Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM

Article *in* JURNAL SISTEM INFORMASI BISNIS · October 2023

DOI: 10.21456/vol13iss2pp127-134

4 authors, including:

Moh. Erkamim
Universitas Tunas Pembangunan Surakarta
29 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Suswadi Suswadi
Universitas Tunas Pembangunan Surakarta
45 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Komparasi Algoritme *Random Forest* dan *XGBoosting* dalam Klasifikasi Performa UMKM

Moh. Erkamim[a,*], Suswadi[a], Muhammad Zidni Subarkah[b], Erni Widarti[a]

[a] Universitas Tunas Pembangunan Surakarta
[b] Universitas Sebelas Maret

Naskah Diterima : 11 Februari 2023; Diterima Publikasi : 1 Oktober 2023
DOI : 10.21456/vol13iss2pp127-134

Abstract

The Covid-19 pandemic has greatly impacted the whole world, especially Indonesia. Various policies have been implemented starting from the implementation of lockdowns, restrictions on large-scale economic activities, and bans from leaving the region. The economic sector is a sector that has been affected quite a lot, one of which is Micro, Small, and Medium Enterprises (MSMEs). As a result of the Covid-19 pandemic, many MSMEs have suffered losses, so many investors have started to consider investing in MSMEs. Therefore, MSMEs need to know their business performance through potential analysis and financial reports to deal with the economic crisis during a pandemic. This study compares two algorithms namely Random Forest and XGBoosting in classifying the good or bad performance of MSME financial conditions. The performance of the developed algorithm will be improved using hyperparameter tuning to obtain the best parameter combination for each algorithm. In this study, the Random Forest algorithm has an accuracy value of 0.944 and an f1-score of 0.944, while the XGBoosting algorithm has an accuracy value of 0.944 and an f1-score of 0.950. Based on the model with the best evaluation metric, six important features are obtained: the 2021 profit and loss variable, 2020 cash, 2020 liabilities, 2020 capital, 2021 sales, and 2021 liabilities.

Keywords: Classification; Algorithm; Random Forest; XGBoosting; MSME; Finance.

Abstrak

Pandemi *Covid-19* memberikan dampak yang besar bagi seluruh dunia, terutama Indonesia. Berbagai kebijakan telah diberlakukan mulai dari pemberlakuan *lockdown*, pembatasan kegiatan ekonomi skala besar, dan larangan keluar wilayah. Sektor perekonomian merupakan sektor yang terdampak cukup besar salah satunya adalah pelaku Usaha Mikro Kecil dan Menengah (UMKM). Akibat adanya pandemi *Covid-19* pelaku UMKM banyak mengalami kerugian sehingga banyak investor yang mulai mempertimbangkan untuk memberikan investasi kepada UMKM. Oleh karena itu, UMKM perlu mengetahui performa bisnisnya melalui analisis potensi dan laporan keuangan untuk menghadapi krisis ekonomi selama pandemi. Penelitian ini mengomparasikan dua Algoritme yakni *Random Forest* dan *XGBoosting* dalam mengklasifikasikan performa baik atau buruk kondisi keuangan UMKM. Algoritme yang dibangun akan ditingkatkan kinerjanya menggunakan *hyperparameter tuning* untuk memperoleh kombinasi parameter terbaik pada masing-masing algoritme. Pada penelitian ini, algoritme *Random Forest* memiliki nilai akurasi 0,944 dan *f1-score* 0,944, sedangkan algoritme *XGBoosting* memiliki nilai akurasi 0,944 dan *f1-score* 0,950. Berdasarkan model dengan metrik evaluasi terbaik maka diperoleh enam *features important* yakni variabel laba rugi 2021, kas 2020, liabilitas 2020, modal 2020, penjualan 2021, dan liabilitas 2021.

Kata kunci: Klasifikasi; Algoritme; Random Forest; XGBoosting; UMKM; Keuangan.

1. Pendahuluan

Masyarakat dunia dikejutkan oleh fenomena *Coronavirus Disease* 19 (*Covid-19*) yang ditetapkan sebagai pandemi oleh *World Health Organization* (WHO) mulai 9 Maret 2020 (Ismawanti, 2021). Seluruh dunia terkena dampak dari wabah *Covid-19*, tak terkecuali Indonesia (Widiyani, 2020). Berbagai kebijakan pencegahan dan penanggulangan wabah ini mulai diberlakukan, antara lain penerapan *lockdown*, pembatasan kegiatan ekonomi skala besar, dan larangan bepergian ke luar wilayah (Dewi dan Tobing, 2021). Mulai dari wabah ini, lahir kebiasaan baru untuk selalu untuk selalu menjaga jarak fisik dan selalu memakai masker agar terhindar dari penyebaran *Covid-19* (Mardiana dan Darmalaksana, 2020). Salah satu sektor yang mengalami dampak terbesar dari kejadian ini adalah sektor perekonomian, yaitu Usaha Mikro Kecil dan Menengah (UMKM) (Nalini, 2021).

