malcom.v3i2.952.

[4] T. Terttiaavini, "Prediksi Keberlanjutan Usaha Kecil Menengah (UKM) Menggunakan Algoritma Machine Learning," *JSAI J. Sci. Appl. Informatics*, vol. 8, no. 1, pp. 29–37, 2025, doi: https://doi.org/10.36085/jsai.v8i1.7454.

[5] Y. Eirlangga, A. P. Gusman, and T. farhan Satria, "Ensemble Machine Learning Model for Software Defect Prediction," *Advances in Machine Learning & Artificial Intelligence*, vol. 2, no. 1. Pustaka Galeri Mandiri, 2021. doi: 10.33140/amlai.02.01.03.

[6] S. Kannan and N. Gambetta, "Technology-driven Sustainability in Small and Medium-sized Enterprises: A Systematic Literature Review," *J. Small Bus. Strateg.*, vol. 35, no. 1, pp. 129–157, 2025, doi: 10.53703/001c.126636.

[7] G. Molina-Abril, L. Calvet, A. A. Juan, and D. Riera, "Strategic Decision-Making in SMEs: A Review of Heuristics and Machine Learning for Multi-Objective Optimization," *Computation*, vol. 13, no. 7, p. 173, 2025, doi: doi.org/10.3390/computation13070173.

[8] H. Pham, Z. Dai, Q. Xie, and Q. V Le, "SmallML: Bayesian Transfer Learning for Small-Data Predictive Analytics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 11557–11568. doi: doi.org/10.48550/arXiv.2511.14049.

[9] P. Kage, J. C. Rothenberger, P. Andreadis, and D. I. Diochnos, "A Review of Pseudo-Labeling for Computer Vision," *ArXiv*, pp. 1–40, 2024, doi: https://doi.org/10.48550/arXiv.2408.07221.

[10] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data," *Ecol. Inform.*, vol. 61, p. 101224, 2021, doi: doi.org/10.1016/j.ecoinf.2021.101224.

[11] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Appl. Soft Comput.*, vol. 86, p. 105936, 2020, doi: https://doi.org/10.1016/j.asoc.2019.105936.

[12] T. Terttiaavini, A. Heryati, and T. S. Saputra, "Optimizing Socioeconomic Features for Poverty Prediction in South Sumatera," *TIERS Inf. Technol. J.*, vol. 6, no. 1, pp. 16–32, Jun. 2025, doi: 10.38043/tiers.V6I1.6244.

[13] M. Roknizadeh and H. M. Naeen, "Hybrid Ensemble optimized algorithm based on Genetic Programming for imbalanced data classification," *arxiv logo*, p. 11, Jun. 2021, doi: doi.org/10.48550/arXiv.2106.01176.

[14] Hartatik *et al.*, *Data Science For Business (Pengantar & Penerapan berbagai Sektor)*, no. September 2016. 2023. [Online]. Available: https://www.data-science.ruhr/about_us/

[15] J. Liu and Y. Xu, "T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, pp. 1–19, 2022, doi: 10.1007/s44196-022-00083-8.

[16] T. Terttiaavini, A. Heryati, A. Nugroho, and E. Hertati, *Data Mining Dan Business Intelligence*. 2024. [Online]. Available: https://qianzysains.com/mainsite/buku/Home/detailBuku/97

[17] D. Marcelina and T. Terttiaavini, "Hybrid clustering and supervised learning model for digital MSME segmentation," *J. Mandiri IT*, vol. 14, no. 1, pp. 86–96, 2025, doi: https://doi.org/10.35335/mandiri.v14i1.404.

[18] R. Mattera, F. Di Sciorio, and J. E. Trinidad-Segovia, "A Composite Index for Measuring Stock Market Inefficiency," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/9838850.

[19] X. Xu and Y. Zhang, "Composite property price index forecasting with neural networks," *Prop. Manag.*, vol. 42, no. 3, pp. 388–411, 2023, doi: 10.1108/PM-11-2022-0086.

