

Klasifikasi Keberhasilan Usaha Mikro, Kecil, dan Menengah (UMKM) di Indonesia Menggunakan Algoritma C4.5 dengan Pendekatan Data Mining

Bunga Nurul Lestari^{a,1,*}, Lusiana Riska^{a,2}

^a Universitas Mahkota Tricom Unggul, Medan, Indonesia,

¹ bungatari2207@gmail.com*

Abstract

This study aims to analyze and classify the success of Micro, Small, and Medium Enterprises (MSMEs) using a data mining approach based on the entropy-driven Decision Tree (C4.5) algorithm, with the Random Forest algorithm employed as a comparative model. The primary objective of this research is to identify the factors influencing MSME success accurately while providing an interpretable classification model to support business decision-making. The dataset consists of MSME data that underwent a preprocessing stage using one-hot encoding to transform categorical variables into numerical representations. The research methodology includes data exploration, data preprocessing, train-test data partitioning, model development using the C4.5 algorithm, and model evaluation based on accuracy, confusion matrix, classification report, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), and k-fold cross-validation. In addition, feature importance analysis was conducted to identify the most influential factors affecting MSME success. The findings indicate that the C4.5 algorithm achieved competitive classification performance with stable predictive accuracy while offering a more interpretable model than the Random Forest algorithm. It can therefore be concluded that the C4.5 algorithm is an effective approach for classifying MSME success, as it provides an appropriate balance between predictive accuracy and model interpretability, thereby supporting data-driven decision-making in the MSME sector.

Keywords: UMKM; Machine Learning; Data Mining; Decision Tree; Random Forest

1. Pendahuluan

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan salah satu sektor yang memiliki peran strategis dalam perekonomian Indonesia. UMKM berkontribusi terhadap pertumbuhan ekonomi nasional, penciptaan lapangan kerja, pemerataan pendapatan masyarakat, serta peningkatan kesejahteraan. Selain itu, UMKM terbukti memiliki ketahanan yang cukup baik dalam menghadapi berbagai tantangan ekonomi, sehingga menjadi salah satu pilar utama pembangunan ekonomi nasional. Meskipun demikian, tidak semua UMKM mampu berkembang dan mencapai tingkat keberhasilan yang sama. Berbagai faktor seperti modal usaha, lama usaha, jumlah tenaga kerja, pemanfaatan teknologi, strategi pemasaran, lokasi usaha, serta kualitas sumber daya manusia menjadi faktor yang memengaruhi keberhasilan suatu UMKM.

Seiring dengan perkembangan teknologi informasi, jumlah data yang dihasilkan oleh aktivitas bisnis UMKM semakin meningkat. Data tersebut dapat dimanfaatkan sebagai sumber informasi untuk mendukung proses pengambilan keputusan apabila diolah menggunakan metode yang tepat. Salah satu pendekatan yang dapat digunakan adalah data mining, yaitu proses menggali pola, hubungan, dan pengetahuan yang tersembunyi dari kumpulan data dalam jumlah besar sehingga menghasilkan informasi yang bermanfaat bagi pengambilan keputusan.

Salah satu teknik dalam data mining yang banyak digunakan adalah klasifikasi. Teknik klasifikasi bertujuan membangun model yang mampu mengelompokkan data ke dalam kelas tertentu berdasarkan karakteristik yang dimiliki. Dalam konteks penelitian ini, klasifikasi digunakan untuk mengidentifikasi tingkat keberhasilan UMKM berdasarkan berbagai faktor yang memengaruhinya. Hasil klasifikasi diharapkan dapat membantu pemerintah, pelaku usaha, maupun lembaga pendukung UMKM dalam menyusun strategi pengembangan usaha yang lebih tepat sasaran.

Algoritma Decision Tree C4.5 merupakan salah satu metode klasifikasi yang banyak digunakan karena memiliki tingkat akurasi yang baik serta menghasilkan model yang mudah dipahami. Algoritma ini membangun pohon keputusan berdasarkan nilai Information Gain dan Gain Ratio sehingga mampu memilih atribut terbaik sebagai dasar pembentukan aturan klasifikasi. Keunggulan lainnya adalah kemampuannya dalam menangani atribut kategorikal maupun numerik, mengatasi data yang memiliki nilai hilang (missing value), serta menghasilkan aturan keputusan (decision rules) yang mudah diinterpretasikan oleh pengguna.

Sebagai pembanding, penelitian ini juga menggunakan algoritma Random Forest yang merupakan pengembangan dari Decision Tree melalui pembentukan banyak pohon keputusan (ensemble learning). Random Forest dikenal memiliki kemampuan prediksi yang tinggi dan lebih tahan terhadap overfitting. Namun, model yang dihasilkan cenderung lebih sulit diinterpretasikan dibandingkan dengan Decision Tree C4.5. Oleh karena itu, perbandingan kedua algoritma dilakukan untuk mengetahui model yang memberikan keseimbangan terbaik antara tingkat akurasi dan kemudahan interpretasi.

