JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 29~37
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Prediksi Keberlanjutan Usaha Kecil Menengah (UKM) Menggunakan Algoritma Machine Learning

Terttiaavini
Universitas Indo Global Mandiri
avini.saputra@uigm.ac.id

Article Info

Article history:

Received, 2024-12-09
Revised, 2024-12-12
Accepted, 2024-12-26

Kata Kunci:

Keberlanjutan UKM
Machine Learning
Klasterisasi
Klasifikasi
SMOTE (*Synthetic Minority Over-sampling Technique*)
5-Fold Cross-Validation

ABSTRAK

Usaha Kecil Menengah (UKM) berkontribusi sekitar 60% terhadap Produk Domestik Bruto (PDB) Indonesia dan menyerap lebih dari 97% tenaga kerja. Namun, UKM menghadapi tantangan yang menghambat keberlanjutan, seperti keterbatasan modal dan ketidakstabilan pasar. Penelitian ini bertujuan mengembangkan model prediksi untuk memetakan keberlanjutan UKM berdasarkan variabel-variabel yang mempengaruhi kelangsungan usaha. Metode yang digunakan mencakup klasterisasi dengan Agglomerative Clustering, K-Means, dan DBSCAN, serta klasifikasi menggunakan algoritma seperti Logistic Regression, Random Forest, dan XGBoost. Hasil menunjukkan bahwa metode Agglomerative Clustering memberikan kinerja terbaik dengan Silhouette Score 0.68. Semua model klasifikasi awalnya mencapai akurasi 1.0 dengan standard deviation 0.0, namun menunjukkan indikasi overfitting akibat ketidakseimbangan kelas antara kategori "Berlanjut" dan "Tidak Berlanjut". Untuk mengatasi masalah ini, penerapan metode SMOTE (*Synthetic Minority Over-sampling Technique*) dan 5-Fold Cross-Validation dilakukan. Hasilnya menunjukkan peningkatan kemampuan model dalam mengenali pola pada kelas minoritas, sehingga akurasi model menjadi lebih representatif terhadap kedua kelas. Penelitian ini diharapkan memberikan wawasan bagi Dinas Koperasi dan UKM Kota Palembang untuk mendukung keberlanjutan sektor UKM di Palembang.

Keywords:

SME Sustainability
Machine Learning
Clustering
Classification
SMOTE (Synthetic Minority Over-sampling Technique)
5-Fold Cross-Validation

ABSTRACT

Small and Medium Enterprises (SMEs) contribute approximately 60% to Indonesia's Gross Domestic Product (GDP) and absorb more than 97% of the workforce. However, SMEs face various challenges that hinder sustainability, such as limited capital and market instability. This study aims to develop a predictive model to map the sustainability of SMEs based on variables that influence business continuity. The methods used include clustering with Agglomerative Clustering, K-Means, and DBSCAN, as well as classification using algorithms such as Logistic Regression, Random Forest, and XGBoost. The results show that the Agglomerative Clustering method provides the best performance with a Silhouette Score of 0.68. All classification models initially achieved an accuracy of 1.0 with a standard deviation of 0.0, but indicated overfitting due to class imbalance between the "Continues" and "Does Not Continue" categories, where the minority class consists of only 16 data points. To address this issue, the application of the SMOTE (Synthetic Minority Over-sampling Technique) method and 5-Fold Cross-Validation was implemented. The results showed an improvement in the model's ability to recognize patterns in the minority class, making the model's accuracy more representative of both classes. This research is expected to provide valuable insights for the Office of Cooperatives and SMEs in Palembang to support the sustainability of the SME sector in Palembang.

Penulis Korespondensi:

Terttiaavini,
Program Studi Magister Ilmu Komputer,
Universitas Indo Global Mandiri,
Email: avini.saputra@uigm.ac.id

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 29~37
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

1. PENDAHULUAN

Usaha Kecil Menengah (UKM) memainkan peranan yang sangat penting dalam perekonomian Indonesia [1]. Berdasarkan data dari Kementerian Koperasi dan Usaha Kecil Menengah, UKM berkontribusi sekitar 60% terhadap Produk Domestik Bruto (PDB) Indonesia dan menyerap lebih dari 97% tenaga kerja [1] . Oleh karena itu, keberlanjutan dan perkembangan UKM sangat memengaruhi stabilitas ekonomi negara. Namun, meskipun memiliki kontribusi yang signifikan, UKM menghadapi berbagai tantangan yang dapat menghambat keberlanjutannya.