UMKM merupakan tulang punggung bisnis ekonomi rakyat yang berperan penting dalam percepatan transformasi struktural, yaitu dengan memperkuat ketahanan ekonomi daerah dan nasional (Siswanti, 2020). Dari tahun ke tahun, perkembangan ekonomi makro maupun mikro cenderung stabil. Di Indonesia, perkembangan UMKM cukup signifikan

*) Corresponding author: erkamim@lecture.utp.ac.id

memegang peran strategis dalam pembangunan ekonomi nasional dan terbukti mampu bertahan ketika krisis ekonomi terjadi pada tahun 1998 (Suci et al., 2017).

Namun, kestabilan berhenti sejak adanya pandemi. Akibat diberlakukannya *lockdown*, sektor UMKM melemah (Thaha, 2020). Banyak terjadi kerugian pada sektor UMKM karena permintaan dan rantai pasokan di seluruh dunia mengalami penurunan. Dalam dunia perekonomian, terdapat beberapa Perseroan Terbatas (PT) yang memberikan pinjaman atau berinvestasi kepada UMKM. Akibat terjadinya pandemi tersebut, PT banyak mengalami kerugian dari investasi ke berbagai UMKM. Sebab, sebagian roda ekonomi yang berjalan dalam PT bergantung pada UMKM. Sebuah UMKM yang mampu mengelola keuangan dan menganalisis potensi dengan baik akan mampu bertahan dalam krisis ekonomi selama pandemi karena UMKM merupakan usaha bisnis rakyat yang sangat tangguh.

Permasalahan di atas disebabkan dari kesalahan sasaran perusahaan dalam berinvestasi. Evaluasi keunggulan, potensi, dan laporan keuangan yang terdapat pada UMKM perlu dianalisis lebih lanjut. Adapun tujuan dari penelitian ini yaitu untuk mengklasifikasikan UMKM berdasarkan laporan keuangan yang dimiliki untuk mengetahui performa baik atau buruknya suatu UMKM. Harapannya dengan adanya analisis klasifikasi ini dapat membantu PT dalam menentukan sasaran investasi yang tepat serta mampu dijadikan evaluasi bagi pelaku UMKM.

Penelitian ini mengomparasikan performa dua algoritme *machine learning* yakni *Random Forest* dan *XGBoosting*. Kedua algoritme tersebut tergolong *supervised learning* karena bekerja menggunakan data yang telah memiliki label sebelumnya. *Random Forest* memiliki keunggulan diantaranya cocok diterapkan pada data berjumlah besar, kuat terhadap data pencilan, dan bekerja dengan baik dengan data nonlinear. Sedangkan *XGBoosting* mampu melakukan pemrosesan paralel, yang dapat mempercepat perhitungan komputasi, memiliki fleksibilitas pengaturan objektif yang besar, dan mengatasi *split* saat *negative loss*. Komparasi performa kedua algoritme menggunakan metrik evaluasi sebagai acuannya yakni nilai akurasi dan *f1-score.*

2. Kerangka Teori

2.1. Random Forest

Metode pengembangan dari *Classification Regression Trees* (CART). Pengembangan dilakukan dengan menerapkan pemilihan fitur secara acak dan teknik *bagging* (agregasi *bootstrap*) (Saragih et al., 2018). *Random Forest* adalah pengklasifikasi yang terdiri dari sekumpulan pohon klasifikasi $\{h(x, S^b), b = 1, \dots, B\}$ dengan $\{S^b\}$ yang tidak berhubungan satu sama lain dan memiliki vektor acak yang sama didistribusikan dan memberikan *vote* x saat masuk kelas terbaik (Zailani & Hanun, 2020). Kami diberi pengklasifikasi *ansemble* $h_1(x)$, $h_2(x)$, ..., h_b dan data latih sampel acak berdasarkan distribusi vektor acak X, Y yang didefinisikan dalam persamaan (1).

$$mg(X,Y) = av_b 1(h_b(X) = Y) - \max_{j \neq Y} av_b\,(h_b(X) = j) \quad (1)$$

Dimana $I(\cdot)$ merupakan fungsi indikator dan av_b merupakan hasil rata-rata dari fungsi indikator, dengan $h_b(X) = Y$ adalah hasil dari prediksi Y dan $h_b(X) = j$ adalah hasil dari prediksi j. *Margin* yang ditentukan digunakan sebagai ukuran seberapa besar nilai rata-rata pada *vote* X, Y untuk kelas yang tepat agar dapat melebihi rata-rata *vote* kelas lainnya. Karena, semakin besar jarak *margin* maka semakin akurat nilainya. Setelah itu, *generalization error* diberikan persamaan (2).