Beberapa penelitian terdahulu menunjukkan bahwa algoritma C4.5 mampu memberikan hasil klasifikasi yang baik pada berbagai bidang, seperti prediksi kelulusan mahasiswa, klasifikasi penyakit, analisis kredit, serta penilaian kinerja usaha. Sementara itu, Random Forest sering menghasilkan akurasi yang lebih tinggi, tetapi interpretasi modelnya lebih kompleks. Berdasarkan kondisi tersebut, penelitian ini bertujuan membangun model klasifikasi keberhasilan UMKM di Indonesia menggunakan algoritma C4.5 dengan pendekatan data mining serta membandingkan kinerjanya dengan Random Forest. Evaluasi model dilakukan menggunakan confusion matrix, accuracy, precision, recall, F1-score, ROC-AUC, dan k-fold cross-validation. Selain itu, analisis feature importance dilakukan untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap keberhasilan UMKM. Hasil penelitian diharapkan dapat menjadi dasar dalam mendukung pengambilan keputusan berbasis data bagi pelaku UMKM maupun pihak terkait dalam merumuskan strategi pengembangan usaha yang lebih efektif.

2. Tinjauan Pustaka

2.1. Data Mining

Data mining merupakan proses menggali informasi, pola, hubungan, dan pengetahuan yang tersembunyi dari kumpulan data dalam jumlah besar menggunakan berbagai teknik statistik, kecerdasan buatan, dan machine learning. Data mining bertujuan mengubah data mentah menjadi informasi yang bernilai sehingga dapat mendukung proses pengambilan keputusan secara lebih efektif. Menurut Han, Kamber, dan Pei (2012), data mining merupakan bagian dari proses Knowledge Discovery in Databases (KDD) yang berfungsi menemukan pola-pola yang bermanfaat dari sekumpulan data. Dalam bidang bisnis, data mining banyak dimanfaatkan untuk analisis pelanggan, prediksi penjualan, deteksi kecurangan, segmentasi pasar, serta evaluasi kinerja usaha.

2.2. Klasifikasi

Klasifikasi merupakan salah satu teknik utama dalam data mining yang digunakan untuk mengelompokkan data ke dalam kelas atau kategori tertentu berdasarkan karakteristik yang dimiliki. Proses klasifikasi dilakukan dengan membangun model menggunakan data yang telah diketahui kelasnya (training data), kemudian model tersebut digunakan untuk memprediksi kelas pada data baru. Menurut Witten, Frank, Hall,

dan Pal (2017), klasifikasi merupakan proses pembelajaran (supervised learning) yang menghasilkan model prediksi berdasarkan pola dari data historis. Dalam penelitian ini, teknik klasifikasi digunakan untuk mengelompokkan tingkat keberhasilan UMKM berdasarkan atribut-atribut yang memengaruhi perkembangan usaha.

2.3. Algoritma Decision Tree C4.5

Algoritma C4.5 merupakan salah satu metode klasifikasi yang dikembangkan oleh J. Ross Quinlan sebagai penyempurnaan dari algoritma ID3. Algoritma ini membangun pohon keputusan (Decision Tree) dengan memilih atribut terbaik berdasarkan nilai Information Gain dan Gain Ratio. Setiap simpul pada pohon merepresentasikan atribut, sedangkan setiap cabang menunjukkan nilai atribut tersebut hingga menghasilkan keputusan pada daun (leaf node). Menurut Quinlan (1993), C4.5 memiliki beberapa keunggulan, antara lain mampu menangani atribut numerik maupun kategorikal, mengatasi data yang memiliki nilai hilang (missing value), melakukan proses pruning untuk mengurangi overfitting, serta menghasilkan aturan keputusan (decision rules) yang mudah dipahami. Karakteristik tersebut menjadikan algoritma C4.5 banyak digunakan dalam berbagai penelitian klasifikasi karena memiliki tingkat interpretabilitas yang tinggi.

2.4. Algoritma Random Forest

Random Forest merupakan algoritma klasifikasi berbasis ensemble learning yang menggabungkan banyak pohon keputusan (Decision Tree) untuk menghasilkan prediksi yang lebih akurat dan stabil. Algoritma ini menggunakan teknik bootstrap aggregating (bagging) dan pemilihan atribut secara acak pada setiap pohon sehingga mampu mengurangi risiko overfitting. Menurut Breiman (2001), Random Forest memiliki performa prediksi yang tinggi pada berbagai jenis dataset serta mampu menangani data dengan jumlah atribut yang banyak. Selain itu, algoritma ini dapat digunakan untuk mengukur tingkat kepentingan setiap atribut (feature importance), sehingga dapat diketahui faktor-faktor yang paling berpengaruh terhadap keberhasilan klasifikasi.