Fenomena ketidakberlanjutan UKM di Indonesia dipengaruhi oleh berbagai faktor internal dan eksternal. Faktor internal meliputi keterbatasan modal, rendahnya efisiensi operasional, kurangnya inovasi dalam produk dan pemasaran, serta masalah manajerial seperti pengelolaan sumber daya manusia yang tidak optimal. Di sisi lain, faktor eksternal seperti ketidakstabilan pasar, perubahan regulasi pemerintah, dan persaingan dengan perusahaan besar yang memiliki sumber daya lebih banyak juga berkontribusi pada kegagalan UKM. Hal ini menciptakan tantangan besar bagi banyak UKM untuk bertahan dalam jangka Panjang [2] .

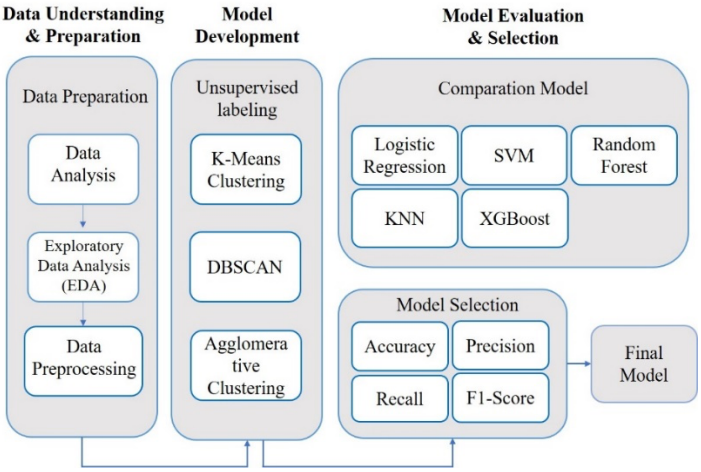
Permasalahan utama yang dihadapi oleh UKM adalah bagaimana memastikan kelangsungan usaha mereka dalam jangka panjang. Untuk itu, diperlukan pendekatan yang dapat membantu pelaku UKM dalam menganalisis dan memprediksi faktor-faktor yang dapat memengaruhi keberlanjutan usaha mereka. Penelitian ini bertujuan untuk mengembangkan model prediksi yang dapat memetakan keberlanjutan UKM berdasarkan variabel-variabel tertentu yang mempengaruhi kelangsungan usaha tersebut .

Model prediksi ini diharapkan dapat memberikan panduan kepada pemilik UKM dalam mengambil keputusan strategis yang dapat memperpanjang umur usaha mereka. Dengan menggunakan pendekatan analisis data berbasis machine learning, penelitian ini akan menganalisis berbagai faktor relevan seperti omset, laba, biaya operasional, jumlah karyawan, jumlah produksi, dan skala usaha. Metode yang digunakan mencakup klasterisasi dengan Agglomerative Clustering [3], [4], K-Means [5], dan DBSCAN [6], [7], [8]., serta klasifikasi menggunakan algoritma seperti Logistic Regression [9] [10], Random Forest, dan XGBoost [11]

Dengan demikian, penelitian ini diharapkan dapat memberikan wawasan mendalam bagi pemilik UKM untuk mengambil langkah-langkah strategis yang mendukung kelangsungan dan perkembangan usaha mereka. Hasil penelitian ini juga dapat menjadi dasar bagi Dinas Koperasi dan UKM Kota Palembang dalam merumuskan kebijakan yang lebih efektif untuk mendukung keberlanjutan sektor ini

2. METODE PENELITIAN

Metodologi penelitian merupakan landasan penting dalam mencapai tujuan yang telah ditetapkan, serta memberikan struktur yang jelas untuk proses analisis data. Dalam konteks penelitian ini, metodologi yang digunakan dirancang untuk mengatasi tantangan ketidakberlanjutan Usaha Kecil Menengah (UKM) di Kota Palembang dengan pendekatan berbasis machine learning. Proses penelitian ini mengikuti tahapan yang sistematis, seperti yang ditunjukkan dalam gambar 1 berikut ini.



Gambar 1. Diagram Alir Model Prediksi keberlanjutan UKM

Tahap Pemahaman dan Persiapan Data (Data Understanding & Preparation)

Tahapan ini merupakan langkah krusial dalam siklus pembangunan model *machine learning*. Seluruh proses ini bertujuan untuk memastikan bahwa data memiliki kualitas yang memadai dan siap digunakan dalam analisis serta pembuatan model prediktif. Seluruh proses ini dilaksanakan menggunakan Python di Google Colab. Tahap ini mencakup empat langkah utama yang dilaksanakan secara berurutan, yaitu

1. **Data Analysis**

Pada tahap ini, karakteristik dataset dianalisis untuk memberikan pemahaman awal. Dilakukan identifikasi terhadap jumlah sampel, atribut yang tersedia, serta tipe data yang digunakan, baik numerik maupun kategorik [12]. Data yang hilang atau tidak konsisten juga diperiksa untuk memastikan kelengkapan dataset. Dalam penelitian ini, atribut yang dipilih meliputi Omset, Laba, Biaya_Operasional, Rata_Jum_Produksi, Skala_Usaha dan Jum_Kar. Hasil Deskripsi Statistik data tersbeut di tampilan pada tabel 1.