$$PE^* = P_{X,Y}(mg(X,Y) = Y) < 0 \quad (2)$$

Sedangkan PE^* merupakan *geleralization error* dan $P_{x,y}$ sebagai indikasi untuk probabilitas yang melebihi ruang X, Y. Sedangkan, dalam RF, $h_b(X) = h(X, Y^b)$. Untuk pohon dengan jumlah yang banyak, terdapat *Tree Structurer* dan *Strong Law of Large Numbers*. Semakin tingginya jumlah pohon, maka hampir dalam semua barisan S^1,... akan menyebabkan nilai PE^* menjadi konvergen ke persamaan (3).

$$P_{X,Y}(P_S(h(X,S^b) = Y) - \max_{j \neq Y}(P_S(h(X,S^b) = j) < 0) \quad (3)$$

2.2. Extreme Gradient Boosting (XGBoost)

Metode pembelajaran mesin untuk analisis dan klasifikasi regresi yang didasarkan pada *Gradient Boosting Decision Tree* (GBDT) (Rosita et al., 2020). Algoritme *XGBoost* dalam penelitiannya, ia menggabungkan peningkatan dan pengoptimalan untuk membuat *Gradient Boosting Machine* (GBM) (Haumahu et al., 2021; Salamah dan Ramayanti, 2018). Pada metode *boosting*, dibangun model baru untuk mengantisipasi kesalahan model sebelumnya. Penyertaan model tambahan dilakukan hingga tidak ada lagi perbaikan kesalahan yang dapat dilakukan. Peningkatan gradien adalah teknik yang menggunakan penurunan gradien untuk mengurangi kesalahan saat membuat model baru (Sheridan et al., 2016).

2.3. Preprocessing Data

Preprocessing data merupakan tahapan awal dalam algoritme *machine learning*, baik *supervised* maupun *unsupervised*, yang umumnya dilakukan sebelum memproses dan menganalisis data. *Preprocessing text* melibatkan beberapa langkah, seperti *case folding*, *punctuation removal, stopwords removal, tokenization,* dan *stemming* (Miftahusalam et al., 2022). *Case holding* adalah proses mengubah semua huruf menjadi bentuk yang sama, misal huruf

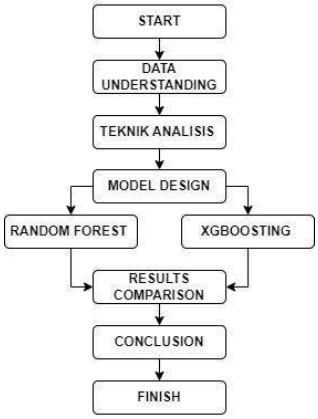
kecil. *Punctuation removal* merupakan proses menghilangkan tanda baca. *Stopwords removal* untuk menghapus kata-kata yang tidak penting. *Tokenization* berupa pembagian kalimat menjadi beberapa bagian yang disebut token. Sedangkan *stemming* berupa proses mendapatkan kata dasar.

2.4. Klasifikasi

Klasifikasi adalah proses mengelompokkan objek atau konsep tertentu ke dalam sejumlah kategori berdasarkan karakteristik yang dimiliki oleh objek atau konsep tersebut (Maxwell et al., 2018). Metode klasifikasi dimaksudkan untuk mempelajari berbagai fungsi yang memetakan semua data terpilih ke dalam salah satu kelompok kelas yang telah ditentukan.

3. Metode

Alur penelitian komparasi Algoritme *Random Forest* dan *XGBoosting* pada data UMKM tahun 2020 dan 2021 ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

3.1. Data

Penelitian ini menggunakan data *cross-section* yang berjumlah 2396 dengan variabel dependennya berupa data kategorik performa UMKM yaitu baik dan buruk, sedangkan untuk variabel lainnya berupa laporan keuangan misalnya penjualan, rasio beban penjualan, sumber laba utama (kategori), laba-rugi, liabilitas, aset, dan kas. Data yang digunakan merupakan data dari tahun 2020 dan 2021.

3.2. Teknik Analisis

Pada penelitian ini, teknis analisis yang digunakan untuk mencapai tujuan penelitian adalah metode klasifikasi dengan mengomparasikan performa dua algoritme yakni *Random Forest* dan *XGBoosting*. Pengukuran performa algoritme didasarkan pada dua matrik evaluasi yakni nilai akurasi dan *f1-score.* Alur pengerjaan analisis pada penelitian ini ditulis sebagai berikut:

3.2.1. Business Understanding

Tahap pertama penelitian ini adalah melakukan pemahaman terkait laporan keuangan atau istilah-istilah dalam akuntansi seperti laba, rugi, kas, aset. Pada tahap ini dibuat *feature* baru seperti ROA, NPM, dan sebagainya. Dalam penelitian ini, variabel laba yang digunakan merupakan laba bersih (setelah pengurangan pajak dan bunga).

3.2.2. Exploratory Data Analysis (EDA)

Tahap kedua yakni data eksplorasi yang dilakukan untuk melihat data secara keseluruhan. Pada tahap ini dilakukan pengecekan karakteristik masing-masing variabel dengan statistik dekriptif dan visualisasi data.