2.5. Usaha Mikro, Kecil, dan Menengah (UMKM)

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan sektor usaha produktif yang memiliki kontribusi besar terhadap pertumbuhan ekonomi nasional, penciptaan lapangan kerja, serta peningkatan kesejahteraan masyarakat. Keberhasilan UMKM dipengaruhi oleh berbagai faktor, antara lain modal usaha, lama usaha, jumlah tenaga kerja, tingkat pendidikan pelaku usaha, pemanfaatan teknologi digital, strategi pemasaran, akses pembiayaan, serta kemampuan inovasi. Dengan banyaknya faktor yang memengaruhi keberhasilan tersebut, diperlukan pendekatan berbasis data untuk mengidentifikasi faktor-faktor dominan sehingga dapat membantu pelaku usaha maupun pemerintah dalam menyusun strategi pengembangan UMKM yang lebih efektif.

2.6. Penelitian Terdahulu

Berbagai penelitian menunjukkan bahwa algoritma C4.5 memiliki performa yang baik dalam menyelesaikan permasalahan klasifikasi. Algoritma ini telah banyak diterapkan pada klasifikasi kelayakan kredit, prediksi kelulusan mahasiswa, diagnosis penyakit, penentuan penerima bantuan sosial, serta evaluasi kinerja usaha. Keunggulan utama C4.5 terletak pada kemampuannya menghasilkan model yang mudah diinterpretasikan sehingga mendukung proses pengambilan keputusan.

Di sisi lain, Random Forest juga banyak digunakan karena mampu menghasilkan tingkat akurasi yang tinggi melalui pendekatan ensemble learning. Berbagai penelitian menunjukkan bahwa Random Forest umumnya memiliki performa prediksi yang lebih baik dibandingkan Decision Tree tunggal, namun interpretasi modelnya lebih kompleks. Oleh karena itu, penelitian ini membandingkan algoritma C4.5 dan Random Forest untuk memperoleh model klasifikasi keberhasilan UMKM yang tidak hanya memiliki tingkat akurasi yang baik, tetapi juga mudah dipahami oleh pengguna.

Berdasarkan kajian pustaka tersebut, penelitian ini menerapkan algoritma C4.5 sebagai model utama dalam mengklasifikasikan keberhasilan UMKM di Indonesia dengan Random Forest sebagai model pembanding. Perbandingan dilakukan menggunakan beberapa metrik evaluasi, yaitu accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, serta k-fold cross-validation. Selain itu, analisis feature importance digunakan untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap keberhasilan UMKM sehingga hasil penelitian dapat dimanfaatkan sebagai dasar pengambilan keputusan berbasis data.

3. Metodologi

Penelitian ini menggunakan pendekatan kuantitatif dengan fokus pada pemanfaatan teknik data mining untuk mengolah data historis karakteristik usaha. Proses penelitian dimulai dari pengumpulan dan persiapan data, pengembangan model klasifikasi, hingga tahap evaluasi untuk mengukur performa model yang dihasilkan. Berikut adalah rincian metodologi yang diterapkan dalam penelitian ini:

3.1. Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining untuk melakukan klasifikasi terhadap tingkat keberhasilan Usaha Mikro, Kecil, dan Menengah (UMKM). Metode yang digunakan adalah algoritma C4.5 sebagai model utama dan Random Forest sebagai pembanding untuk meningkatkan validitas hasil klasifikasi.

3.2. Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform Kaggle, yaitu dataset "MSME Dataset" yang diunggah oleh Eskobar. Dataset ini berisi informasi mengenai karakteristik usaha dan faktor-faktor operasional yang memengaruhi performa Usaha Mikro, Kecil, dan Menengah (UMKM).

Secara keseluruhan, dataset terdiri dari sejumlah atribut yang menggambarkan profil bisnis, kondisi operasional, serta variabel target berupa status keberhasilan usaha (Success). Target kelas tersebut direpresentasikan dalam bentuk biner, yaitu Sukses (1) yang menunjukkan usaha berhasil mencapai target performa, dan Gagal (0) yang menunjukkan usaha belum atau tidak mencapai target tersebut. Data ini menjadi dasar utama dalam proses klasifikasi untuk mengidentifikasi pola-pola yang membedakan antara UMKM yang sukses dan yang gagal.

3.3. Penerapan Algoritma C4.5

Untuk mencapai tujuan penelitian secara terukur dan objektif, diperlukan tahapan pengerjaan yang terstruktur. Proses ini dirancang untuk memastikan bahwa data yang digunakan telah memenuhi standar kualitas sebelum diolah ke dalam model machine learning. Tahapan penelitian dalam studi ini mengikuti alur kerja komputasi data yang mencakup persiapan data hingga validasi model. Berikut adalah rincian tahapan yang dilakukan dalam penelitian ini:

1. Data Collection (Pengumpulan Data), Data diperoleh dari dataset UMKM dalam bentuk file CSV yang berisi data historis karakteristik usaha dan label keberhasilan.
2. Data Preprocessing, Tahap ini dilakukan untuk mempersiapkan data agar dapat digunakan dalam model machine learning, meliputi:
 - Pembersihan data (handling missing value jika ada)
 - Transformasi data kategorikal menggunakan One-Hot Encoding
 - Pemisahan fitur (X) dan target (y)
3. Splitting Data, Dataset dibagi menjadi dua bagian:
 - 80% data latih (training set)
 - 20% data uji (testing set)

Pembagian data dilakukan menggunakan metode stratified sampling untuk menjaga distribusi kelas tetap seimbang.