Pada tahap ini, dilakukan analisis deskriptif terhadap dataset untuk memperoleh pemahaman awal. Analisis meliputi identifikasi jumlah sampel, tipe data (numerik dan kategorik), serta pemeriksaan terhadap keberadaan missing values dan data duplikat. Variabel yang digunakan dalam penelitian ini adalah Omset, Laba, Biaya_Operasional, Rata_Jum_Produksi, Skala_Usaha, dan Jum_Kar. Hasil analisis deskriptif disajikan pada Tabel 1."

Tabel 1. Deskrisi statistic data UKM

Atribut	Mean	Median	Min	Max	Std
Omset	6,139,268.00	4,000,000	40,000	75,000,000	7,515,395
Laba	3,144,558.00	2,000,000	0	5,000,000	4,440,955
Biaya_Operasional	2,859,655.00	1,000,000	0	130,000,000	7,762,152
Rata_Jum_Produksi	9,050.36	30	1	3,000,000	157,728
Skala_Usaha	1.09	1	1	3	0.42
Jum_Kar	1.93	1	1	50	3.05

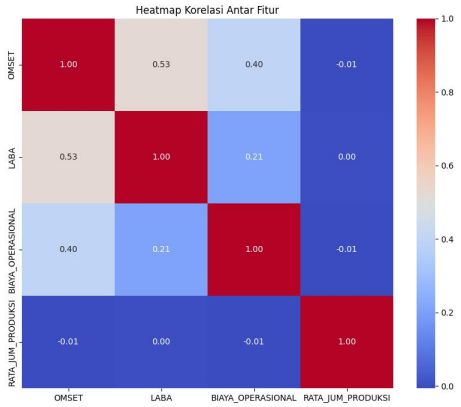
Secara keseluruhan, dataset ini mencakup berbagai jenis UMK dengan ukuran dan performa yang bervariasi, baik dalam hal omset, biaya, laba, maupun produksi, yaitu

1) Omset**:** Terdapat variasi yang besar dalam omset usaha, dengan sebagian besar UKM menghasilkan pendapatan jauh lebih rendah daripada yang tertinggi (sekitar 75 juta). Ini menunjukkan adanya perbedaan signifikan antara usaha besar dan kecil.
2) Laba: Sebagian besar UKM tidak menghasilkan laba besar, bahkan beberapa di antaranya tidak menghasilkan laba sama sekali. Namun, ada juga yang mencatatkan laba sangat tinggi (hingga 50 juta).
3) Biaya Operasional: Ada UKM yang memiliki biaya operasional rendah atau bahkan nol, namun ada pula yang mengeluarkan biaya operasional yang sangat besar, menunjukkan variasi dalam struktur biaya antar usaha.
4) Rata-Jum-Produksi: Rata-rata jumlah produksi cukup tinggi, namun banyak UKM yang memproduksi dalam jumlah yang lebih sedikit. Hal ini bisa menunjukkan perbedaan skala produksi antar UKM.
5) Skala Usaha: Mayoritas UKM di dataset ini beroperasi dalam skala kecil, dengan nilai 1 menunjukkan usaha kecil. Variasi dalam skala usaha tidak terlalu besar.
6) Jum_Kar: Kebanyakan UKM memiliki sedikit karyawan, dengan mayoritas beroperasi dengan hanya satu atau dua karyawan. Namun, ada beberapa yang memiliki jumlah karyawan yang jauh lebih banyak.

***Exploratory Data Analysis* (EDA)**

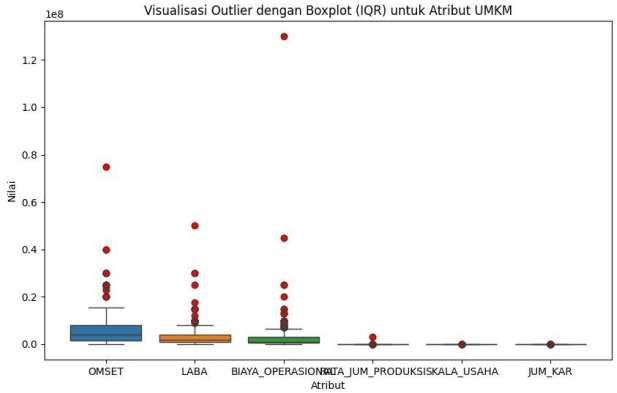
Pada tahap *Exploratory Data Analysis* (EDA), hubungan antar variabel dalam dataset dianalisis secara mendalam untuk mengidentifikasi pola, tren, atau anomali yang dapat memengaruhi hasil analisis lebih lanjut. Salah satu langkah penting dalam EDA adalah analisis korelasi, yang dilakukan untuk mengukur kekuatan dan arah hubungan antara setiap pasangan fitur. Korelasi ini dapat membantu memahami sejauh mana perubahan dalam satu variabel memengaruhi variabel lainnya, sehingga dapat digunakan untuk menentukan fitur-fitur yang relevan atau yang memiliki potensi multikolinieritas dalam model prediktif [13].
Hasil analisis korelasi disajikan dalam bentuk heatmap, seperti yang ditunjukkan pada gambar 1. Heatmap ini menunjukkan bahwa terdapat korelasi positif yang kuat antara Omset dan Laba, yang mengindikasikan bahwa peningkatan Omset cenderung diikuti oleh peningkatan Laba. Temuan ini mendukung hipotesis bahwa Omset merupakan salah satu faktor utama yang mempengaruhi keberlanjutan UKM.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 29~37
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085