3.2.3. Pre-processing Data

Tahap ketiga yakni *data pre-processing* yang dilakukan untuk menyiapkan data agar lebih terstruktur dan siap digunakan dalam proses klasifikasi (Septianingrum et al., 2021). Pada tahap ini dilakukan pengecekan *missing-value*, penanganan terhadap data yang tidak konsisten, mengubah karakter *decimal* dari (,) menjadi (.), membuat varibel *new features*, melakukan *label encoder*, menganalisis korelasi, membagi data *training* dan *testing*, serta penanganan terhadap data yang tidak seimbang.

3.2.4. Pembentukan Model Klasifikasi

Tahap keempat yakni pembentukan model klasifikasi menggunakan algoritme *machine learning*. Pada penelitian ini digunakan dua algoritme sebagai komparasi performa klasifikasi yakni *Random Forest* dan *XGBoosting* (Mursianto et al., 2021). *Random Forest* merupakan bagian metode dalam *Decision Tree* yang dibangun dengan mengombinasikan setiap *tree* ke dalam satu model. Sedangkan *XGBoosting* merupakan pengembangan algoritme GBM dengan beberapa fitur tambahan yang berguna dalam mempercepat proses komputasi dan mencegah terjadinya *overfitting* (Rosita et al., 2020). Di samping itu, penelitian ini juga melakukan *hyperparameter tuning* untuk menentukan parameter terbaik pada masing-masing algoritme klasifikasi.

3.2.5. Pemilihan Model Terbaik dan Interpretasi

Tahap kelima yakni melakukan pemilihan model terbaik yang didasarkan pada metrik evaluasi. Pada penelitian ini, metrik evaluasi yang digunakan adalah nilai akurasi dan *f1-score*. Nilai akurasi adalah ukuran yang menentukan tingkat kemiripan antara hasil prediksi dengan nilai yang sebenarnya diukur. Sedangkan *f1-score* menggambarkan perbandingan rata-rata *precision* (tingkat ketepatan prediksi) dan *recall* (rasio prediksi benar) yang dibobotkan. Setelah diperoleh model terbaik maka akan dilakukan interpretasi hasil klasifikasi.

4. Hasil dan Pembahasan

4.1. Business Understanding

Business Understanding dalam *data science* adalah pemahaman komprehensif mengenai tujuan,

tantangan, dan kebutuhan spesifik dari perspektif bisnis. Adapun pemahaman dalam konteks bisnis melalui analisis data sebagai berikut:

4.1.1. Net Profit Margin (NPM)

Net profit margin merupakan rasio untuk menunjukkan penjualan bersih perusahaan atas penjualan. Rasio ini membandingkan pendapatan setelah bunga dan pajak dengan pendapatan untuk menentukan profitabilitas perusahaan. Rata-rata standar *margin* laba bersih industri adalah 20%. Rumus NPM yang digunakan ditunjukkan oleh persamaan (4).

$$NPM = \frac{\text{laba bersih}}{penjualan} \times 100\% \quad (4)$$

4.1.2. Return On Asset (ROA)

Return On Assets digunakan untuk menentukan potensi perusahaan dalam menciptakan laba bersih berdasarkan tingkat aset tertentu. Rumus ROA yang digunakan ditunjukkan oleh persamaan (5).

$$ROA = \frac{\text{laba bersih}}{total\ aset} \times 100\% \quad (5)$$

4.1.3. Modal

Modal atau dana induk yang digunakan ditunjukkan oleh persamaan (6).

$$Model = aset - utang\ (liabilitas) \quad (6)$$

4.2. Exploratory Data Analysis (EDA)

4.2.1. Performa UMKM

Data yang membandingkan kinerja UMKM berdasarkan 2396 data dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan Performa UMKM

Performa	Jumlah	Persentase
Baik	168	7%
Buruk	2228	93%

Berdasarkan Tabel 1, dapat dilihat dari 2396 data UMKM 93% memiliki performa yang buruk, sedangkan untuk 7% memiliki performa yang baik. Melihat besarnya perbedaan proporsi status performa, perlu dilakukan penyeimbangan kelas data agar model klasifikasi tetap mampu memprediksi kelas minor.

4.2.2. Sumber Laba Utama 2020

Sumber laba utama pada tahun 2020 dapat disajikan dalam bentuk tabel sebagaimana terlihat pada Tabel 2.

Tabel 2. Sumber Laba Utama 2020

Sumber Laba	Jumlah
Bidang Usaha	2055
Bidang lain	341

Berdasarkan Tabel 2, bahwa pada tahun 2020 sumber laba utama berasal dari "Bidang Usaha" dengan jumlah sebanyak 2055 kasus, sedangkan "Bidang lain" berkontribusi dalam jumlah yang lebih kecil, yaitu sebanyak 341 kasus. Ini menunjukkan bahwa sebagian besar laba pada tahun tersebut berasal dari bidang usaha tertentu, sementara bidang lainnya memberikan kontribusi yang lebih sedikit terhadap laba utama.