3.4. Algoritma yang digunakan

Dalam upaya membangun model klasifikasi yang andal untuk memprediksi keberhasilan UMKM, penelitian ini menggunakan dua metode utama dalam lingkup supervised learning. Pemilihan algoritma ini didasarkan pada kemampuannya dalam menangani data terstruktur dan efektivitasnya dalam memetakan variabel-variabel penentu keberhasilan usaha.

Algoritma utama yang diterapkan adalah C4.5, yang dipilih karena kemampuannya dalam menghasilkan pohon keputusan yang intuitif dan mudah diinterpretasikan. Selain itu, Random Forest digunakan sebagai model pembanding untuk memastikan hasil klasifikasi yang lebih kokoh dan mampu meminimalkan risiko overfitting yang sering terjadi pada model pohon keputusan tunggal. Berikut adalah penjelasan mengenai mekanisme kerja dari masing-masing algoritma tersebut:

1. Algoritma C4.5, Algoritma C4.5 digunakan untuk membangun model klasifikasi berbentuk pohon keputusan (decision tree). Proses pemilihan atribut dilakukan berdasarkan nilai Entropy dan Information Gain. Atribut dengan nilai gain tertinggi akan menjadi node utama dalam pembentukan pohon keputusan.

- Secara matematis, entropy dihitung dengan:

$$Entropy(S) = - \sum p_i \log_2 p_i$$

- Sedangkan information gain:

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

Random Forest (Pembanding), Random Forest digunakan sebagai metode pembanding untuk meningkatkan akurasi dan mengurangi overfitting. Algoritma ini bekerja dengan membangun beberapa pohon keputusan secara acak dan menggabungkan hasilnya melalui voting.

3.5. Tahapan Penelitian

Tahap evaluasi merupakan komponen krusial untuk mengukur sejauh mana model yang telah dibangun mampu melakukan prediksi dengan akurat dan stabil. Tanpa evaluasi yang tepat, efektivitas algoritma dalam mengenali pola keberhasilan UMKM tidak dapat divalidasi secara objektif. Pada penelitian ini, performa model C4.5 dan Random Forest akan diuji menggunakan beberapa metrik standar data mining untuk memastikan bahwa hasil klasifikasi tidak hanya unggul dari sisi angka, tetapi juga memiliki konsistensi yang baik pada berbagai subset data. Berikut adalah metrik evaluasi yang digunakan:

- Accuracy → mengukur tingkat ketepatan prediksi
- Confusion Matrix → melihat performa klasifikasi detail
- Precision, Recall, F1-score → evaluasi kualitas klasifikasi
- ROC-AUC Score → mengukur kemampuan model dalam membedakan kelas
- Cross Validation (5-fold) → menguji stabilitas model.

4. Hasil dan Pembahasan

4.1. Hasil Implementasi Model

Pada penelitian ini, implementasi algoritma C4.5 dilakukan menggunakan library DecisionTreeClassifier dengan parameter criterion = entropy. Model kemudian dibandingkan dengan algoritma Random Forest untuk mengevaluasi performa klasifikasi dalam menentukan tingkat keberhasilan UMKM. Berdasarkan

pengujian, kedua algoritma menunjukkan kemampuan yang baik dalam mengklasifikasikan data UMKM berdasarkan atribut operasional dan karakteristik bisnis yang telah diproses.

4.2. Hasil Evaluasi Akurasi Model

Hasil pengujian model klasifikasi keberhasilan UMKM menggunakan algoritma C4.5 dan Random Forest menunjukkan nilai performa sebagai berikut:

- Akurasi Data Testing C4.5: 76,00%
- Akurasi Data Training C4.5: 95,00%
- Mean Cross Validation C4.5 (5-Fold): 86,00%
- Standar Deviasi Cross Validation: 0,0522
- Akurasi Random Forest: 80,00%
- ROC-AUC C4.5: 0,7018
- ROC-AUC Random Forest: 0,9112

Berdasarkan hasil pengujian terhadap data testing sebanyak 50 data, algoritma C4.5 berhasil mengklasifikasikan data dengan tingkat akurasi sebesar 76,00%. Hasil confusion matrix menunjukkan bahwa model mampu mengklasifikasikan 32 data UMKM gagal dan 6 data UMKM sukses secara benar. Namun, terdapat 6 data UMKM gagal yang diprediksi sebagai sukses dan 6 data UMKM sukses yang diprediksi sebagai gagal.