Gambar 3. Headmap Korelasi antar Fitur

Selain itu, deteksi outlier perlu dilakukan untuk mengidentifikasi data yang menyimpang secara signifikan dari pola umum. Deteksi outlier dilakukan menggunakan metode *Interquartile Range* (IQR). Gambar 2 menunjukkan hasil perhitungan IQR yang divisualisasikan dalam bentuk boxplot.



Gambar 2. Deteksi outlier menggunakan metode *Interquartile Range* (IQR) dengan *boxplot*.

Berdasarkan visualisasi boxplot, dapat disimpulkan bahwa terdapat ketidaksamaan yang signifikan dalam kinerja UKM yang diteliti. Terdapat beberapa UKM dengan kinerja yang sangat baik (terlihat dari outlier pada nilai omset dan laba yang tinggi), namun juga banyak UKM dengan kinerja yang relatif rendah. Adanya outlier ini mengindikasikan adanya variasi yang cukup besar dalam data. Selain itu, distribusi data untuk beberapa atribut menunjukkan adanya kemiringan, yang menandakan bahwa sebagian besar data terkonsentrasi pada nilai tertentu. Hal ini menyiratkan bahwa perlu dilakukan analisis lebih lanjut untuk memahami faktor-faktor yang menyebabkan perbedaan kinerja antar UKM dan untuk mengembangkan strategi yang lebih tepat dalam mendukung pertumbuhan UKM.

Preprocessing data

Pada tahap *preprocessing data*, beberapa langkah penting dilakukan untuk mempersiapkan data agar siap digunakan dalam model *machine learning*. Langkah pertama adalah penanganan terhadap data yang hilang (*missing values*). Nilai yang hilang pada kolom Jum_Kar diimputasi menggunakan nilai minimum dari kolom tersebut.
Selanjutnya, duplikasi data diidentifikasi dan dihapus untuk menghindari adanya baris yang identik, yang dapat mempengaruhi kualitas model.
Terakhir, normalisasi dilakukan pada variabel numerik seperti Omset, Laba, dan Biaya_Operasional menggunakan metode *Min-Max Scaling* [14]. Metode ini mengubah nilai data ke dalam rentang [0, 1] dengan rumus:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

di mana X' adalah nilai yang dinormalisasi, X adalah nilai asli, $\min(X)$ adalah nilai minimum, dan $\max(X)$ adalah nilai maksimum dari fitur tersebut. Langkah ini memastikan bahwa semua fitur memiliki skala yang seragam, yang sangat penting agar model yang sensitif terhadap skala data dapat bekerja secara optimal.

Dengan data yang dinormalisasi, model dapat bekerja secara lebih efisien dan akurat karena fitur-fitur tersebut memiliki kontribusi yang setara dalam perhitungan. Dari hasil tahap processing jumlah dataset menjadi 345.