4.2.3. Sumber Laba Utama 2021

Sumber laba utama pada tahun 2021 dapat disajikan dalam bentuk tabel sebagaimana terlihat pada Tabel 3.

Tabel 3. Sumber Laba Utama 2021

Sumber Laba	Jumlah
Bidang Usaha	2032
Bidang lain	364

Berdasarkan Tabel 3, dapat diketahui bahwa sumber laba utama pada tahun 2021 dari UMKM merupakan dari bidang usaha sebanyak 2032 dan dari bidang lainnya sebanyak 364.

4.2.4. Status Laba Rugi 2020

Data yang menunjukkan perbandingan laba dan rugi pada tahun 2020 ditunjukkan pada Tabel 4.

Tabel 4. Status Laba Rugi 2020

Status	Persentase
Laba	70%
Rugi	30%

Berdasarkan Tabel 4, menunjukkan bahwa pada tahun 2020, sebagian besar entitas atau perusahaan mengalami laba, dengan persentase laba mencapai 70%, sementara persentase rugi adalah sekitar 30%. Artinya, mayoritas entitas atau perusahaan dalam sampel data ini berhasil mencapai keuntungan, meskipun ada juga sebagian yang mengalami kerugian. Hal ini mencerminkan keragaman kondisi keuangan di pasar pada tahun tersebut.

4.2.5. Status Laba Rugi 2021

Data yang menunjukkan perbandingan laba dan rugi pada tahun 2021 ditunjukkan oleh Tabel 5.

Tabel 5. Status Laba Rugi 2021

Status	Persentase
Laba	68%
Rugi	32%

Berdasarkan Tabel 5, menunjukkan bahwa pada tahun 2021, sebagian besar entitas atau perusahaan dalam sampel data mengalami laba, dengan persentase laba mencapai 68%, sementara persentase rugi adalah sekitar 32%. Meskipun mayoritas entitas masih berhasil mencapai keuntungan, terdapat peningkatan sedikit dalam persentase perusahaan yang mengalami kerugian dibandingkan dengan tahun sebelumnya. Hal

ini menunjukkan adanya perubahan dalam kondisi keuangan di pasar antara tahun 2020 dan 2021.

4.2.6. Statistik Deskriptif

Statistik deskriptif dari data numerik disajikan dalam Tabel 6.

Tabel 6. Statistik Deskriptif dari Data Numerik

Variabel	*Mean*	*Min*	*Max*
Penjualan_2021	9,584888e+07	5,000499e+07	1, 699342e+08
Penjualan_2020	9,641617e+07	5,001518e+07	1,698822e+08
Ratio_Beban-Penjualan_2021	0,509183	0,000098	0,999648
Ratio_Beban-Penjualan_2020	0,498286	0,000007	1,113085e+08
Laba_Rugi_2021	1,819757e+07	-4,775745e+07	1,1753046+08
Laba_Rugi_2020	1,903650e+07	-4,869758e+07	1,175304e+8
Liabilitas_2021	6,009127e+07	2,003301e+07	9,998582e+07

4.3. Pre-Processing Data

4.3.1. Checking Missing Value

Hasil pengecekan *missing value* di dalam data ditampilkan pada Tabel 7.

Tabel 7. *Checking Missing Value*

Data *Missing Value*	Nilai
Penjualan_2021	0
Penjualan_2020	0
Ratio_Beban-Penjualan_2021	0
Ratio_Beban-Penjualan_2020	0
Sumber_Laba_Utama_2021	0
Sumber_Laba_Utama_2020	0
Laba_Rugi_2021	0
Laba_Rugi_2020	0
Liabilitas_2021	0
Liabilitas_2020	0
Aset_2021	0
Aset_2020	0
Kas_2021	0
Kas_2020	0

4.3.2. Penanganan Data Tidak Konsisten

Data yang tidak konsisten yakni terkait penyamarataan penggunaan huruf kapital sebagai contoh, kata 'Baik' dan 'baik' memiliki definisi yang sama, sehingga dilakukan pendefinisian nama yang ditunjukkan oleh Tabel 8, 9, dan 10.

Tabel 8. Penanganan Data yang Tidak Konsisten dalam Permorma UMKM

Sebelum Penanganan		Setelah Penanganan	
Keterangan	Nilai	Keterangan	Nilai
Buruk	2230	Buruk	2230
Baik	133	Baik	166
Baik	33	Baik	0

Tabel 9. Penanganan Data yang Tidak Konsisten dalam Sumber Laba Utama 2020

Sebelum Penanganan		Setelah Penanganan	
Keterangan	Nilai	Keterangan	Nilai
Bidang Usaha	1420	Bidang Usaha	2055
Bidang Lain	48	Bidang Lain	341
Bidang usaha	635	Bidang usaha	0
Bidang Lainnya	291	Bidang Lainnya	0
Bidang lain	2	Bidang lain	0

Tabel 10. Penanganan Data yang Tidak Konsisten dalam Sumber Laba Utama 2021

Sebelum Penanganan		Setelah Penanganan	
Keterangan	Nilai	Keterangan	Nilai
Bidang Usaha	1296	Bidang Usaha	2032
Bidang Lain	736	Bidang lain	364
Bidang usaha	360	Bidang usaha	0
Bidang Lainnya	2	Bidang Lainnya	0
Bidang lain	2	Bidang lain	0

4.3.3. Mengubah Tanda Desimal (,) Menjadi (.)

Dalam *python* untuk definisi koma menggunakan tanda titik, sehingga data yang sebelumnya berbentuk (,) diubah menjadi tanda (.) dan berikut lima data teratas setelah penanganan ditunjukkan pada Tabel 11.