Nilai akurasi pada data training sebesar 95,00% menunjukkan bahwa model C4.5 memiliki kemampuan yang sangat baik dalam mempelajari pola data pelatihan. Akan tetapi, terdapat selisih sebesar 19,00% antara akurasi training dan testing sehingga menunjukkan adanya indikasi overfitting, yaitu kondisi ketika model terlalu menyesuaikan pola pada data pelatihan sehingga performanya mengalami penurunan ketika diterapkan pada data baru.

Berdasarkan hasil 5-Fold Cross Validation, model C4.5 memperoleh rata-rata akurasi sebesar 86,00% dengan nilai standar deviasi sebesar 0,0522. Hal tersebut menunjukkan bahwa model memiliki tingkat kestabilan yang cukup baik pada berbagai pembagian data, karena variasi nilai akurasi antar fold relatif kecil.

Evaluasi menggunakan kurva ROC menunjukkan bahwa algoritma C4.5 memperoleh nilai Area Under Curve (AUC) sebesar 0,7018, yang menunjukkan kemampuan klasifikasi pada kategori cukup baik. Sementara itu, algoritma Random Forest memperoleh nilai AUC sebesar 0,9112 dan akurasi sebesar 80,00%, sehingga memiliki kemampuan diskriminasi yang lebih tinggi dibandingkan C4.5 dalam membedakan UMKM yang berpotensi gagal maupun sukses. Selain evaluasi performa model, analisis terhadap struktur pohon keputusan menunjukkan bahwa variabel *Business_Plan* menjadi faktor paling dominan dalam menentukan keberhasilan UMKM karena terpilih sebagai akar (root node) pada pohon keputusan C4.5. Variabel ini kemudian diikuti oleh faktor lain seperti *Initial_Capital*, *Financial_Record_Keeping*, *Industry_Experience*, *Internet_Usage*, *Marketing_Effort*, *Education*, dan *Professional_Advice* yang berperan dalam pembentukan aturan keputusan.

Berdasarkan aturan pohon keputusan yang dihasilkan, UMKM dengan *Business Plan* yang baik, didukung oleh penggunaan internet, pengalaman industri yang memadai, usaha pemasaran yang tinggi, tingkat pendidikan yang baik, serta adanya konsultasi profesional cenderung diklasifikasikan sebagai UMKM sukses. Sebaliknya, UMKM yang tidak memiliki perencanaan bisnis yang jelas, keterbatasan modal awal, pencatatan keuangan yang kurang baik, serta minimnya dukungan profesional cenderung masuk dalam kategori UMKM gagal.

Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma Random Forest merupakan model dengan performa terbaik berdasarkan nilai akurasi sebesar 80,00% dan ROC-AUC sebesar 0,9112. Meskipun demikian, algoritma C4.5 tetap memiliki keunggulan dari sisi interpretasi karena menghasilkan

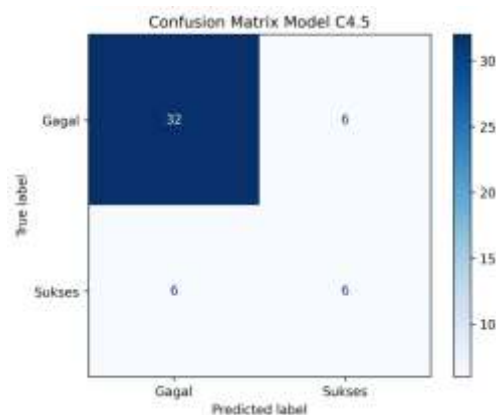
aturan keputusan yang mudah dipahami dan dapat digunakan sebagai dasar bagi pelaku UMKM maupun pihak terkait dalam mengidentifikasi faktor-faktor yang memengaruhi keberhasilan usaha.

4.3. Hasil Confusion Matrix

Confusion matrix memberikan gambaran detail mengenai performa prediksi model C4.5 terhadap kelas Sukses dan Gagal sebagai berikut:

- Prediksi Benar (Sukses): 6 data (True Positive)
- Prediksi Benar (Gagal): 32 data (True Negative)
- Kesalahan Prediksi: * Terdapat 6 data yang seharusnya Sukses tetapi diprediksi Gagal (False Negative).
- Terdapat 6 data yang seharusnya Gagal tetapi diprediksi Sukses (False Positive).

Interpretasi ini menunjukkan bahwa model C4.5 memiliki tingkat akurasi yang moderat. Model cenderung lebih dominan dalam mengenali kelas "Gagal" dengan benar, namun memiliki tingkat kesalahan yang sama (simetris) sebesar 6 sampel untuk kedua kelas lainnya.