Model Development

Pada tahap model *development* dalam machine learning, tujuan utamanya adalah membangun dan melatih model yang dapat memprediksi atau mengelompokkan data berdasarkan pola yang ada. Salah satu aspek penting dalam tahap ini adalah pemilihan metode yang tepat untuk menangani dataset tanpa label, yang akan digunakan untuk melatih model. Salah satu pendekatan yang dapat digunakan untuk tujuan ini adalah unsupervised learning. Metode-metode yang digunakan dalam hal ini meliputi *K-Means Clustering* [15], DBSCAN [16], dan *Agglomerative Clustering*.
Proses klasterisasi bertujuan untuk mengelompokkan data ke dalam dua kategori, yaitu 'Berlanjut' dan 'Tidak Berlanjut', dengan menggunakan ketiga metode klasterisasi yang berbeda. Setiap metode ini menganalisis data tanpa label target untuk mengidentifikasi pola atau kelompok yang memiliki karakteristik serupa. Hasil dari klasterisasi ini kemudian dipetakan ke dalam dua kategori tersebut, yang memungkinkan pemahaman mengenai apakah data cenderung menunjukkan potensi untuk 'Berlanjut' atau 'Tidak Berlanjut'. Ketiga metode yang digunakan adalah *K-Means*, DBSCAN, dan *Agglomerative Clustering*. Adapun kode Python yang digunakan adalah sebagai berikut:

```
# 1.K-Means Clustering
kmeans = KMeans(n_clusters=2, random_state=42)
kmeans_labels = kmeans.fit_predict(X_scaled)
kmeans_mapped = ['Berlanjut' if label == 1 else 'Tidak Berlanjut' for
    label in kmeans_labels]
# 2.DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=3)
dbscan_labels = dbscan.fit_predict(X_scaled)
dbscan_mapped = ['Berlanjut' if label != -1 else 'Tidak Berlanjut' for
    label in dbscan_labels]
# 3.Agglomerative Clustering (2 klaster: Berlanjut dan Tidak Berlanjut)
agg_clustering = AgglomerativeClustering(n_clusters=2)
agg_labels = agg_clustering.fit_predict(X_scaled)
# Memetakan label klaster ke kategori 'Berlanjut' dan 'Tidak Berlanjut'
agg_mapped = ['Berlanjut' if label == 1 else 'Tidak Berlanjut' for label
    in agg_labels]
```

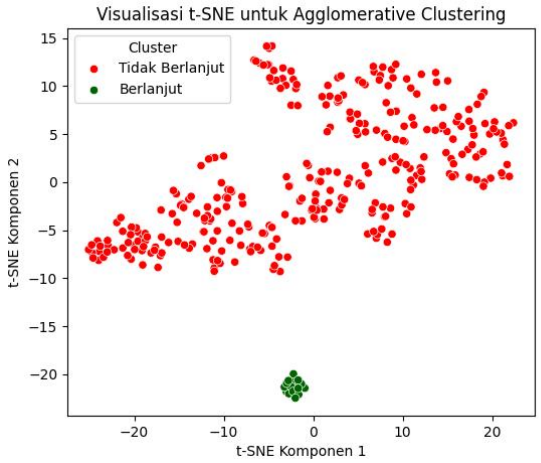
Hasil klasterisasi tersebut kemudian dievaluasi untuk menentukan model terbaik dengan menghitung beberapa metrik evaluasi, yaitu: 1) *Silhouette Score*, 2) *Davies-Bouldin Score*, dan 3) *Calinski-Harabasz Score* [17]. Metrik-metrik ini digunakan untuk menilai kualitas dan kesesuaian klaster yang dihasilkan oleh masing-masing model. Tabel 1 menyajikan hasil perhitungan metrik evaluasi tersebut.

Tabel 1. Matrik evaluasi model Klasterisasi

Metode	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
K-Means	0.61	1.21	80.75
DBSCAN	0.37	2.57	26.95
Agglomerative	0.68	0.41	69.53

Berdasarkan hasil evaluasi klasterisasi, metode *Agglomerative* menunjukkan kinerja terbaik dengan *Silhouette Score* tertinggi (0.68), *Davies-Bouldin Score* terendah (0.41), dan *Calinski-Harabasz Score* yang cukup baik (69.53), yang menunjukkan klaster yang jelas terpisah dan terstruktur. K-Means juga memberikan hasil yang baik, tetapi sedikit kurang optimal dibandingkan Agglomerative, dengan Silhouette Score 0.61 dan Davies-Bouldin Score 1.21. Sementara itu, DBSCAN menghasilkan nilai evaluasi yang lebih rendah, terutama dalam hal *Silhouette Score* (0.37) dan *Calinski-Harabasz Score* (26.95), yang menunjukkan bahwa klaster yang dihasilkan kurang terpisah dengan baik dan cenderung mengandung lebih banyak noise. Dengan demikian, Agglomerative merupakan metode klasterisasi yang paling efektif untuk dataset ini. Jumlah data berdasarkan hasil klasterisasi *Agglomerative Clustering* : tidak Berlanjut: 329 data, berlanjut: 16 data.
Hasil dari pemisahan klaster divisualisasikan menggunakan t-SNE Plot. *t-Distributed Stochastic Neighbor Embedding* (t-SNE). Plot ini memberikan gambaran yang lebih jelas tentang distribusi dan pemisahan antar klaster, serta bagaimana hubungan antar data dalam masing-masing klister [18]. Dengan menggunakan t-SNE, kita dapat mengamati apakah klaster-klaster yang dihasilkan memiliki pemisahan yang jelas atau ada tumpang tindih antara klaster, yang memberikan wawasan tambahan dalam memilih model klasterisasi yang paling efektif. Gambar 3 merupakan Visualisasi hasil klasterisasi menggunakan t-SNE Plot.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 29~37
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085