Tabel 11. Pengubahan Tanda Desimal Koma menjadi Titik

Sebelum penanganan		Setelah penanganan	
Index	Rasio Beban Penjualan	*Index*	Rasio Beban Penjualan
0	0,855825664	0	0.855825664
1	0,731139801	1	0.731139801
2	0,080935538	2	0.080935538
3	0,989462669	3	0.989462669
4	0,162016308	4	0.162016308

4.3.4. Membuat New Features

Features tambahan yakni modal, ROA, dan NPM yang ditunjukkan oleh persamaan di bawah.

$$MODAL\_2020 = ASET\_2020 - ASET\_2020 \quad (7)$$

$$MODAL\_2021 = ASET\_2021 - ASET\_2021 \quad (8)$$

$$NPM\_2020'] = \frac{\text{LABA\_2020}}{\text{PENJUALAN\_2020}} x100 \quad (9)$$

$$NPM\_2021 = \frac{\text{LABA\_2021}}{\text{PENJUALAN\_2021}} x100 \quad (10)$$

$$Roa = \frac{laba\ bersih}{total\ aset}\ x\ 100\% \quad (11)$$

4.3.5. Label Encoder

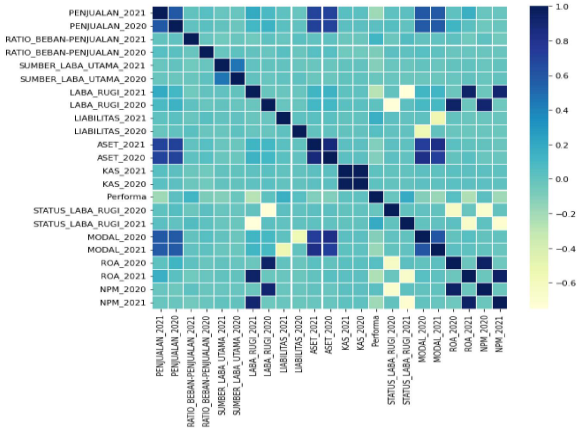
Label *encoder* untuk mengubah variabel kategorik menjadi bentuk numerik yang ditunjukkan pada Gambar 6.

```
from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
df['Performa'] = labelencoder.fit_transform(df['Performa'])
df['SUMBER_LABA_UTAMA_2021'] = labelencoder.fit_transform(df['SUMBER_LABA_UTAMA_2021'])
df['SUMBER_LABA_UTAMA_2020'] = labelencoder.fit_transform(df['SUMBER_LABA_UTAMA_2020'])
df['STATUS_LABA_RUGI_2020'] = labelencoder.fit_transform(df['STATUS_LABA_RUGI_2020'])
df['STATUS_LABA_RUGI_2021'] = labelencoder.fit_transform(df['STATUS_LABA_RUGI_2021'])
```

Gambar 2. Label *encoder*

4.3.6. Analisis Korelasi

Analisis korelasi pada beberapa variabel UMKM ditunjukkan pada Gambar 3.



Gambar 3. Analisis korelasi

Pada Gambar 3 ditunjukkan bahwa hubungan variabel apabila mendekati nilai 1 atau -1 maka korelasinya akan semakin kuat, sebaliknya jika mendekati nilai 0 maka korelasinya akan semakin lemah. Adapun nilai korelasi pada setiap variabel ditunjukkan pada Tabel 7.

Tabel 7. Nilai Variabel

Variabel	Nilai
Performa	1.000000
Status_Laba_Rugi_2021	0.186855
Liabilitas_2021	0.162787
Ratio_Beban-Penjualan_2021	0.123039
Roa_2020	0.048209
Liabilitas_2020	0.045755
Penjualan_2020	0.025341
NPM_2020	0.024494
Kas_2021	0.020908
Kas_2020	0.020398
Laba_Rugi_2020	0.018868
Status_Laba_Rugi_2020	-0.011680
Sumber_Laba_Utama_2020	-0.021758
Ratio_Beban-Penjualan_2020	-0.094683
Aset_2021	-0.101107
Aset_2020	-0.106052
Modal_2020	-0.114113
Sumber_Laba_Utama_2021	-0.115476
Modal_2021	-0.177387
NPM_2021	-0.187789
Penjualan_2021	-0.187812
Roa_2021	-0.242454
Laba_rugi_2021	-0.262893

4.3.7. *Pembagian Data Training dan Testing*

Pembagian data *training* dan *testing* yakni menggunakan 80% untuk data *training* dan 20% untuk data *testing*.