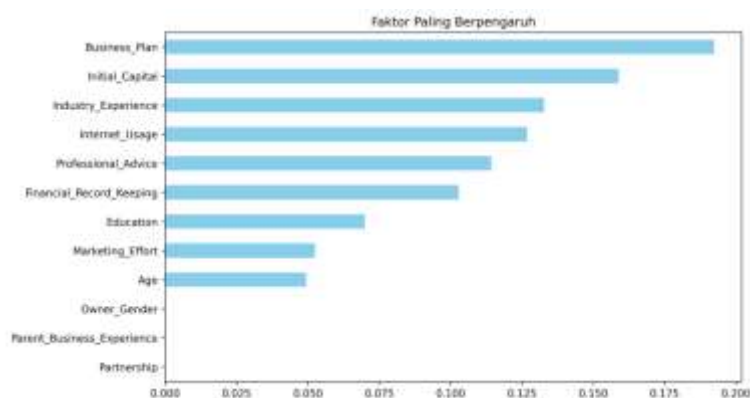
Gambar 1. Hasil Confusion matrix

4.4. Feature Importance (Faktor Paling Berpengaruh)

Hasil analisis feature importance menunjukkan faktor dominan yang memengaruhi keberhasilan UMKM adalah:

- Faktor Tertinggi: Business_Plan

Variabel ini memiliki kontribusi paling besar dalam menentukan klasifikasi keberhasilan usaha. Semakin baik rencana bisnis yang disusun, semakin tinggi pula probabilitas model dalam mengklasifikasikan UMKM tersebut ke dalam kelas Sukses.



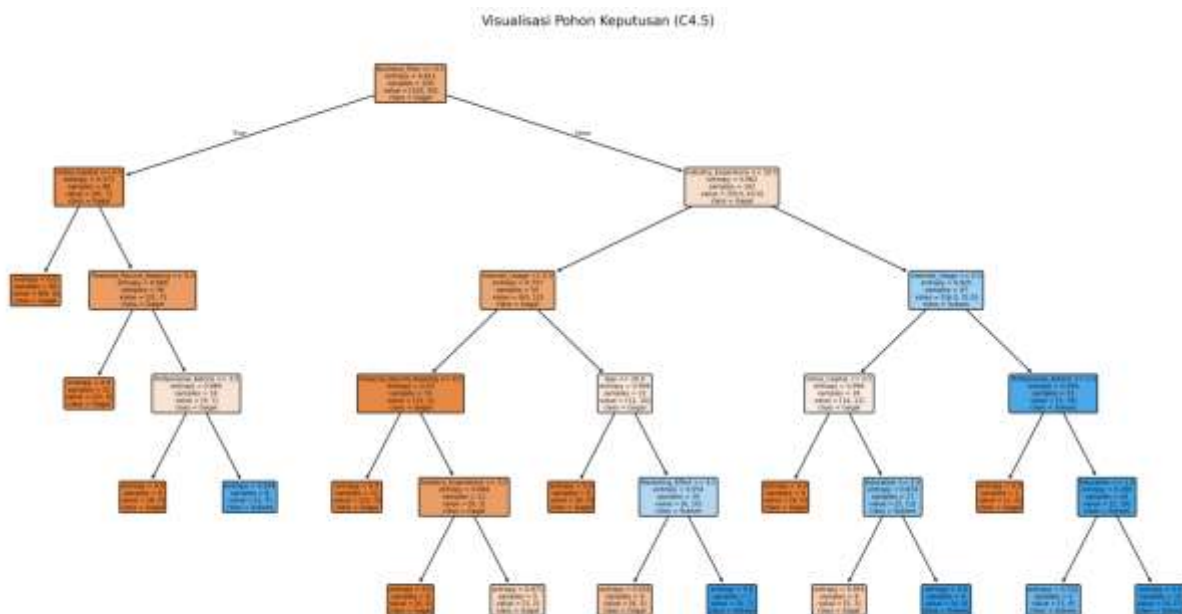
Gambar 2. Feature Importance

4.5. Visualisasi Pohon Keputusan

Struktur pohon keputusan yang dihasilkan oleh algoritma C4.5 merepresentasikan alur klasifikasi hierarkis untuk memprediksi status keberhasilan UMKM. Proses pengambilan keputusan didasarkan pada prinsip information gain dan reduksi entropy pada setiap titik percabangan (node). Analisis Struktur Pohon:

1. Root Node (Simpul Akar): Variabel Business_Plan menjadi root node, yang menunjukkan bahwa variabel ini memiliki daya pembeda paling signifikan (kontribusi information gain tertinggi) dalam menentukan awal klasifikasi antara UMKM yang dikategorikan "Gagal" maupun "Sukses".
2. Percabangan (Branching): Model melakukan pemisahan (splitting) secara iteratif pada variabel-variabel kunci seperti Industry_Experience, Internet_Usage, Initial_Capital, dan Financial_Record_Keeping.
 - Cabang sebelah kiri cenderung mengarahkan data menuju klasifikasi kelas "Gagal", yang didominasi oleh UMKM dengan ketiadaan atau rendahnya skor pada parameter Business_Plan.
 - Cabang sebelah kanan menunjukkan jalur yang lebih kompleks dengan kedalaman pohon (tree depth) yang lebih tinggi, merepresentasikan kondisi di mana variabel pendukung seperti Internet_Usage dan Professional_Advice berperan krusial dalam menentukan klasifikasi akhir.