Gambar 3. Visualisasi hasil klasterisasi menggunakan t-SNE Plot

Comparation Model

Setelah melakukan klasterisasi menggunakan *Agglomerative Clustering* dan memperoleh hasil klaster terbaik, langkah berikutnya adalah mengevaluasi performa model klasifikasi dengan berbagai algoritma untuk menentukan metode paling efektif dalam memprediksi keberlanjutan UKM [19]. Tujuan evaluasi ini adalah membandingkan kinerja algoritma klasifikasi seperti *Logistic Regression, Decision Tree, Random Forest, K-Neighbors Classifier (KNN), Support Vector Classifier (SVC), dan XGBoost Classifier (XGBClassifier)* [20]. Masing-masing model dianalisis berdasarkan metrik evaluasi, yaitu akurasi, presisi, *recall*, dan *F1-score*, untuk menentukan model yang paling akurat dan efisien dalam memprediksi keberlanjutan usaha. Tabel 3 menyajikan perbandingan hasil evaluasi dari berbagai model klasifikasi yang telah digunakan.

Tabl 3. Hasil Evaluasi Model Klasifikasi

Model	Akurasi	Precision	Recall	F1-Score
Logistic Regression	1.0	1.0	1.0	1.0
Decision Tre	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0
KNN	1.0	1.0	1.0	1.0
SVM	1.0	1.0	1.0	1.0
XGBoos	1.0	1.0	1.0	1.0

Hasil matriks evaluasi ini menunjukkan bahwa semua model yang diuji—*Logistic Regression, Decision Tree, Random Forest, KNN, SVM, dan XGBoost*—memperoleh rata-rata akurasi 1.0 dengan standard deviation 0.0 saat diuji menggunakan validasi silang (cross-validation). Akurasi 1.0 pada semua model berarti bahwa setiap model berhasil memprediksi seluruh data dengan benar dalam setiap fold validasi silang, menunjukkan bahwa model sangat efektif dalam memisahkan data sesuai dengan kelas yang ada, baik kelas "Berlanjut" maupun "Tidak Berlanjut".

Hal ini dapat disebabkan oleh beberapa faktor, seperti model yang sangat cocok dengan data yang ada atau data yang digunakan terlalu sederhana dan homogen, sehingga model dapat dengan mudah mempelajari pola yang ada. Namun, model ini bisa terindikasi mengalami overfitting, di mana model hanya menghafal data pelatihan tanpa kemampuan generalisasi yang baik terhadap data baru. *Overfitting* ini lebih mungkin terjadi jika data pelatihan yang tersedia terbatas atau sangat mirip, terutama dengan adanya ketidakseimbangan kelas yang kuat antara kelas mayoritas dan kelas minoritas [21].

Standard deviation 0.0 menunjukkan bahwa hasil evaluasi akurasi sangat konsisten, yang mengindikasikan bahwa model tidak menunjukkan variasi dalam performanya. Faktor lain yang berkontribusi adalah ketidakseimbangan kelas, di mana model lebih cenderung memprediksi kelas mayoritas ("Tidak Berlanjut"), terutama karena kelas minoritas ("Berlanjut") hanya terdiri dari 16 data. Hal ini menyebabkan model mengingat pola dari kelas mayoritas dan menghasilkan akurasi yang tinggi, meskipun ini tidak mencerminkan kualitas model dalam memprediksi kelas minoritas.

Oleh karena itu, meskipun akurasi 1.0 tampak menggembirakan, hasil ini dapat menjadi indikasi bahwa model mengalami overfitting dan mungkin tidak akan bekerja dengan baik pada data baru. Untuk mengatasi permasalahan ini, beberapa langkah perbaikan perlu dilakukan, antara lain:

1) Penggunaan Teknik SMOTE
   Teknik SMOTE (*Synthetic Minority Over-sampling Technique*) yang diterapkan untuk mengatasi ketidakseimbangan kelas terbukti efektif dalam meningkatkan jumlah sampel pada kelas minoritas

[22]. Meskipun hasil evaluasi menunjukkan rata-rata akurasi tetap sempurna di angka 1.0 dengan standar deviasi 0.0, hal ini mengindikasikan bahwa model mampu melakukan klasifikasi dengan sangat baik, baik untuk kelas mayoritas (Tidak Berlanjut) maupun kelas minoritas (Berlanjut).

2) *5-fold cross-validation* (Validasi silang dengan 5 lipatan)
Evaluasi model juga dilakukan menggunakan *5-fold cross-validation* [23]. Hasil pengujian menunjukkan bahwa semua model yang diuji—*Logistic Regression, Decision Tree, Random Forest*, KNN, dan SVM—mencapai rata-rata akurasi 1.0 dengan standar deviasi 0.0. Hasil ini mengindikasikan bahwa model mampu memprediksi data dengan sangat baik, menghasilkan akurasi sempurna di setiap lipatan validasi.