4.3.8. *Penanganan Data Tidak Seimbang*

Algoritme *Synthetic Minority Oversampling Technique* (SMOTE) digunakan untuk menangani ketidakseimbangan data. Penangan ini dilakukan agar data label untuk klasifikasi dapat seimbang pada data *training* sehingga analisis pada data *testing* yang dihasilkan tidak bisa dan menghasilkan akurasi yang baik. Jumlah data sebelum dan sesudah penyeimbangan ditunjukkan pada Tabel 8.

Tabel 8. Penanganan Data Tidak Seimbang

Performa	Jumlah Data	
	Sebelum Penyeimbangan	Setelah Penyeimbangan
Baik	137	1779
Buruk	1779	1779

4.4. *Pembentukan Model*

Penelitian ini menggunakan metode klasifikasi dengan dua algoritme, yaitu *Random Forest* dan *XGBoosting*. Masing-masing algoritme ditingkatkan performanya dengan mengombinasikan parameter terbaik menggunakan metode *hyperparameter tuning*. Pada algoritme *Random Forest* parameter yang dikombinasikan adalah *criterion* (untuk mengukur kualitas split), *max_features* (jumlah fitur yang perlu dipertimbangkan saat mencari pemisahan terbaik), dan *n_estimators* (jumlah *tree* yang dibangun). Kombinasi nilai parameter terbaik pada algoritme *Random Forest* ditunjukkan pada Tabel 9.

Tabel 9. Parameter Terbaik pada Algoritme *Random Forest*

Parameter	Nilai
Criterion	gini
Max_features	auto
N_estimators	1000

Kemudian pada algoritme *XGBoosting* parameter yang dikombinasikan adalah *learning_rate* (nilai koreksi bobot pada waktu proses *training*), *max_depth* (maksimum kedalaman *tree*), dan *subsample* (Rasio subsampel dari *training instances*). Tabel 10 menunjukkan kombinasi nilai parameter terbaik pada algoritme *XGBoosting*.

Tabel 10. Parameter Terbaik pada Algoritme *XGBoosting*

Parameter	Nilai
Learning_Rate	0,1
Max_Depth	5
Subsample	0,5

4.5. *Pemilihan Model Terbaik*

Komparasi performa algoritme *Random Forest* dan *XGBoosting* mengacu pada nilai akurasi dan *f1-score* yang ditunjukkan oleh Tabel 11.

Tabel 11. Ringkasan Model Terbaik

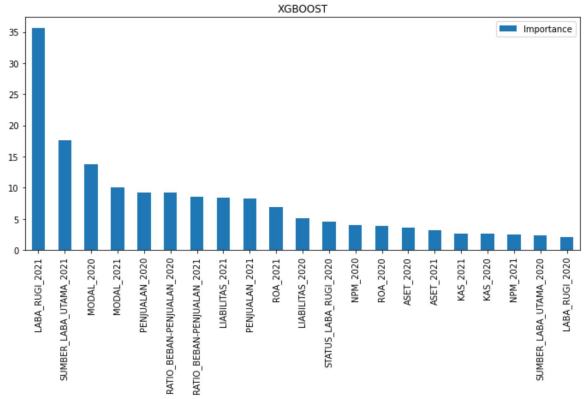
Algoritme	Akurasi	F_1-*score*
Random Forest	0,944	0,944
XGBoosting	0,944	0,950

Berdasarkan Tabel 11, algoritme *Random Forest* dan *XGBoosting* memperoleh nilai akurasi yang sama, sedangkan nilai *f1-score* yang diperoleh *XGBoosting* lebih tinggi dibandingkan *Random Forest*. Dengan demikian, algoritme *XGBoosting* memiliki performa lebih baik dibandingkan algoritme *Random Forest* dengan nilai akurasi 0,944 dan nilai *f1-score* 0,950

4.6. *Features Important*

Setelah memperoleh model terbaik maka langkah selanjutnya adalah menentukan *features important* dari model yang telah dibuat yakni variabel laba rugi

2021, kas 2020, liabilitas 2020, modal 2020, penjualan 2021, dan liabilitas 2021 yang ditunjukkan oleh Gambar 4.