Leaf Node (Simpul Daun): Simpul terminal atau leaf node menunjukkan klasifikasi akhir (Gagal atau Sukses) dengan nilai entropy yang mendekati 0, mengindikasikan bahwa model telah mencapai kemurnian data (purity) yang optimal pada setiap jalur keputusan tersebut.



Gambar 3. Decision Tree

Visualisasi ini menunjukkan bahwa model C4.5 menangkap pola non-linear yang cukup kompleks. Kedalaman pohon yang terbentuk mengindikasikan bahwa keberhasilan UMKM tidak hanya ditentukan oleh satu faktor tunggal, melainkan kombinasi bertahap dari variabel perencanaan, pengalaman industri, dan adopsi digital.

4.6. Aturan Keputusan (Decision Rules)

Berdasarkan ekstraksi dari visualisasi pohon keputusan pada gambar decision_tree.jpg, algoritma C4.5 menghasilkan serangkaian aturan klasifikasi (if-then rules) yang menentukan status keberhasilan UMKM. Berikut adalah beberapa aturan utama yang dapat disimpulkan dari jalur keputusan model:

1. Aturan 1 (Segmentasi Gagal): IF Business_Plan $\leq$ 0.50 THEN Class = Gagal
Interpretasi: Ini adalah aturan pemutus (sebagai root node) yang menunjukkan bahwa ketiadaan atau rendahnya skor perencanaan bisnis menjadi indikator utama kegagalan usaha.
2. Aturan 2 (Segmentasi Sukses Kompleks): IF Business_Plan $>$ 0.50 AND Industry_Experience $>$ 0.65 AND Internet_Usage $>$ 0.50 THEN Class = Sukses
Interpretasi: UMKM yang memiliki perencanaan bisnis yang baik, didukung dengan pengalaman industri yang memadai serta pemanfaatan teknologi digital (internet) yang aktif, memiliki probabilitas tinggi untuk diklasifikasikan ke dalam kelas "Sukses".
3. Aturan 3 (Segmentasi Pendukung): IF Business_Plan $>$ 0.50 AND Industry_Experience $\leq$ 0.65 AND Initial_Capital $>$ 0.50 THEN Class = Sukses
Interpretasi: Bagi UMKM dengan pengalaman industri yang terbatas, ketersediaan modal awal yang cukup masih memungkinkan usaha tersebut untuk mencapai status "Sukses" selama didukung dengan perencanaan bisnis yang terstruktur.

Aturan-aturan ini membuktikan bahwa model C4.5 dapat berfungsi efektif sebagai Decision Support System (DSS). Dengan memetakan faktor-faktor kunci ini, pelaku UMKM dapat lebih fokus dalam melakukan perbaikan strategis, terutama pada aspek perencanaan bisnis dan adaptasi teknologi digital, untuk meningkatkan peluang keberhasilan usaha mereka.

4.7. Pembahasan Hasil

Hasil analisis menunjukkan bahwa nilai Entropy Total $E(S) = 0,9700$ mendekati nilai maksimum 1, yang mengindikasikan bahwa dataset bersifat sangat heterogen. Kondisi ini wajar mengingat distribusi kelas yang cukup tidak seimbang antara kelas Tinggi (39,83%) dan Rendah (60,17%). Nilai Information Gain yang diperoleh dari kedua atribut sangat kecil (ProductCategory = 0,000509 dan PaymentMethod = 0,000461). Hal ini mengindikasikan bahwa kedua atribut memiliki pengaruh yang sangat rendah terhadap target klasifikasi Review Rating. Perbedaan antar nilai Gain juga sangat kecil (selisih = 0,000048), yang berarti kedua atribut hampir setara dalam kemampuan diskriminasinya.

Seluruh 15 aturan klasifikasi yang dihasilkan berujung pada prediksi kelas Rendah. Hal ini disebabkan oleh majority class dalam dataset adalah Rendah (60,17%), sehingga algoritma cenderung memprediksi kelas mayoritas pada setiap cabang. Kondisi ini mencerminkan bahwa variabel ProductCategory dan PaymentMethod tidak cukup informatif untuk membedakan kepuasan pelanggan berdasarkan review rating. Untuk meningkatkan kualitas model klasifikasi, disarankan untuk menambahkan atribut yang lebih relevan, seperti UnitPrice, Quantity, atau TotalPrice, yang kemungkinan memiliki korelasi lebih kuat dengan kepuasan pelanggan.