3. HASIL DAN ANALISIS

Berdasarkan pengujian model yang telah dilakukan, dapat disimpulkan hasil penelitian ini sebagai berikut : Hasil dari tahap *model development* menunjukkan bahwa *Agglomerative Clustering* merupakan metode klasterisasi yang paling efektif dalam memisahkan data ke dalam dua kategori, yaitu 'Berlanjut' dan 'Tidak Berlanjut', dengan hasil evaluasi yang superior. *Metrik Silhouette Score* yang tinggi (0.68), *Davies-Bouldin Score* yang rendah (0.41), dan *Calinski-Harabasz Score* yang cukup baik (69.53) menunjukkan klaster yang terpisah dengan jelas dan terstruktur. Dibandingkan dengan K-Means dan DBSCAN, yang memberikan hasil lebih rendah dalam hal pemisahan klaster, *Agglomerative Clustering* menunjukkan kinerja yang lebih stabil dan konsisten. Visualisasi dengan t-SNE juga mendukung temuan ini, dengan menunjukkan pemisahan yang jelas antar klaster. Oleh karena itu, *Agglomerative Clustering* adalah metode klasterisasi yang paling cocok untuk dataset ini, menghasilkan pemisahan data yang lebih baik dan lebih terstruktur.

Penggunaan Teknik SMOTE:
Teknik SMOTE telah diterapkan untuk mengatasi ketidakseimbangan kelas dalam dataset. Setelah penerapan SMOTE, hasil evaluasi menunjukkan bahwa model mampu mengklasifikasikan kedua kelas (mayoritas dan minoritas) dengan sangat baik, meskipun standar deviasi yang sangat kecil menandakan bahwa model mungkin hanya lebih cenderung memprediksi kelas mayoritas. Oleh karena itu, meskipun SMOTE membantu memperbaiki keseimbangan kelas, masih ada kecenderungan untuk memprediksi kelas mayoritas lebih sering daripada kelas minoritas.

Validasi Silang (*5-Fold Cross-Validation*)
Hasil evaluasi model menggunakan *5-Fold Cross-Validation* menunjukkan bahwa semua model yang diuji mencapai akurasi rata-rata sebesar 1.0 dengan standar deviasi 0.0. Nilai akurasi yang sempurna pada setiap lipatan validasi ini mencerminkan kemampuan prediksi model yang sangat baik. Namun, akurasi sempurna tersebut juga dapat menjadi indikasi overfitting, terutama jika data yang digunakan terbatas dan kurang beragam. Ketidakseimbangan kelas dalam dataset dapat memengaruhi kemampuan model untuk mengenali pola pada kelas minoritas secara efektif. Standar deviasi yang bernilai 0.0 menunjukkan tingkat konsistensi yang sangat tinggi dalam hasil, tetapi juga mengindikasikan bahwa model mungkin terlalu terlatih pada data pelatihan, sehingga berpotensi kurang mampu menggeneralisasi ke dataset yang lebih luas. Oleh karena itu, diperlukan evaluasi lebih lanjut menggunakan dataset yang lebih besar dan lebih bervariasi untuk memastikan keandalan model

Dalam penelitian ini, meskipun hasil evaluasi model terlihat positif, analisis lebih lanjut mengungkapkan adanya ketidakseimbangan kelas yang signifikan. Dari total 345 UKM yang dianalisis, hanya 16 UKM (4,6%) yang dikategorikan sebagai "Berlanjut," sementara 329 UKM (95,4%) berada dalam kategori "Tidak Berlanjut." Ketidakseimbangan ini menunjukkan bahwa mayoritas UKM di Palembang masih menghadapi tantangan besar untuk mempertahankan keberlanjutan usaha mereka. Temuan ini menegaskan perlunya pembinaan yang lebih intensif dan strategi khusus untuk membantu UKM meningkatkan daya saing mereka di pasar yang semakin kompetitif.

4. KESIMPULAN

Melalui pendekatan berbasis machine learning, penelitian ini menggunakan beberapa metode klasterisasi dan klasifikasi untuk memetakan keberlanjutan UKM. Hasil analisis menunjukkan bahwa metode Agglomerative Clustering memberikan hasil terbaik dalam mengelompokkan data berdasarkan potensi keberlanjutan. Selain itu, semua model klasifikasi yang diuji menunjukkan akurasi 1.0, tetapi juga mengindikasikan potensi overfitting akibat ketidakseimbangan kelas antara kategori "Berlanjut" dan "Tidak Berlanjut". Dengan demikian, meskipun model menunjukkan performa yang sangat baik pada data pelatihan, penting untuk melakukan langkah-langkah perbaikan seperti meningkatkan jumlah data pelatihan dan menerapkan teknik

resampling untuk mengatasi masalah ketidakseimbangan kelas. Penelitian ini diharapkan dapat memberikan wawasan berharga bagi pemilik UKM dan pembuat kebijakan dalam merumuskan strategi yang mendukung keberlanjutan sektor UKM di Indonesia.