Gambar 4. *Features Important*

5. Kesimpulan

Berdasarkan hasil dan pembahasan dapat disimpulkan bahwa performa algoritme *XGBoosting* lebih baik dibandingkan dengan algoritme *Random Forest* dalam melakukan klasifikasi performa UMKM menurut faktor-faktor keuangan tahun 2020-2021. Algoritme *XGBoosting* yang dibangun memiliki nilai parameter *learning_rate* 0,1, *max_depth* 5, dan *subsample* 0,5. Kemudian diperoleh metrik evaluasi hasil klasifikasi yakni nilai akurasi 0,944 dan nilai *f1-score* 0,950. Di samping itu, berdasarkan model yang telah dibuat diperoleh enam *features important* yakni variabel laba rugi 2021, kas 2020, liabilitas 2020, modal 2020, penjualan 2021, dan liabilitas 2021. *Features important* yang diperoleh dapat digunakan PT dalam menentukan sasaran investasi yang tepat berdasarkan enam faktor tersebut serta sebagai bahan evaluasi keuangan bagi para pelaku UMKM.

Daftar Pustaka

Dewi, D. S., & Tobing, T. N. W. (2021). Optimalisasi Penyelenggaran Pelayanan Publik Dalam Masa Perubahan Melawan Covid-19 Di Indonesia. Journal of Information System, Applied, Management, Accounting and Research, 5(1), 210. https://doi.org/10.52362/jisamar.v5i1.362

Haumahu, J. P., Permana, S. D. H., & Yaddarabullah, Y. (2021). Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost). IOP Conference Series: Materials Science and Engineering, 1098(5), 052081. https://doi.org/10.1088/1757-899X/1098/5/052081

Ismawanti, R. (2021). Dampak Manajemen Perubahan Lingkungan Kerja Masa Pandemi Covid-19 Terhadap Pegawai PT Telkom Indonesia Tbk DIVREG 3 Jawa Barat. Kebijakan: Jurnal Ilmu Administrasi, 12(1), 57-62. https://doi.org/10.23969/kebijakan.v12i1.3468

Mardiana, D., & Darmalaksana, W. (2020). Relevansi syahid ma'nawi dengan peristiwa pandemic covid-19. Jurnal Perspektif, 4(1), 12-20.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784-2817. https://doi.org/10.1080/01431161.2018.1433343

Miftahusalam, A., Nuraini, A. F., Khoirunisa, A. A., & Pratiwi, H. (2022). Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine Pada Analisis Sentimen Twitter Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer. Seminar Nasional Official Statistics, 2022(1), 563-572. https://doi.org/10.34123/semnasoffstat.v2022i1.1410

Mursianto, G. A., Falih, I. M., Irfan, M., Sakinah, T., & Prasvita, D. S. (2021). Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan. Senamika, 2(2), 41-50.

Nalini, S. N. L. (2021). Dampak Dampak covid-19 terhadap Usaha MIkro, Kecil dan Menengah. Jesya (Jurnal Ekonomi & Ekonomi Syariah), 4(1), https://doi.org/10.36778/jesya.v4i1.278

Rosita, R., Zailani, A. U., Hanun, N. L., Maxwell, A. E., Warner, T. A., Fang, F., Kotsiantis, S. B., Zaharakis, I. D., Pintelas, P. E., Haumahu, J. P., Permana, S. D. H., Yaddarabullah, Y., Salamah, U., Ramayanti, D., Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., Gifford, E. M., … He, T. (2020). Supervised Classification of Indonesian Text DocumentUsing Extreme Gradient Boosting (XGBoost). International Journal of Remote Sensing, 56(5), 79-84.

Salamah, U., & Ramayanti, D. (2018). Supervised Classification of Indonesian Text DocumentUsing Extreme Gradient Boosting (XGBoost). 5(5), 79-84.

Saragih, G. S., Rustam, Z., Bustaman, examiner A., & Sarwinda, examiner D. (2018). Prediksi kebangkrutan bank dengan menggunakan random forest = Predict bank failures using random forest.

Septianingrum, F., Jaman, J. H., & Enri, U. (2021). Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier. Jurnal Media Informatika Budidarma, 5(4), 1431. https://doi.org/10.30865/mib.v5i4.3260

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. Journal of Chemical Information and Modeling, 56(12), 2353-2360. https://doi.org/10.1021/acs.jcim.6b00591

Siswanti, T. (2020). Analisis Pengaruh Faktor Internal Dan Eksternal Terhadap Kinerja Usaha Mikro Kecil Dan Menengah (Umkm). Jurnal Bisnis &

Akuntansi Unsurya, 5(2), 61-76. https://doi.org/10.35968/jbau.v5i2.430

Suci, Y. R., Tinggi, S., & Ekonomi, I. (2017). Perkembangan UMKM (Usaha Mikro Kecil Menengah) di Indonesia. Jurnal Ilmiah Fakultasi Ekonomi, 6(1), 51-58.

Thaha, A. F. (2020). Dampak Covid-19 Terhadap UMKM Di Indonesia [The Impact of Covid-19 on MSMEs in Indonesia]. Jurnal Brand, 2(1), 148-153.

Widiyani. (2020). Latar Belakang Virus Corona, Perkembangan hingga Isu Terkini. detikNews.

Zailani, A. U., & Hanun, N. L. (2020). Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera. Infotech: Journal of Technology Information, 6(1), 7-14. https://doi.org/10.37365/jti.v6i1.61