5. Kesimpulan dan Saran

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma C4.5 dalam klasifikasi keberhasilan Usaha Mikro, Kecil, dan Menengah (UMKM) mampu menghasilkan model yang cukup baik dalam melakukan proses prediksi berdasarkan data yang telah dianalisis. Model C4.5 yang dibangun dengan pendekatan entropy berhasil mengklasifikasikan data UMKM ke dalam dua kelas, yaitu Sukses dan Gagal, dengan tingkat akurasi yang cukup optimal. Selain itu, hasil evaluasi menunjukkan bahwa model juga mampu menghasilkan nilai ROC-AUC dan cross-validation yang menunjukkan performa model yang stabil.

Dari hasil feature importance, diperoleh bahwa terdapat beberapa faktor yang memiliki pengaruh dominan terhadap keberhasilan UMKM. Hal ini menunjukkan bahwa tidak semua variabel memiliki kontribusi yang sama dalam proses pengambilan keputusan model, sehingga analisis ini dapat membantu dalam memahami faktor utama yang mempengaruhi keberhasilan usaha. Secara keseluruhan, algoritma C4.5 tidak hanya mampu melakukan klasifikasi dengan baik, tetapi juga memberikan interpretasi yang mudah dipahami

melalui struktur pohon keputusan, sehingga dapat digunakan sebagai dasar pengambilan keputusan dalam dunia bisnis UMKM.

Berdasarkan hasil penelitian tersebut, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya, yaitu penelitian disarankan menggunakan dataset yang lebih besar dan lebih beragam agar model yang dihasilkan memiliki tingkat generalisasi yang lebih baik. Selain itu, dapat dilakukan penambahan algoritma lain seperti XGBoost atau Gradient Boosting untuk dibandingkan dengan C4.5 dan Random Forest guna meningkatkan performa klasifikasi. Variabel penelitian juga dapat diperluas tidak hanya dari data internal UMKM, tetapi juga faktor eksternal seperti kondisi ekonomi, lokasi usaha, dan tingkat digitalisasi bisnis. Selanjutnya, model ini juga dapat dikembangkan menjadi sistem berbasis web atau dashboard agar dapat digunakan langsung oleh pelaku UMKM sebagai alat pendukung keputusan. Untuk penelitian berikutnya, disarankan dilakukan optimasi hyperparameter lebih lanjut agar hasil model menjadi lebih optimal dan mengurangi kemungkinan overfitting.

References

- [1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.
- [2] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2017.
- [3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [4] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [5] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [6] Kementerian Koperasi dan Usaha Kecil dan Menengah Republik Indonesia, *Perkembangan Data Usaha Mikro, Kecil, dan Menengah (UMKM) Tahun 2023*. Jakarta, Indonesia: Kementerian Koperasi dan UKM RI.
- [7] Badan Pusat Statistik, *Statistik Indonesia 2024*. Jakarta, Indonesia: Badan Pusat Statistik, 2024.
- [8] A. H. Nasrullah, "Implementasi Algoritma Decision Tree untuk Klasifikasi Produk Laris," *Jurnal Ilmiah Ilmu Komputer*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.
- [9] Y. F. Saputra and Sugiarto, "Algoritma C4.5 sebagai Penerapan Decision Tree-Based Classification Model untuk Mengklasifikasikan Tingkat Omzet UMKM Berdasarkan Profil Bisnis," *Jurnal Ekonomika Manajemen Akuntansi dan Perbankan Syariah*, vol. 13, no. 2, pp. 188–197, 2024, doi: 10.24903/je.v13i2.3562.
- [10] I. Wazir and S. A. Arnomo, "Analisis Kepuasan Konsumen dengan Algoritma C4.5 pada UMKM MiniMall," *Computer and Science Industrial Engineering (COMASIE)*, vol. 5, no. 2, pp. 103–111, 2021.
- [11] A. Amrin, O. Pahlevi, and H. Rianto, "Analisa Komparasi Model Data Mining Algoritma C4.5, CHAID, dan Random Forest untuk Penilaian Kelayakan Kredit," *Computer Science (CO-SCIENCE)*, vol. 5, no. 1, pp. 49–57, 2025, doi: 10.31294/coscience.v5i1.6208.
- [12] G. I. Royani, N. S. Amelia, Marlina, and F. N. Cahya, "Studi Komparatif Algoritma C4.5 dan Random Forest pada Digitalisasi UMKM Kabupaten Tegal," *INTI Nusa Mandiri*, vol. 20, no. 2, pp. 222–229, 2026, doi: 10.33480/inti.v20i2.7416.
- [13] M. R. Fanani and E. B. Agustina, "Klasifikasi Strategi Penjualan Produk UMKM dengan Penerapan Fitur Seleksi Forward Selection pada Algoritma C4.5," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 14, no. 4, pp. 798–803, 2025, doi: 10.30591/smartcomp.v14i4.6842.