REFERENSI

[1] F. Baderi, "UMKM Pilar Pemulihan dan Pertumbuhan Ekonomi Nasional," Harian Ekonomi Neraca. Accessed: Dec. 07, 2024. [Online]. Available: https://www.neraca.co.id/article/209137/umkm-pilar-pemulihan-dan-pertumbuhan-ekonomi-nasional

[2] Terttiaavini, Sofyan, and T. S. Saputra, "Pendampingan Penyusunan Program Rencana Kerja Badan Usaha Milik Desa Dalam Rangka Optimalisasi Potensi Desa Serijabo Ogan Ilir Sumatera Selatan," *JMM (Jurnal Masyarakat Mandiri)*, vol. 5, no. 1, pp. 3536–3546, 2021.

[3] Y. Rong and Y. Liu, "Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, Jun. 2020, pp. 124–127. doi: 10.1109/ICAICA50127.2020.9182394.

[4] A. Chhabra and P. Mohapatra, "Fair Algorithms for Hierarchical Agglomerative Clustering," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2022, pp. 206–211. doi: 10.1109/ICMLA55696.2022.00036.

[5] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[6] N. F. Sahamony, T. Terttiaavini, and H. Rianto, "Analisis Perbandingan Kinerja Model Machine Learning untuk Memprediksi Risiko Stunting pada Pertumbuhan Anak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, 2024, doi: 10.57152/malcom.v4i2.1210.

[7] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simul Model Pract Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/J.SIMPAT.2015.03.003.

[8] G. Zotteri, M. Kalchschmidt, and F. Caniato, "The impact of aggregation level on forecasting performance," *Int J Prod Econ*, vol. 93–94, no. SPEC.ISS., pp. 479–491, Jan. 2005, doi: 10.1016/J.IJPE.2004.06.044.

[9] Y. Geng, Q. Li, G. Yang, and W. Qiu, "Logistic Regression," in *Practical Machine Learning Illustrated with KNIME*, Singapore: Springer Nature Singapore, 2024, pp. 99–132. doi: 10.1007/978-981-97-3954-7_4.

[10] K. Furmańczyk, K. Paczutkowski, M. Dudziński, and D. Dziewa-Dawidczyk, "Classification Methods Based on Fitting Logistic Regression to Positive and Unlabeled Data," in *Computational Science – ICCS 2022: 22nd International Conference, London, UK,* D. Groen, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., London: Springer Cham, Jun. 2022, pp. 31–45. doi: 10.1007/978-3-031-08751-6_3.

[11] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *International Journal of Information Technology*, vol. 13, no. 4, pp. 1503–1511, Aug. 2021, doi: 10.1007/s41870-020-00430-y.

[12] D. Antoni, T. Avini, and A. H. Heryati, *Business Process Reengineering*. 2024.

[13] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA: ACM, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[14] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[15] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 293–301, Nov. 2023, doi: 10.57152/malcom.v3i2.952.

[16] B. J. J. Kremers, A. Ho, J. Citrin, and K. L. van de Plassche, "Two step clustering for data reduction combining DBSCAN and k-means clustering," *Journal Metrics: Contributions to Plasma Physics*, Nov. 2021, doi: 10.1002/ctpp.202200177.

[17] T. Avini, Z. Alamin, G. Maulani, and E. B. Perkasa, *Fundamental Algorithma*. 2024.

[18] M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz, and D. K. Hartline, "t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis," *Mar Genomics*, vol. 51, p. 100723, Jun. 2020, doi: 10.1016/j.margen.2019.100723.

[19] A. Avram, O. Matei, C.-M. Pintea, P. C. Pop, and C. A. Anton, "Comparative Analysis of Clustering Techniques for a Hybrid Model Implementation," in *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)*, Á. Herrero, C. Cambra, D. Urda, J. JSedano, H. Quintián, and E. Corchado, Eds., Springer, Cham, 2021, pp. 22–32. doi: 10.1007/978-3-030-57802-2_3.

[20] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[21] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Overfitting, Model Tuning, and Evaluation of Prediction Performance*. Springer International Publishing, 2022. doi: 10.1007/978-3-030-89010-0.

[22] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[23] C. Mahlich, T. Vente, and J. Beel, "From Theory to Practice: Implementing and Evaluating e-Fold Cross-Validation," in *International Conference on Artificial Intelligence and Machine Learning Research (CAIMLR). 2024*, 2024. doi: https://doi.org/10.48550/arXiv.2410.09463.