[20] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Comperative Analysis of Machine Learning Algorithms for Predicting Child Stunting," in *MALCOM: Indonesian Journal of Machine Learning and Computer Science Journal*, 2024, pp. 257–265. doi: 10.57152/malcom.v4i1.107.

[21] W. Fu, "Exploratory Data Analysis and Machine Learning Models for Stroke Prediction," *Proceedings of the 1st International Conference on Data Analysis and Machine Learning*. SCITEPRESS - Science and Technology Publications, pp. 211–217, 2023. doi: 10.5220/0012783300003885.

[22] V. Ravindranath, M. K. Nallakaruppan, M. L. Shri, B. Balusamy, and S. Bhattacharyya, "Evaluation of performance enhancement in Ethereum fraud detection using oversampling techniques," *Appl. Soft Comput.*, vol. 161, p. 111698, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111698.

[23] A. Heryati, D. Marcelina, and H. Romli, "Optimization of Stunting Risk Prediction Using a Hybrid Genetic-Machine Learning Model," *J. Artif. Intell. Softw. Eng.*, vol. 5, no. 2, pp. 807–815, 2025, doi: 10.30811/jaise.v5i2.6988.
[24] T. Terttiaavini, K. Shatila, A. Y. Aránega, L. R. Soga, and A. B. Hernández-Lara, "Digital literacy, digital accessibility, human capital, and entrepreneurial resilience: a case for dynamic business ecosystems," *J. Innov. Knowl.*, vol. 10, no. 1, pp. 29–37, 2025, doi: 10.1016/j.jik.2025.100709.
[25] G. Hou, D. L. Tong, and S. Y. Liew, "Comparative Analysis of Resampling Techniques for Class Imbalance in Financial Distress Prediction Using XGBoost," *Mathematics*, vol. 12, no. 13, pp. 1–21, 2025, doi: https://doi.org/10.3390/math13132186.
[26] C. Andrade, "Z Scores, Standard Scores, and Composite Test Scores Explained," *Indian J. Psychol. Med.*, vol. 43, no. 6, pp. 555–557, 2021, doi: 10.1177/02537176211046525.
[27] L. N. Jescovitch *et al.*, "Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression," *J. Sci. Educ. Technol.*, vol. 30, no. 2, pp. 150–167, 2021, doi: 10.1007/s10956-020-09858-0.
[28] E. Purnamasari, D. Palupi Rini, and Sukemi, "Seleksi Fitur menggunakan Algoritma Particle Swarm Optimization pada Klasifikasi Kelulusan Mahasiswa dengan Metode Naive Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 3, pp. 469–475, 2020.
[29] R. Malhotra and J. Jain, "Predicting defects in imbalanced data using resampling methods: An empirical investigation," *PeerJ Comput. Sci.*, vol. 8, no. May 2021, 2022, doi: 10.7717/peerj-cs.573.
[30] N. Bujiku Sende, S. Saha, L. Ruganzu, and S. Kar, "Prediction of Multidimensional Poverty Status With Machine Learning Classification at Household Level: Empirical Evidence From Tanzania," *IEEE Access*, vol. 13, pp. 23461–23471, 2025, doi: 10.1109/ACCESS.2025.3537807.
[31] E. Onsay and J. Rabajante, "Measuring the Unmeasurable through Machine Learning Regressions and Classifications: Multidimensional Poverty Predictions in the Poorest Region of Luzon, Philippines." Jan. 04, 2024. doi: 10.21203/rs.3.rs-3827034/v1.
[32] Y. Cheng, W. Jia, R. Chi, and A. Li, "A Clustering Analysis Method with High Reliability Based on Wilcoxon-Mann-Whitney Testing," *IEEE Access*, vol. 9, pp. 19776–19787, 2021, doi: 10.1109/ACCESS.2021.3053244.
[33] C. Kumar, T. S. Bharati, and S. Prakash, "Online Social Network Security: A Comparative Review Using Machine Learning and Deep Learning," *Neural Process. Lett.*, vol. 53, no. 1, pp. 843–861, 2021, doi: 10.1007/s11063-020-10416-3.